A First Course in Abstract Algebra

Rings, Groups, and Fields

Second Edition

Marlow Anderson
Todd Feil

Chapman & Hall/CRC
Taylor & Francis Group
Boca Raton London New York Singapore

Library of Congress Cataloging-in-Publication Data

Anderson, Marlow, 1950-
A first course in abstract algebra : rings, groups, and fields / by Marlow Anderson and Todd Feil.—2nd ed.
p. cm.
Includes index.
ISBN 1-58488-515-7 (alk. paper)
1. Algebra, Abstract. I. Feil, Todd, 1951- II. Title.

QA162.A53 2004
512'.02--dc22 2004061805

Direct all inquiries to CRC Press, 2000 N.W. Corporate Blvd., Boca Raton, Florida 33431.

No claim to original U.S. Government works
International Standard Book Number 1-58488-515-7
Library of Congress Card Number 2004061805
Printed in the United States of America 1 2 3 4 5 6 7 8 9 0
Printed on acid-free paper

Contents

Preface xiii

I Numbers, Polynomials, and Factoring

1 The Natural Numbers 3
1.1 Operations on the Natural Numbers 3
1.2 Well Ordering and Mathematical Induction 4
1.3 The Fibonacci Sequence 7
1.4 Well Ordering Implies Mathematical Induction 9
1.5 The Axiomatic Method 9

2 The Integers 15
2.1 The Division Theorem 15
2.2 The Greatest Common Divisor 17
2.3 The GCD Identity 20
2.4 The Fundamental Theorem of Arithmetic 22
2.5 A Geometric Interpretation 24

3 Modular Arithmetic 31
3.1 Residue Classes 31
3.2 Arithmetic on the Residue Classes 34
3.3 Properties of Modular Arithmetic 36

4 Polynomials with Rational Coefficients 41
4.1 Polynomials 41
4.2 The Algebra of Polynomials 43
4.3 The Analogy between $\mathbb{Z}$ and $\mathbb{Q}[x]$ 45
4.4 Factors of a Polynomial 47
4.5 Linear Factors 47
4.6 Greatest Common Divisors 49

5 Factorization of Polynomials 55

5.1 Factoring Polynomials 55

5.2 Unique Factorization 57

5.3 Polynomials with Integer Coefficients 59

Section I in a Nutshell 69

II Rings, Domains, and Fields

6 Rings 73

6.1 Binary Operations 73

6.2 Rings 74

6.3 Arithmetic in a Ring 80

6.4 Notational Conventions 81

6.5 The Set of Integers is a Ring 82

7 Subrings and Unity 89

7.1 Subrings 89

7.2 The Multiplicative Identity 93

8 Integral Domains and Fields 101

8.1 Zero Divisors 101

8.2 Units 103

8.3 Fields 105

8.4 The Field of Complex Numbers 105

8.5 Finite Fields 110

9 Polynomials over a Field 119

9.1 Polynomials with Coefficients from an Arbitrary Field . 119

9.2 Polynomials with Complex Coefficients 121

9.3 Irreducibles in $\mathbb{R}[x]$ 124

9.4 Extraction of Square Roots in $\mathbb{C}$ 125

Section II in a Nutshell 135

III Unique Factorization

10 Associates and Irreducibles 141

10.1 Associates 141

10.2 Irreducibles 142

10.3 Quadratic Extensions of the Integers 144

10.4 Units in Quadratic Extensions 145

10.5 Irreducibles in Quadratic Extensions 149

11 Factorization and Ideals 155

11.1 Factorization for Quadratic Extensions 155

11.2 How Might Factorization Fail? 157

11.3 Ideals 158

11.4 Principal Ideals 160

12 Principal Ideal Domains 169

12.1 Ideals that are not Principal 169

12.2 Principal Ideal Domains 171

13 Primes and Unique Factorization 177

13.1 Primes 177

13.2 UFDs 179

13.3 Expressing Properties of Elements in Terms of Ideals . 180

13.4 Ideals in $\mathbb{Z}[\sqrt{-5}]$ 183

13.5 A Comparison between $\mathbb{Z}$ and $\mathbb{Z}[\sqrt{-5}]$ 183

13.6 All PIDs are UFDs 184

14 Polynomials with Integer Coefficients 189

14.1 The Proof that $\mathbb{Q}[x]$ is a UFD 189

14.2 Factoring Integers out of Polynomials 190

14.3 The Content of a Polynomial 191

14.4 Irreducibles in $\mathbb{Z}[x]$ are Prime 193

15 Euclidean Domains 197

15.1 Euclidean Domains 197

15.2 The Gaussian Integers 199

15.3 Euclidean Domains are PIDs 201

15.4 Some PIDs are not Euclidean 203

Section III in a Nutshell 207

IV Ring Homomorphisms and Ideals

16 Ring Homomorphisms 211

16.1 Homomorphisms 211

16.2 One-to-one and Onto Functions 214

16.3 Properties Preserved by Homomorphisms 215

16.4 More Examples 216
16.5 Making a Homomorphism Onto 218

17 The Kernel **225**
17.1 Ideals 226
17.2 The Kernel 226
17.3 The Kernel is an Ideal 228
17.4 All Pre-images Can Be Obtained from the Kernel . . . 229
17.5 When is the Kernel Trivial? 232
17.6 A Summary and Example 232

18 Rings of Cosets **237**
18.1 The Ring of Cosets 237
18.2 The Natural Homomorphism 240

19 The Isomorphism Theorem for Rings **247**
19.1 Isomorphism 247
19.2 The Fundamental Isomorphism Theorem 249
19.3 Examples 251

20 Maximal and Prime Ideals **259**
20.1 Maximal Ideals 259
20.2 Prime Ideals 262

21 The Chinese Remainder Theorem **271**
21.1 Direct Products of Domains 271
21.2 Chinese Remainder Theorem 274

Section IV in a Nutshell **283**

V Groups

22 Symmetries of Figures in the Plane **287**
22.1 Symmetries of the Equilateral Triangle 287
22.2 Permutation Notation 290
22.3 Matrix Notation 292
22.4 Symmetries of the Square 294

23 Symmetries of Figures in Space **299**
23.1 Symmetries of the Regular Tetrahedron 300
23.2 Symmetries of the Cube 304

24 Abstract Groups **313**
24.1 Definition of Group 314
24.2 Examples of Groups 314
24.3 Multiplicative Groups 316

25 Subgroups **329**
25.1 Arithmetic in an Abstract Group 329
25.2 Notation 330
25.3 Subgroups 331
25.4 Characterization of Subgroups 333

26 Cyclic Groups **339**
26.1 The Order of an Element 339
26.2 Rule of Exponents 342
26.3 Cyclic Subgroups 345
26.4 Cyclic Groups 347

Section V in a Nutshell **353**

VI Group Homomorphisms and Permutations

27 Group Homomorphisms **357**
27.1 Homomorphisms 357
27.2 Examples 358
27.3 Direct Products 361

28 Group Isomorphisms **367**
28.1 Structure Preserved by Homomorphisms 368
28.2 Uniqueness of Cyclic Groups 369
28.3 Symmetry Groups 371
28.4 Characterizing Direct Products 372

29 Permutations and Cayley's Theorem **379**
29.1 Permutations 379
29.2 The Symmetric Groups 380
29.3 Cayley's Theorem 383

30 More About Permutations **389**
30.1 Cycles 389
30.2 Cycle Factorization of Permutations 391
30.3 Orders of Permutations 394

31 Cosets and Lagrange's Theorem **399**
31.1 Cosets . . . 399
31.2 Lagrange's Theorem . . . 401
31.3 Applications of Lagrange's Theorem . . . 404

32 Groups of Cosets **413**
32.1 Left Cosets . . . 414
32.2 Normal Subgroups . . . 415
32.3 Examples of Groups of Cosets . . . 417

33 The Isomorphism Theorem for Groups **425**
33.1 The Kernel . . . 425
33.2 Cosets of the Kernel . . . 428
33.3 The Fundamental Theorem . . . 429

34 The Alternating Groups **435**
34.1 Transpositions . . . 435
34.2 The Parity of a Permutation . . . 436
34.3 The Alternating Groups . . . 438
34.4 The Alternating Subgroup is Normal . . . 439
34.5 Simple Groups . . . 442

35 Fundamental Theorem for Finite Abelian Groups **449**
35.1 The Fundamental Theorem . . . 449
35.2 p-groups . . . 452

36 Solvable Groups **455**
36.1 Solvability . . . 455
36.2 New Solvable Groups from Old . . . 457

Section VI in a Nutshell **461**

VII Constructibility Problems

37 Constructions with Compass and Straightedge **465**
37.1 Construction Problems . . . 465
37.2 Constructible Lengths and Numbers . . . 467

38 Constructibility and Quadratic Field Extensions **475**
38.1 Quadratic Field Extensions . . . 475
38.2 Sequences of Quadratic Field Extensions . . . 477

38.3 The Rational Plane . . . 479
38.4 Planes of Constructible Numbers . . . 480
38.5 The Constructible Number Theorem . . . 484

39 The Impossibility of Certain Constructions **489**
39.1 Doubling the Cube . . . 489
39.2 Trisecting the Angle . . . 490
39.3 Squaring the Circle . . . 493

Section VII in a Nutshell **499**

VIII Vector Spaces and Field Extensions

40 Vector Spaces I **503**
40.1 Vectors . . . 504
40.2 Vector Spaces . . . 505

41 Vector Spaces II **511**
41.1 Spanning Sets . . . 511
41.2 A Basis for a Vector Space . . . 514
41.3 Finding a Basis . . . 518
41.4 Dimension of a Vector Space . . . 520

42 Field Extensions and Kronecker's Theorem **527**
42.1 Field Extensions . . . 527
42.2 Kronecker's Theorem . . . 528
42.3 The Characteristic of a Field . . . 530

43 Algebraic Field Extensions **537**
43.1 The Minimal Polynomial for an Element . . . 537
43.2 Simple Extensions . . . 539
43.3 Simple Transcendental Extensions . . . 544
43.4 Dimension of Simple Algebraic Extensions . . . 545

44 Finite Extensions and Constructibility Revisited **551**
44.1 Finite Extensions . . . 551
44.2 Constructibility Problems . . . 556

Section VIII in a Nutshell **561**

IX Galois Theory

45 The Splitting Field **565**
45.1 The Splitting Field 566
45.2 Fields with Characteristic Zero 570

46 Finite Fields **577**
46.1 Existence and Uniqueness 577
46.2 Examples 580

47 Galois Groups **585**
47.1 The Galois Group 585
47.2 Galois Groups of Splitting Fields 588

48 The Fundamental Theorem of Galois Theory **599**
48.1 Subgroups and Subfields 599
48.2 Symmetric Polynomials 601
48.3 The Fixed Field and Normal Extensions 602
48.4 The Fundamental Theorem 604
48.5 Examples 607

49 Solving Polynomials by Radicals **615**
49.1 Field Extensions by Radicals 615
49.2 Refining the Root Tower 618
49.3 Solvable Galois Groups 622

Section IX in a Nutshell **629**

Hints and Solutions **633**

Guide to Notation **661**

Index **665**

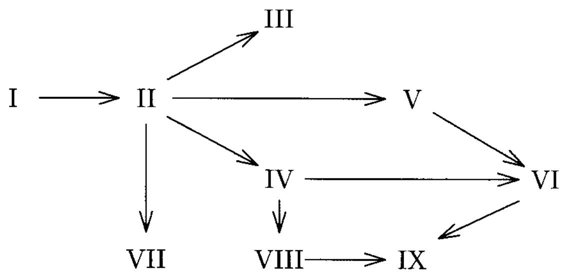
Preface

Traditionally, a first course in abstract algebra introduces groups, rings, and fields, in that order. In contrast, we have chosen to develop ring theory first, in order to draw upon the student's familiarity with integers and with polynomials, which we use as the motivating examples for studying rings.

This approach has worked well for us in motivating students in the study of abstract algebra and in showing them the power of abstraction. Our students have found the process of abstraction easier to understand, when they have more familiar examples to base it upon. We introduce groups later on, again by first looking at concrete examples, in this case symmetries of figures in the plane and space. By this time students are more experienced, and they handle the abstraction much more easily. Indeed, these parts of the text move quite quickly, which initially surprised (and pleased) the authors.

There is more material here than can be used in one semester. Those teaching a one-semester course may choose among various topics they might wish to include. There is sufficient material in this text for a two-semester course, probably more, in most cases. The text is divided into large sections (numbered with Roman numerals), each containing a number of short chapters (numbered with Arabic numerals). Each chapter is in turn divided into subsections, which are numbered using a decimal system: 38.2 is the second subsection of Chapter 38. Within a given chapter each mathematical statement (theorem, lemma or proposition) is numbered consecutively in decimal style, to make cross referencing easy. The examples in a chapter are also numbered consecutively in the same style. Cross references to exercises are done in the same way: Exercise 38.2 refers to the second exercise in Chapter 38.

The diagram below indicates the dependency of the large sections.



Section I (*Numbers, Polynomials, and Factoring*) introduces the integers $\mathbb{Z}$, and the polynomials $\mathbb{Q}[x]$ over the rationals. In both cases we emphasize the idea of factoring into irreducibles, pointing out the structural similarities. We also introduce the rings of integers modulo n in this section. Induction, the most important proof technique used in the early part of this text, is introduced in Chapter 1.

In Section II (*Rings, Domains, and Fields*) we define a ring as the abstract concept encompassing our specific examples from Section I. We define integral domains and fields and then look at polynomials over an arbitrary field. We make the point that the important properties of $\mathbb{Q}[x]$ are really due to the fact that we have coefficients from a field; this gives students a nice example of the power of abstraction.

In Section III (*Unique Factorization*) we explore more general contexts in which unique factorization is possible. Along the way, we introduce the important notion of ideal of a commutative ring. We find that some experience with ideals, prior to encountering the Fundamental Isomorphism Theorem, helps make that difficult topic more understandable. Chapter 13 concludes with the theorem that every principal ideal domain is a unique factorization domain. In the interest of time, many instructors may wish to skip the last two chapters of this section.

Section IV (*Ring Homomorphisms and Ideals*) has as its main goal the proof of the Fundamental Isomorphism Theorem. This section does not logically depend upon Section III, but we much prefer that students first encounter the notion of ideal in the first few chapters of Section III. Section IV also includes a chapter about the connection between maximal ideals and fields, and prime ideals and domains. There is also an optional chapter on the Chinese Remainder Theorem.

Section V (*Groups*) begins with two chapters on symmetries, those in the plane and those in space, in order to give students some concrete examples of non-abelian groups. We then define abstract groups. By this time, students have some experience with abstract algebra, and so the instructor should find this part of the text moves fairly quickly, with students anticipating results and ideas.

Section VI (*Permutations and Group Homomorphisms*) does enough permutation group theory to give students plenty of groups to compute with and has as one of its goals the Fundamental Isomorphism Theorem for groups. The last three chapters are largely optional, unless you wish to include Section IX.

Section VII (*Constructibility Problems and Field Extensions*) is an optional section that is a great example of the power of abstract algebra. In it, we show that the three Greek constructibility problems using a compass and straightedge are impossible. This section does not use Kronecker's Theorem and is very computational in flavor. It does not depend on knowing any group theory and can be taught immediately after Section IV, if the instructor wishes to delay the introduction of groups.

We revisit the impossibility proofs in Section VIII (*Vector Spaces and Field Extensions*), where we give enough vector space theory to introduce students to the theory of algebraic field extensions. Seeing the impossibility proofs again, in a more abstract context, emphasizes the power of abstract field theory.

Section IX develops Galois Theory with the goal of showing the impossibility of solving the quintic with radicals. This section depends heavily on Section VIII, as well as Chapters 34 and 36.

Each chapter includes Quick Exercises, which are intended to be done by the student as the text is read (or perhaps on the second reading) to make sure the topic just covered is understood. Quick Exercises are typically straightforward, rather short verifications of facts just stated that act to reinforce the information just presented. They also act as an early warning to the student that something basic was missed. We often use some of them as a basis for in-class discussion. The exercises following each chapter begin with the Warm-up Exercises, which test fundamental comprehension and should be done by all students. These are followed by the regular exercises, which include both computational problems and supply-the-proof problems. Answers to most exercises that do not require proof are given in the Hints and Answers section. Hints to many problems are also given there.

Historical remarks follow many of the chapters. For the most part we try to make use of the history of algebra to make certain pedagogical points. We find that students enjoy finding out a bit about the history of the subject, and how long it took for some of the concepts of abstract algebra to evolve. We've relied on such authors as Boyer & Merzbach, Eves, Burton, Kline, and Katz for this material.

We find that in a first (or second) course, students lose track of the forest, getting bogged down in the details of new material. With this in mind, we've ended each section with a short synopsis that we've called a "Nutshell" in which we've laid out the important definitions and theorems developed in that section, sometimes in an order slightly altered from the text. It's a way for the student to organize their thoughts on the material and see what the major points were, in case they missed them the first time through.

We include an appendix entitled "Guide to Notation", which provides a list of mathematical notations used in the book, and citations to where they are introduced in the text. We group the notations together conceptually. There is also a complete index, which will enable readers to find theorems, definitions and biographical citations easily in the text.

There are a couple of suggested tracks through the text. One way we use the book, for a one-semester course, is to tackle Sections I through VI (except Chapters 14, 15, 21, 34, 35, and 36). If time permits, we either include some of those skipped chapters or Section VII.

For a more leisurely pace, covering less material, one might use Sections I, II, and Sections IV through VI, except Chapter 21. (This has the disadvantage of missing our first discussion of ideals in Chapters 11 and 12, but that just gives a bit less motivation to Section IV, which isn't a big problem.) One could also use this basic track with Section VII or selected topics from Section VIII.

A second semester could pick up wherever the first semester left off with the goal of completing Section IX on Galois Theory.

We assume that the students using the text have had the usual calculus sequence; this is mostly an assumption of a little mathematical maturity, since we only occasionally make any real use of the calculus. We do not assume any familiarity with linear algebra, although it would be helpful. We regularly use the multiplication of 2×2 matrices, mostly as an example of a non-commutative operation; we find that a short in-class discussion of this (perhaps supplemented with some of our exercises) is sufficient even for students who've never seen matrices before. We make heavy use of complex numbers in the text but do not assume any prior acquaintance with them; our introduction to them in Chapter 8 should be quite adequate.

This book is a substantial revision of the first edition, which appeared in 1995. We have carefully edited the entire manuscript and added many exercises throughout the book; we're grateful for the readers of the first edition who pointed out various typographical and mathematical errors. The major change in the book is the addition of Section IX about Galois theory, which follows naturally from our discussion of field extensions in Section VIII. As a requirement for Section IX, we have also added a new chapter on solvable groups (Chapter 36). We have deleted the last section of the earlier edition, which consisted of several unrelated applications, although we have included a revised version of the finite fields chapter (Chapter 46) from that edition. We have also added Nutshells for this edition.

Acknowledgments

Each of us has spent twenty-five years studying, reading, teaching and doing scholarship in abstract algebra, and thinking about the undergraduate course in this subject. Over the course of that time we have been influenced importantly by many textbooks on the subject. These books have inspired us to think about what topics to include, the order in which they might be presented, and techniques of proof and presentation. Both of us took courses from I. N. Herstein's text *Topics in Algebra.* We've had many occasions to look into the classic works by Van der Waerden and Jacobson. We've taught and learned from texts by such authors as Goldstein, Fraleigh, Burton and Gallian. Burton's text on elementary number theory and the Maxwells' little book on Galois theory have also been important. We offer our thanks here to those mentioned, and the many influences on our approach we cannot trace.

We owe considerable gratitude to the many people who have helped us in writing this book. We have discussed this book informally with many of our colleagues, immeasurably shaping the final product. Special mention should be made of Mike Westmoreland and Robin Wilson, who gave parts of the original manuscript a close reading. The charts and diagrams in Chapter 23 owe their origin to handouts prepared by C. C. Edwards. We're grateful to our respective institutions (The Colorado College and Denison University), which provided financial support to this project at various crucial stages.

Most of all, we thank Audrey and Robin, for putting up with the years of phone calls and e-mail that went into this book! Their moral support was (and is) indispensable.

Marlow Anderson
Mathematics Department
The Colorado College
Colorado Springs, CO 80903
(719) 389-6543
manderson@coloradocollege.edu

Todd Feil
Department of Mathematics and Computer Science
Denison University
Granville, OH 43023
(740) 587-6248
feil@denison.edu

I

Numbers, Polynomials, and Factoring

Chapter 1

The Natural Numbers

All mathematics begins with counting. This is the process of putting the set of objects to be counted in one-to-one correspondence with the first several **natural numbers** (or **counting numbers**):

$$1, 2, 3, 4, 5, \cdots .$$

We denote by $\mathbb{N}$ the infinite set consisting of all these numbers. Amazingly, despite the antiquity of its study, humankind has barely begun to understand the algebra of this set. This introduction is intended to provide you with a fund of examples and principles that we will generalize in later chapters.

1.1 Operations on the Natural Numbers

We encounter no trouble as long as we restrict ourselves to *adding* natural numbers, because more natural numbers result. Accordingly we say our set is *closed under addition.* However, consider what happens when we attempt to *subtract* a natural number a from b, or, equivalently, we seek a solution to the equation $a + x = b$ in the unknown x. We discover that our set of natural numbers is inadequate to the task. This naturally leads to the set of all integers, which we denote by $\mathbb{Z}$ (for 'Zahlen,' in German):

$$\cdots - 3, -2, -1, 0, 1, 2, 3, \cdots .$$

This is the smallest set of numbers containing $\mathbb{N}$ and closed under subtraction.

It is easy to make sense of *multiplication* in $\mathbb{N}$, by viewing it as repeated addition:

$$na = \underbrace{a + a + \cdots + a.}_{n \text{ times}}$$

This operation is easily extended to $\mathbb{Z}$ by using the sign conventions with which you are probably familiar. Why minus multiplied by minus needs to be plus is something you might reflect on now. We will return to this question in a more general context later.

We now have a whole new class of equations, many of which lack solutions: $ax = b$. This leads to *division*, and to the rational numbers $\mathbb{Q}$, which are precisely the quotients of one integer by another. The reason why we don't allow division by 0 is because if we let $a = 0$ and $b \neq 0$ in the equation above, we obtain $0 = 0x = b \neq 0$. Why $0x = 0$ is another question you might reflect on now – we will return to this later too.

But to address the algebra of $\mathbb{Q}$ takes us too far afield from our present subject. For the present we shall be more than satisfied in considering $\mathbb{Z}$ and its operations.

1.2 Well Ordering and Mathematical Induction

A fundamental property of $\mathbb{N}$ (which has a profound influence on the algebra of $\mathbb{Z}$) is that this set is **well ordered**, a property that we state formally as follows, and which we shall accept as an axiom about $\mathbb{N}$:

The Well-ordering Principle *Every non-empty subset of* $\mathbb{N}$ *has a least element.*

For any subset of $\mathbb{N}$ that we might specify by listing the elements, this seems obvious, but the principle applies even to sets that are more indirectly defined. For example, consider the set of all natural numbers expressible as $12x + 28y$, where x and y are allowed to be any integers. The extent of this set is not evident from the definition. Yet the Well-ordering Principle applies and thus there is a smallest natural number expressible in this way. We shall meet this example again, when we prove something called the GCD identity in the next chapter. (See also Exercise 1.9.)

Suppose we wish to apply the Well-ordering Principle to a particular subset X of $\mathbb{N}$. We may then consider a sequence of yes/no questions of the following form:

$$\text{is } 1 \in X?$$

$$\text{is } 2 \in X?$$
$$\vdots$$

Because X is non-empty, sooner or later one of these questions must be answered yes. The first such occurrence gives the least element of X. Of course, such questions might not be easily answerable in practice. But nevertheless, the Well-ordering Principle asserts the existence of this least element, without identifying it explicitly.

The Well-ordering Principle allows us to prove one of the most powerful techniques of proof that you will encounter in this book. (See Theorem 1.1 later in this chapter.) This is the *Principle of Mathematical Induction*:

Principle of Mathematical Induction *Suppose* X *is a subset of* $\mathbb{N}$ *that satisfies the following two criteria:*

1. $1 \in X$, *and*
2. *If* $k \in X$ *for all* $k < n$, *then* $n \in X$.

Then $X = \mathbb{N}$.

The Principle of Mathematical Induction is used to prove that certain sets X equal the entire set $\mathbb{N}$. In practice, the set X will usually be "the set of all natural numbers with property such-and-such." To apply it we must check two things:

1. the 'base case': that the least element of $\mathbb{N}$ belongs to X, and
2. the 'bootstrap': a general statement which asserts that a natural number belongs to X whenever all its predecessors do.

You should find the Principle of Mathematical Induction plausible because successively applying the bootstrap allows you to conclude that

$$2 \in X,\ 3 \in X,\ 4 \in X,\ \cdots.$$

When checking the 'bootstrap', we assume that all predecessors of n belong to X and must infer that n belongs to X. In practice we often need only that certain predecessors of n belong to X. For instance, many times we will need only that $n - 1$ belongs to X. Indeed, the form of induction you have used before probably assumed only that $n - 1$ was in X, instead of all $k < n$. It turns out that the version you learned before and the version we will be using are equivalent, although they don't appear to be at first glance. We will find the version given above of more use. (See Exercise 1.17.)

Before proving the Principle of Mathematical Induction itself, let us look at some simple examples of its use.

Example 1.1

A finite set with n elements has exactly 2^n subsets.

Proof by Induction: Let X be the set of those positive integers for which this is true. We first check that $1 \in X$. But a set with exactly one element has two subsets, namely, the empty set $\emptyset$ and the set itself. This is $2 = 2^1$ subsets, as required.

Now suppose that $n > 1$, and $k \in X$ for all $k < n$. We must prove that $n \in X$. Suppose then that S is a set with n elements; we must show that S has 2^n subsets. Because S has at least one element, choose one of them and call it s. Now every subset of S either contains s or it doesn't. Those subsets that don't contain s are precisely the subsets of $S\backslash\{s\} = \{x \in S : x \neq s\}$. But this latter set has $n-1$ elements, and so by our assumption that $n-1 \in X$ we know that $S\backslash\{s\}$ has 2^{n-1} subsets. Now those subsets of S that *do* contain s are of the form $A \cup \{s\}$, where A is a subset of $S\backslash\{s\}$. There are also 2^{n-1} of these subsets. Thus, there are $2^{n-1} + 2^{n-1} = 2^n$ subsets of S altogether. In other words, $n \in X$. Thus, by the Principle of Mathematical Induction, any finite set with n elements has exactly 2^n subsets. □

Notice that this formula for counting subsets also works for a set with zero elements because the empty set has exactly one subset (namely, itself). We could have easily incorporated this fact into the proof above by starting at $n = 0$ instead. This amounts to saying that the set $\{0, 1, 2, \cdots\}$ is well ordered too. In the future we will feel free to start an induction proof at any convenient point, whether that happens to be $n = 1$ or $n = 0$. (We can also start induction at, say, $n = 2$, but in such a case remember that we would then have proved only that $X = \mathbb{N}\backslash\{1\}$.)

Example 1.2

The sum of the first n odd integers is n^2. That is,

$$1 + 3 + 5 + \cdots + (2n-1) = n^2, \text{ for } n \geq 1.$$

Proof by Induction: In this proof we proceed slightly less formally than before and suppress explicit mention of the set X.

▷ **Quick Exercise.** What is the set X in this proof? ◁Because $2 \cdot 1 - 1 = 1^2$, our formula certainly holds for $n = 1$. We now assume that the formula holds for $k < n$ and show that it holds for n. But then, putting $k = n - 1$, we have

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

Thus,

$$1 + 3 + 5 + \cdots + (2(n-1) - 1) + (2n-1) = \\ (n-1)^2 + (2n-1) = n^2,$$

which shows that the formula holds for n. Thus, by the Principle of Mathematical Induction, the formula holds for *all* n. □

Students new to mathematical induction often feel that in verifying (2) they are assuming exactly what they are required to prove. This feeling arises from a misunderstanding of that fact that (2) is an *implication*: that is, a statement of the form $p \Rightarrow q$. To prove such a statement we must assume p, and then derive q. Indeed, assuming that k is in X for all $k < n$ is often referred to as **the induction hypothesis.**

Mention should also be made of the fact that mathematical induction is a deductive method of proof and so should not be confused with the notion of inductive reasoning discussed by philosophers. The latter involves inferring likely general principles from particular cases. This sort of reasoning has an important role in mathematics, and we hope you will apply it to make conjectures regarding the more general principles that lie behind many of the particular examples which we will discuss. However, for a mathematician an inductive inference of this sort does not end the story. What is next required is a deductive proof that the conjecture (which might have been verified in particular instances) is always true.

1.3 The Fibonacci Sequence

To provide us with another example of proof by induction, we consider a famous sequence of integers, called the **Fibonacci sequence** in honor of the medieval mathematician who first described it. The first several terms are

$$1, 1, 2, 3, 5, 8, 13, \cdots.$$

You might have already detected the pattern: a typical element of the sequence is the sum of its two immediate predecessors. This means that we can *inductively* define the sequence by setting

$$\begin{aligned} a_1 &= 1, \\ a_2 &= 1, \text{ and} \\ a_{n+2} &= a_{n+1} + a_n, \text{ for } n \geq 1. \end{aligned}$$

This sort of inductive or *recursive* definition of a sequence is often very useful. However, it would still be desirable to have an explicit formula for a_n in terms of n. It turns out that the following surprising formula does the job:

$$a_n = \frac{(1+\sqrt{5})^n - (1-\sqrt{5})^n}{2^n\sqrt{5}}.$$

At first (or even second) glance, it does not even seem clear that this formula gives integer values, much less the particular integers that make up the Fibonacci sequence. You will prove this formula in Exercise 1.13. The proof uses the Principle of Mathematical Induction, because the Fibonacci sequence is defined in terms of its two immediate predecessors. We now prove another simpler fact about the Fibonacci sequence:

Example 1.3

$a_{n+1} \leq 2a_n$, for all $n \geq 1$.

Proof by Induction: This is trivially true when $n = 1$. In the argument which follows we rely on two successive true instances of our formula—as might be expected because the Fibonacci sequence is defined in terms of two successive terms. Consequently, you should check that the inequality holds when $n = 2$.

▷ **Quick Exercise.** Verify that the inequality $a_{n+1} \leq 2a_n$ holds for $n = 1$ and $n = 2$. ◁

We now assume that $a_{k+1} \leq 2a_k$ for all $k < n$, where $n > 2$. We must show that this inequality holds for $k = n$. Now

$$a_{n+1} = a_n + a_{n-1} \leq 2a_{n-1} + 2a_{n-2},$$

where we have applied the induction hypothesis for both $k = n-1$ and $k = n-2$. But because $a_{n-1} + a_{n-2} = a_n$, we have $a_{n+1} \leq 2a_n$, as required. □

1.4 Well Ordering Implies Mathematical Induction

We now prove the Principle of Mathematical Induction, using the Well-ordering Principle.

Theorem 1.1 *The Well-ordering Principle implies the Principle of Mathematical Induction.*

Proof: Suppose that X is a subset of $\mathbb{N}$ satisfying both (1) and (2). Our strategy for showing that $X = \mathbb{N}$ is 'reductio ad absurdum' (or 'proof by contradiction'): we assume the contrary and derive a contradiction.

In this case we assume that X is a proper subset of $\mathbb{N}$, and so $Y = \mathbb{N}\backslash X$ is a non-empty subset of $\mathbb{N}$. By the Well-ordering Principle, Y possesses a least element m. Clearly $m \neq 1$ by (1). All natural numbers $k < m$ belong to X because m is the *least* element of Y. However, by (2) we conclude that $m \in X$. But now we have concluded that $m \in X$ and $m \notin X$; this is clearly a contradiction. Our assumption that X is a proper subset of $\mathbb{N}$ must have been false. Hence, $X = \mathbb{N}$. □

The converse of this theorem is also true (see Exercise 1.16).

1.5 The Axiomatic Method

Our careful proof of the Principle of Mathematical Induction from the Well-ordering Principle is part of a general program we are beginning in this chapter. We wish eventually to base our analysis of the arithmetic of the integers on as few assumptions as possible. This will be an example of the *axiomatic method* in mathematics. By making our assumptions clear and our proofs careful, we will be able to accept with confidence the truth of statements about the integers which we will prove later, even if the statements themselves are not obviously true. We eventually will also apply the axiomatic method to many algebraic systems other than the integers.

The first extended example of an axiomatic approach to mathematics appears in *The Elements* of Euclid, who was a Greek mathematician living circa 300 B.C. In his book he developed much of ordinary plane geometry by means of a careful logical string of theorems, based on only five axioms and some definitions. The logical structure of Euclid's book is a model of mathematical economy and elegance. So much mathematics is inferred from so few underlying assumptions!

Note of course that we must accept *some* statements without proof (and we call these statements axioms)—for otherwise we'd be led into circular reasoning or an infinite regress.

One cost of the axiomatic method is that we must sometimes prove a statement that already seems 'obvious'. But if we are to be true to the axiomatic method, a statement we believe to be true must either be proved, or else added to our list of axioms. And for reasons of logical economy and elegance, we wish to rely on as few axioms as possible.

Unfortunately, we are not yet in a position to proceed in a completely axiomatic way. We shall accept the Well-ordering Principle as an axiom about the natural numbers. But in addition, we shall accept as given facts your understanding of the elementary arithmetic in $\mathbb{Z}$: that is, addition, subtraction and multiplication. In Chapter 6, we will finally be able to enumerate carefully the abstract properties which make this arithmetic work. The role of $\mathbb{Z}$ as a familiar, motivating example will be crucial.

The status of division in the integers is quite different. It is considerably trickier (because it is not always possible). We will examine this carefully in the next chapter.

Chapter Summary

In this chapter we introduced the natural numbers $\mathbb{N}$ and emphasized the following facts about this set:

- $\mathbb{N}$ is closed under addition;
- Multiplication in $\mathbb{N}$ can be defined in terms of addition, and under this definition $\mathbb{N}$ is closed under multiplication;

- The *Well-ordering Principle* holds for $\mathbb{N}$.

We then used the Well-ordering Principle to prove the *Principle of Mathematical Induction* and provided examples of its use.

We also introduced the set $\mathbb{Z}$ of all integers, as the smallest set of numbers containing $\mathbb{N}$ that is closed under subtraction.

Warm-up Exercises

a. Explain the arithmetic advantages of $\mathbb{Z}$, as compared with $\mathbb{N}$. How about $\mathbb{Q}$, as compared with $\mathbb{Z}$?

b. Why isn't $\mathbb{Z}$ well ordered? Why isn't $\mathbb{Q}$ well ordered? Why isn't the set of all rational numbers x with $0 \leq x \leq 1$ well ordered?

c. Suppose we have an infinite row of dominoes, set up on end. What sort of induction argument would convince us that knocking down the first domino will knock them all down?

d. Explain why any finite subset of $\mathbb{Q}$ is well ordered.

Exercises

1. Prove using mathematical induction that for all positive integers n,
$$1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}.$$

2. Prove using mathematical induction that for all positive integers n,
$$1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(2n+1)(n+1)}{6}.$$

3. You probably recall from your previous mathematical work the *triangle inequality*: for any real numbers x and y,
$$|x + y| \leq |x| + |y|.$$
Accept this as given (or see a calculus text to recall how it is proved). Generalize the triangle inequality, by proving that
$$|x_1 + x_2 + \cdots + x_n| \leq |x_1| + |x_2| + \cdots + |x_n|,$$
for any positive integer n.

4. Given a positive integer n, recall that $n! = 1 \cdot 2 \cdot 3 \cdots n$ (this is read as n *factorial*). Provide an inductive definition for $n!$. (It is customary to actually start this definition at $n = 0$, setting $0! = 1$.)

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

6. Prove that for all positive integers n,

$$1^3 + 2^3 + \cdots + n^3 = \left(\frac{n(n+1)}{2}\right)^2.$$

7. Prove the familiar *geometric progression* formula. Namely, suppose that a and r are real numbers with $r \neq 1$. Then show that

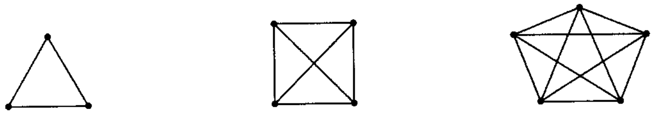
$$a + ar + ar^2 + \cdots + ar^{n-1} = \frac{a - ar^n}{1-r}.$$

8. Prove that for all positive integers n,

$$\frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \cdots + \frac{1}{n(n+1)} = \frac{n}{n+1}.$$

9. By trial and error, try to find the smallest positive integer expressible as $12x + 28y$, where x and y are allowed to be any integers.

10. A **complete graph** is a collection of n points, each of which is connected to each other point. The complete graphs on 3, 4, and 5 points are illustrated below:



Use mathematical induction to prove that the complete graph on n points has exactly $n(n-1)/2$ lines.

11. Consider the sequence $\{a_n\}$ defined inductively as follows:

$$a_1 = a_2 = 1, \quad a_{n+2} = 2a_{n+1} - a_n.$$

Use mathematical induction to prove that $a_n = 1$, for all natural numbers n.

12. Consider the sequence $\{a_n\}$ defined inductively as follows:

$$a_1 = 5, \quad a_2 = 7, \quad a_{n+2} = 3a_{n+1} - 2a_n.$$

Use mathematical induction to prove that $a_n = 3 + 2^n$, for all natural numbers n.

13. Consider the Fibonacci sequence $\{a_n\}$.

(a) Prove that $a_{n+1}a_{n-1} = (a_n)^2 + (-1)^n$.

(b) Use mathematical induction to prove that

$$a_n = \frac{(1+\sqrt{5})^n - (1-\sqrt{5})^n}{2^n\sqrt{5}}.$$

14. In this problem you will prove some results about the **binomial coefficients**, using induction. Recall that

$$\binom{n}{k} = \frac{n!}{(n-k)!k!},$$

where n is a positive integer, and $0 \leq k \leq n$.

(a) Prove that

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1},$$

$n \geq 2$ and $k < n$. *Hint:* You do not need induction to prove this. Bear in mind that $0! = 1$.

(b) Verify that $\binom{n}{0} = 1$ and $\binom{n}{n} = 1$. Use these facts, together with part a, to prove by induction on n that $\binom{n}{k}$ is an integer, for all k with $0 \leq k \leq n$. (*Note:* You may have encountered $\binom{n}{k}$ as the count of the number of k element subsets of a set of n objects; it follows from this that $\binom{n}{k}$ is an integer. What we are asking for here is an inductive proof based on algebra.)

(c) Use part a and induction to prove the **Binomial Theorem**: For non-negative n and variables x, y,

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

15. Criticize the following 'proof' showing that all cows are the same color.

    It suffices to show that any herd of n cows has the same color. If the herd has but one cow, then trivially all the cows in the herd have the same color. Now suppose that we have a herd of n cows and $n > 1$. Pick out a cow and remove it from the herd, leaving $n-1$ cows; by the induction hypothesis these cows all have the same color. Now put the cow back and remove another cow. (We can do so because $n > 1$.) The remaining $n-1$ again must all be the same color. Hence, the first cow selected and the second cow selected have the same color as those not selected, and so the entire herd of n cows has the same color.

16. Prove the converse of Theorem 1.1; that is, prove that the Principle of Mathematical Induction implies the Well-ordering Principle. (This shows that these two principles are logically equivalent, and so from an axiomatic point of view it doesn't matter which we assume is an axiom for the natural numbers.)

17. The *Strong* Principle of Mathematical Induction asserts the following. Suppose that X is a subset of $\mathbb{N}$ that satisfies the following two criteria:

    (a) $1 \in X$, and

    (b) If $n > 1$ and $n-1 \in X$, then $n \in X$.

    Then $X = \mathbb{N}$. Prove that the Principle of Mathematical Induction holds if and only if the Strong Principle of Mathematical Induction does.

Chapter 2

The Integers

In this chapter we analyze how multiplication works in the integers $\mathbb{Z}$, and in particular when division is possible. This is more interesting than asking how multiplication works in the rational numbers $\mathbb{Q}$, where division is always possible (except for division by zero).

We all learned at a very young age that we can always divide one integer by another non-zero integer, as long as we allow for a remainder. For example, $326 \div 21$ gives quotient 15 with remainder 11. The actual computation used to produce this result is our usual long division. Note that the division process halts when we arrive at a number less than the divisor. In this case 11 is less than 21, and so our division process stops. We can record the result of this calculation succinctly as

$$326 = (21)(15) + 11, \text{ where } 0 \leq 11 < 21.$$

2.1 The Division Theorem

The following important theorem describes this situation formally. This is the first of many examples in this book of an *existence and uniqueness theorem*: We assert that something exists, and that there is only one such. Both assertions must be proved. We will use induction for the existence proof.

Theorem 2.1 Division Theorem for $\mathbb{Z}$ *Let $a, b \in \mathbb{Z}$, with $a \neq 0$. Then there exist unique integers q and r (called the* quotient *and* remainder*, respectively), with $0 \leq r < |a|$, such that $b = aq + r$.*

Proof: We first prove the theorem in case $a > 0$ and $b \geq 0$. To show the existence of q and r in this case, we use induction on b.

We must first establish the base case for the induction. You might expect us to check that the theorem holds in case $b = 0$ (the smallest possible value for b). But actually, we can establish the theorem for all b where $b < a$; for in this case the quotient is 0 and the remainder is b. That is, $b = a \cdot 0 + b$.

We may now assume that $b \geq a$. Our induction hypothesis is that there exist a quotient and remainder whenever we attempt to divide an integer $c < b$ by a. So let $c = b - a$. Since $c < b$ we have by the induction hypothesis that $c = aq' + r$, where $0 \leq r < a$. But then

$$b = aq' + r + a = a(q' + 1) + r, \text{ where } 0 \leq r < a.$$

We therefore have a quotient $q = q' + 1$ and a remainder r.

We now consider the general case, where b is any integer, and a is any non-zero integer. We apply what we have already proved to the integers $|b|$ and $|a|$ to obtain unique integers q' and r' so that $|b| = q'|a| + r'$, with $r < |a|$. We now obtain the quotient and remainders required, depending on the signs of a and b, in the following three cases:

Case (i): Suppose that $a < 0$ and $b \leq 0$. Then let $q = q' + 1$ and $r = -a - r'$. Note first that $0 \leq r < |a|$. Now

$$\begin{aligned} aq + r &= a(q' + 1) + -a - r' = aq' + a - a - r' \\ &= aq' - r' = -(|a|q' + r') = -|b| = b, \end{aligned}$$

as required.

You can now check the remaining two cases:

Case (ii): If $a < 0$ and $b \geq 0$, then let $q = -q'$ and $r = r'$.

Case (iii): If $a > 0$ and $b \leq 0$, then let $q = -q' - 1$ and $r = a - r'$.

▷ **Quick Exercise.** Verify that the quotients and remainders specified in Cases (ii) and (iii) actually work. ◁

Now we prove the uniqueness of q and r. Our strategy is to assume that we have two potentially different quotient-remainder pairs, and then show that the different pairs are actually the same. So, suppose that $b = aq + r = aq' + r'$, where $0 \leq r < |a|$ and $0 \leq r' < |a|$. We hope that $q = q'$ and $r = r'$.

Since $aq + r = aq' + r'$, we have that $a(q - q') = r' - r$. Now $|r' - r| < |a|$, and so $|a||q - q'| = |r' - r| < |a|$. Hence, $|q - q'| < 1$. Thus, $q - q'$ is an integer whose absolute value is less than 1, and so $q - q' = 0$. That is, $q = q'$. But then $r' - r = a \cdot 0 = 0$ and so $r' = r$, proving uniqueness. □

You should exercise some care in applying the Division Theorem with negative integers. The fact that the remainder must be positive leads to some answers that may be surprising.

Example 2.1

For example, while 326 divided by 21 gives quotient 15 and remainder 11, -326 divided by 21 gives quotient -16 and remainder 10, and -326 divided by -21 gives quotient 16 and remainder 10.

We say an integer a **divides** an integer b if $b = aq$ for some integer q. In this case, we say a is a **factor** of b, and write $a|b$. In the context of the Division Theorem, $a|b$ means that the remainder obtained is 0.

Example 2.2

Thus, $-6|126$, because $126 = (-6)(-21)$. Note that *any* integer divides 0, because $0 = (a)(0)$.

2.2 The Greatest Common Divisor

In practice, it may be *very* difficult to find the factors of a given integer, if it is large. However, it turns out to be relatively easy to determine whether two given integers have a common factor. To understand this, we must introduce the notion of greatest common divisor: Given two integers a and b (not both zero), then the integer d is the **greatest common divisor** (gcd) of a and b if d divides both a and b, and it is the largest positive integer that does. We will often write $\gcd(a, b) = d$ to express this relationship.

For example $6 = \gcd(42, -30)$, as you can check directly by computing all possible common divisors, and picking out the largest one. Because all integers divide 0, we have not allowed ourselves to consider the meaningless expression $\gcd(0, 0)$. However, if $a \neq 0$, it does make sense to consider $\gcd(a, 0)$.

▷ **Quick Exercise.** Argue that for all $a \neq 0$, $\gcd(a, 0) = |a|$. ◁

But why should an arbitrary pair of integers (not both zero) have a gcd? That is, does the definition we have of gcd really make sense?

Note that if $c > 0$ and $c|a$ and $c|b$, then $c \leq |a|$ and $c \leq |b|$. This means that there are only finitely many positive integers that could possibly be the gcd of a and b, and because 1 *does* divide both a and b, a and b do have at least one common divisor. This means that the gcd of any pair of integers exists (and is unique).

To actually determine $\gcd(a, b)$ we would rather not check all the possibilities less than $|a|$ and $|b|$. Fortunately, we don't have to, because there is an algorithm that determines the gcd quite efficiently. This first appears as Proposition 2 of Book 7 of Euclid's *Elements* and depends on repeated applications of the Division Theorem; we call it **Euclid's Algorithm**. We present the algorithm below but first need the following lemma:

Lemma 2.2 *Suppose that a, b, q, r are integers and $b = aq + r$. Then* $\gcd(b, a) = \gcd(a, r)$.

Proof: To show this, we need only check that every common divisor of b and a is a common divisor of a and r, and vice versa, for then the greatest element of this set will be both $\gcd(b, a)$ and $\gcd(a, r)$. But if $d|a$ and $d|b$ then $d|r$, because $r = b - aq$. Conversely, if $d|a$ and $d|r$, then $d|b$, because $b = aq + r$. □

We will now give an example of Euclid's Algorithm, before describing it formally below. This example should make the role of the lemma clear.

Example 2.3

Suppose we wish to determine the gcd of 285 and 255. If we successively apply the Division Theorem until we reach a remainder of 0, we obtain the following:

$$\begin{aligned} 285 &= 255 \cdot 1 + 30 \\ 255 &= 30 \cdot 8 + 15 \\ 30 &= 15 \cdot 2 + 0 \end{aligned}$$

By the lemma we have that

$$\gcd(285, 255) = \gcd(255, 30) = \gcd(30, 15) = \gcd(15, 0),$$

and by the Quick Exercise above, this last is equal to 15.

Explicitly, to compute the gcd of b and a using Euclid's Algorithm, where $|b| \geq |a|$, we proceed inductively as follows. First, set $b_0 = b, a_0 = a$, and let q_0 and r_0 be the quotient and remainder that result when b_0 is divided by a_0. Then, for $n \geq 0$, let $b_n = a_{n-1}$ and $a_n = r_{n-1}$, and let q_n and r_n be the quotient and remainder that result when b_n is divided by a_n. We then continue until $r_n = 0$, and claim that $r_{n-1} = \gcd(b, a)$. Setting aside for a moment the important question of why r_n need ever reach 0, the general form of the algorithm looks like this:

$$\begin{aligned} b_0 &= a_0 q_0 + r_0 \\ b_1 &= a_1 q_1 + r_1 \\ &\vdots \\ b_{n-1} &= a_{n-1} q_{n-1} + r_{n-1} \\ b_n &= a_n q_n + 0 \end{aligned}$$

We can now formally show that Euclid's Algorithm does indeed compute $\gcd(b, a)$:

Theorem 2.3 *Euclid's Algorithm computes* $\gcd(b, a)$.

Proof: Using the general form for Euclid's Algorithm above, the lemma says that

$$\begin{gathered} \gcd(b, a) = \gcd(b_0, a_0) = \\ \gcd(a_0, r_0) = \gcd(b_1, a_1) = \\ \gcd(a_1, r_1) = \cdots = \\ \gcd(a_{n-1}, r_{n-1}) = \gcd(b_n, a_n) = \\ \gcd(a_n, 0) = a_n = r_{n-1}. \end{gathered}$$

It remains only to understand why this algorithm halts. That is, why must some remainder $r_n = 0$? But $a_{i+1} = r_i < |a_i| = r_{i-1}$. Thus, the r_i's form a strictly decreasing sequence of non-negative integers. By the Well-ordering Principle, such a sequence is necessarily finite. This means that $r_n = 0$ for some n. □

We have thus proved that after finitely many steps Euclid's Algorithm will produce the gcd of any pair of integers. In fact, this algorithm reaches the gcd quite rapidly, in a sense we cannot make precise here. It is certainly much more rapid than considering all possible common factors case by case.

2.3 The GCD Identity

In the equations describing Euclid's algorithm above, we can start with the bottom equation $b_{n-1} = a_{n-1}q_{n-1} + r_{n-1}$ and solve this for $\gcd(b,a) = r_{n-1}$ in terms of b_{n-1} and a_{n-1}. Plugging this into the previous equation, we can express $\gcd(b,a)$ in terms of b_{n-2} and a_{n-2}. Repeating this process, we can eventually obtain an equation of the form $\gcd(b,a) = ax + by$, where x and y are integers. That is, $\gcd(b,a)$ can be expressed as a **linear combination** of a and b. (Here the coefficients of the linear combination are integers x and y; we will use this terminology in a more general context later.)

Example 2.4

In the case of 285 and 255 we have the following:

$$\begin{aligned} 15 &= 255 - 30(8) \\ &= 255 - (285 - 255 \cdot 1)(8) \\ &= 255(9) + 285(-8) \end{aligned}$$

This important observation we state formally:

Theorem 2.4 The GCD identity for integers *Given integers a and b (not both zero), there exist integers x and y for which $\gcd(b,a) = ax + by$.*

▷ **Quick Exercise.** Try using Euclid's Algorithm to compute

$$\gcd(120, 27),$$

and then express this gcd as a linear combination of 120 and 27. ◁

What we have described above is a *constructive* (or *algorithmic*) approach to expressing the gcd of two integers as a linear combination of them. We will now describe an alternative proof of the GCD identity, which shows the existence of the linear combination, without giving us an explicit recipe for finding it. This sort of proof is inherently more abstract than the constructive proof, but we are able to conclude a bit more about the gcd from it. We will also find it valuable when we generalize these notions to more general algebraic structures than the integers.

Existential proof of the GCD identity: We begin by considering the set of all linear combinations of the integers a and b. That is, consider the set

$$S = \{ax + by : x, y \in \mathbb{Z}\}.$$

This is obviously an infinite subset of $\mathbb{Z}$. If the GCD identity is to be true, then the gcd of a and b belongs to this set. But which element is it? By the Well-ordering Principle, S contains a unique smallest positive element which we will call d.

▷ **Quick Exercise.** To apply the Well-ordering Principle, the set S must contain at least one positive element. Why is this true? ◁

Since $d \in S$, we can write it as $d = ax_0 + by_0$, for some particular integers x_0 and y_0. We claim that d is the gcd of a and b.

To prove this, we must first check that d is a common divisor, that is, that it divides both a and b. If we apply the Division Theorem 2.1 to d and a, we obtain $a = dq + r$. We must show that r is zero. But

$$r = a - dq = a - (ax_0 + by_0)q = a(1 - qx_0) + b(-qy_0),$$

and so $r \in S$. Because $0 \leq r < d$, and d is the smallest positive element of S, $r = 0$, as required. A similar argument shows that $d|b$ too.

Now suppose that $c > 0$ and $c|a$ and $c|b$. Then $a = nc$ and $b = mc$. But then $ax + by = ncx + mcy = c(nx + my)$, and so c divides any linear combination of a and b. Thus, c divides d. But because c and d are both positive, $c \leq d$. That is, d is the gcd of a and b. □

Example 2.5

Thus, the gcd of 12 and 28 is 4, because $4 = 12 \cdot (-2) + 28(1)$ is the smallest positive integer expressible as a linear combination of 12 and 28. We referred to this example when introducing the Well-ordering Principle in the previous chapter; see Exercise 1.9.

We conclude from this proof the following:

Corollary 2.5 *The gcd of two integers (not both zero) is the least positive linear combination of them.*

2.4 The Fundamental Theorem of Arithmetic

We are now ready to tackle the main business of this chapter: proving that every non-zero integer can be factored uniquely as a product of integers that cannot be further factored. This theorem's importance is emphasized by the fact that it is usually known as the *Fundamental Theorem of Arithmetic.* It first appears (in essence) as Proposition 14 of Book 9 in Euclid's *Elements.*

We first need a formal definition. An integer p (other than ± 1) is **irreducible** if whenever $p = ab$, then a or b is ± 1. We are thus allowing the always possible 'trivial' factorizations $p = (1)(p) = (-1)(-p)$. We are not allowing ± 1 to be irreducible because it would unnecessarily complicate the formal statement of the Fundamental Theorem of Arithmetic that we make below. Because $0 = (a)(0)$ for any integer a, it is clear that 0 is not irreducible. Finally, notice that if p is irreducible, then so is $-p$. This means that in the arguments that follow we can often assume that p is positive.

The positive integers that are irreducible form a familiar list:

$$2, \quad 3, \quad 5, \quad 7, \quad 11, \quad 13, \quad 17, \cdots.$$

You are undoubtedly familiar with these numbers, under the name *prime* integers, and it may seem perverse for us to call them 'irreducible'. But this temporary perversity now will allow us to be consistent with the more general terminology we'll use later.

We reserve the term 'prime' for another definition: an integer p (other than 0 and ± 1) is **prime** if, whenever p divides ab, then either p divides a or p divides b. (Notice that when we say 'or' here, we mean one or the other or both. This is what logicians call the *inclusive* 'or', and is the sense of this word that we will always use.)

Example 2.6

For instance, we know that 2 is a prime integer. For if $2|ab$, then ab is even. But a product of integers is even exactly if at least one of the factors is even, and so $2|a$ or $2|b$.

The prime property generalizes to more than two factors:

Theorem 2.6 *If p is prime and $p|a_1 a_2 \cdots a_n$, then $p|a_i$ for some i.*

Proof: This is Exercise 2.5. Prove it using induction on n. □

For the integers, the ideas of primeness and irreducibility coincide. This is the content of the next theorem.

Theorem 2.7 *An integer is prime if and only if it is irreducible.*

Proof: This theorem asserts that the concepts of primeness and irreducibility are equivalent for integers. This amounts to two implications which must be proved: primeness implies irreducibility, and the converse statement that irreducibility implies primeness.

Suppose first that p is prime. To show that it is irreducible, suppose that p has been factored: $p = ab$. Then $p|ab$, and so (without loss of generality) $p|a$. Thus, $a = px$, and so $p = pxb$. But then $1 = xb$, and so both x and b can only be ± 1. This shows that the factorization $p = ab$ is trivial, as required.

Conversely, suppose that p is irreducible, and $p|ab$. We will suppose also that p does not divide a. We thus must prove that p does divide b. Suppose that d is a positive common divisor of p and a. Then, because p is irreducible, d must be p or 1. Because p doesn't divide a, it must be that $\gcd(p, a) = 1$. So by the GCD identity 2.4, there exist x and y with $1 = ax + py$. But then $b = abx + bpy$, and because p clearly divides both abx and bpy it thus divides b, as required. □

Again, it may seem strange to have both of the terms 'prime' and 'irreducible', because for $\mathbb{Z}$ we have proved that they amount to the same thing. But we will later discover more general contexts where these concepts are distinct.

We now prove half of the Fundamental Theorem of Arithmetic:

Theorem 2.8 *Every non-zero integer (other than ± 1) is either irreducible or a product of irreducibles.*

Proof: Let n be an integer other than ± 1, which we may as well suppose is positive. We proceed by induction on n. We know that $n \neq 1$, and if $n = 2$, then it is irreducible. Now suppose the theorem holds true for all $m < n$. If n is irreducible already, we are done. If not, then $n = bc$, where, without loss of generality, both factors are positive and greater than 1. But then by the induction hypothesis both b and c can be factored as a product of irreducibles, and thus so can their product n. □

For example, we can factor the integer 120 as $2 \cdot 2 \cdot 2 \cdot 3 \cdot 5$. Now $(-5) \cdot 2 \cdot (-2) \cdot 3 \cdot 2$ is a distinct factorization of 120 into irreducibles, but it is clearly essentially the same, where we disregard order and factors of -1. The uniqueness half of the Fundamental Theorem of Arithmetic asserts that all distinct factorizations into irreducibles of a given integer are essentially the same, in this sense. To prove this we use the fact that irreducible integers are prime.

Theorem 2.9 Unique Factorization Theorem for Integers
If an integer $x = a_1a_2 \cdots a_n = b_1b_2 \cdots b_m$ where the a_i and b_j are all irreducible, then $n = m$ and the b_j may be rearranged so that $a_i = \pm b_i$, for $i = 1, 2, \cdots, n$.

Proof: We use induction on n. If $n = 1$, the theorem follows easily.

▷ **Quick Exercise.** Check this. ◁

So we assume $n > 1$. By the primeness property of the irreducible a_1, a_1 divides one of the b_j. By renumbering the b_j if necessary, we may assume a_1 divides b_1. So, because b_1 is irreducible, $a_1 = \pm b_1$. Therefore, by dividing both sides by a_1, we have

$$a_2a_3 \cdots a_n = \pm b_2 \cdots b_m.$$

(Because b_2 is irreducible, so is $-b_2$, and we consider $\pm b_2$ as an irreducible factor.) We now have two factorizations into irreducibles, and the number of a_i factors is $n - 1$. So by the induction hypothesis $n - 1 = m - 1$, and by renumbering the b_j as necessary, $a_i = \pm b_i$ for $i = 1, 2, \cdots, n$. This proves the theorem. □

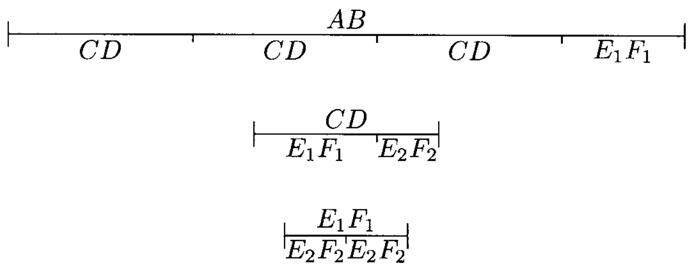
2.5 A Geometric Interpretation

As we have indicated already, both Euclid's Algorithm and the Fundamental Theorem of Arithmetic have their origins in the work of the

Greek geometer Euclid. It is important to note that Euclid viewed both of these theorems as *geometric* statements about line segments.

To understand this requires a definition: A line segment AB **measures** a line segment CD, if there is a positive integer n, so that we can use a compass to lay exactly n copies of AB next to one another, to make up the segment CD. In modern language, we would say that the length of CD is n times that of AB, but this notion of *length* was foreign to Euclid.

Euclid's Algorithm can now be viewed in the following geometric way: Given two line segments AB and CD, can we find a line segment EF, which measures both AB and CD? In the diagram below, we see by example how Euclid's Algorithm accomplishes this.



We can recapitulate this geometry in algebraic form, which makes the connection with Euclid's Algorithm clear:

$$AB = 3CD + E_1F_1,$$
$$CD = 1E_1F_1 + E_2F_2,$$
$$E_1F_1 = 2E_2F_2.$$

Thus, AB and CD are both measured by E_2F_2. In fact, $CD = 3E_2F_2$ and $AB = 11E_2F_2$. In modern language, we would say that the ratio of the length of AB to the length of CD is 11/3. Note that in this context Euclid's Algorithm halts only in case this ratio of lengths is a *rational number* (that is, a ratio of integers). In fact, it is possible to prove that the ratio of the diagonal of a square to one of the sides is irrational, by showing that in this case Euclid's Algorithm never halts (see Exercises 2.14 and 2.15).

Euclid's proposition that is closest to the Fundamental Theorem of Arithmetic says that *if a number be the least that is measured by prime numbers, it will not be measured by any other prime number except those originally measuring it.* This seems much more obscure than our statement, in part because of the geometric language that Euclid uses. Euclid's proposition does assert that if a number is a product of certain primes, it is then not divisible by any other prime, which certainly follows from the Fundamental Theorem, and is indeed the most important idea contained in our theorem. However, Euclid lacked both our flexible notation, and the precisely formulated tool of Mathematical Induction, to make his statement clearer and more modern. It wasn't until the eighteenth century, with such mathematicians as Euler and Legendre, that a modern statement was possible, and a careful proof in modern form did not appear until the work of Gauss, in the early 19th century.

Chapter Summary

In this chapter we examined division and factorization in $\mathbb{Z}$. We proved the *Division Theorem* by induction and then used it to obtain *Euclid's Algorithm* and the *GCD identity.* We defined the notions of *primeness* and *irreducibility* and showed that they are equivalent. We then proved the *Fundamental Theorem of Arithmetic*, which asserts that all integers other than $0, 1, -1$ are irreducible or can be factored uniquely into a product of irreducibles.

Warm-up Exercises

a. Find the quotient and remainder, as guaranteed by the Division Theorem 2.1, for 13 and -120, -13 and 120, and -13 and -120.

b. What are the possible remainders when you divide by 3, using the Division Theorem 2.1? Choose one such remainder, and make a list describing all integers that give this remainder, when dividing by 3.

c. What are the possible answers to $\gcd(a, p)$, where p is prime, and a is an arbitrary integer?

d. Let m be a fixed integer. Describe succinctly the integers a where

$$\gcd(a, m) = m.$$

e. Give the prime factorizations of 92, 100, 101, 102, 502, and 1002.

f. Suppose that we have two line segments. One has length 11/6 units, and the other has length 29/15. What length is the longest segment that measures both?

g. We proved the GCD identity 2.4 twice. Explain the different approaches of the two proofs to finding the appropriate linear combination. Which is easier to describe in words? Which is computationally more practical?

Exercises

1. (a) Find the greatest common divisor of 34 and 21, using Euclid's Algorithm. Then express this gcd as a linear combination of 34 and 21.

   (b) Now do the same for 2424 and 772.

2. Prove that $\gcd(a, b)$ divides $a - b$. This sometimes provides a short cut in finding gcds. Use this to find $\gcd(1962, 1965)$. Now find $\gcd(1961, 1965)$.

3. Prove that the set of all linear combinations of a and b are precisely the multiples of $\gcd(a, b)$.

4. Two numbers are said to be **relatively prime** if their gcd is 1. Prove that a and b are relatively prime if and only if every integer can be written as a linear combination of a and b.

5. Prove Theorem 2.6. That is, use induction to prove that if the prime p divides $a_1 a_2 \cdots a_n$, then p divides a_i, for some i.

6. Suppose that a and b are positive integers. If $a+b$ is prime, prove that $\gcd(a, b) = 1$.

7. (a) A natural number greater than 1 that is not prime is called **composite**. Show that for any n, there is a run of n consecutive composite numbers. *Hint:* Think factorial.

   (b) Therefore, there is a string of 5 consecutive composite numbers starting where?

8. Prove that two consecutive members of the Fibonacci sequence are relatively prime.

9. Notice that $\gcd(30,50) = 5\gcd(6,10) = 5 \cdot 2$. In fact, this is always true; prove that if $a \neq 0$, then $\gcd(ab, ac) = a \cdot \gcd(b,c)$.

10. Suppose that two integers a and b have been factored into primes as follows:

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

and

$$b = p_1^{m_1} p_2^{m_2} \cdots p_r^{m_r},$$

where the p_i's are primes, and the exponents m_i and n_i are non-negative integers. It is the case that

$$\gcd(a,b) = p_1^{s_1} p_2^{s_2} \cdots p_r^{s_r},$$

where s_i is the smaller of n_i and m_i. Show this with $a = 360 = 2^3 3^2 5$ and $b = 900 = 2^2 3^2 5^2$. Now prove this fact in general.

11. The **least common multiple** of natural numbers a and b is the smallest positive common multiple of a and b. That is, if m is the least common multiple of a and b, then $a|m$ and $b|m$, and if $a|n$ and $b|n$ then $n \geq m$. We will write $\text{lcm}(a,b)$ for the least common multiple of a and b. Find $\text{lcm}(20,114)$ and $\text{lcm}(14,45)$. Can you find a formula for the lcm of the type given for the gcd in the previous exercise?

12. Show that if $\gcd(a,b) = 1$, then $\text{lcm}(a,b) = ab$. In general, show that

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

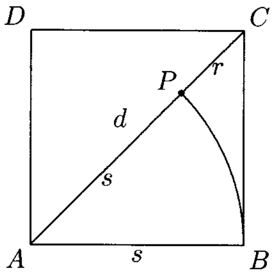
13. Prove that if m is a common multiple of both a and b, then $\text{lcm}(a,b)|m$.

14. Prove that $\sqrt{2}$ is irrational.

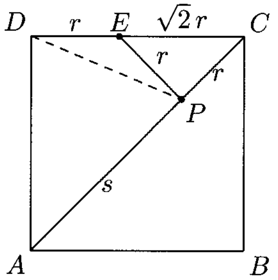
15. This problem outlines another proof that $\sqrt{2}$ is irrational. We show that Euclid's Algorithm never halts if applied to a diagonal d and side s of a square. The first step of the algorithm yields

$$d = 1 \cdot s + r,$$

as shown in the picture below:



(a) Now find the point E by intersecting the side CD with the perpendicular to the diagonal AC at P. It is obvious that the length of segment EC is $\sqrt{2}r$. (Why?) Now prove that the length of segment DE is r, by showing that the triangle DEP is an isosceles triangle.



Why does this mean that the next step in Euclid's Algorithm yields

$$s = 2r + (\sqrt{2} - 1)r?$$

(b) Argue that the next step of the algorithm yields

$$r = 2(\sqrt{2} - 1)r + (\sqrt{2} - 1)^2 r.$$

Hint: Consider the square with three vertices E, P, and C, and use part a. Why does this mean that the algorithm never halts?

16. State Euclid's version of the Fundamental Theorem of Arithmetic in modern language, and prove it carefully as a Corollary of the Fundamental Theorem.

17. (a) As with many algorithms, one can easily write a recursive version of Euclid's Algorithm. This version is for nonnegative a and b. (The symbol $\leftarrow$ is the assignment symbol and $a \bmod b$ is the remainder when dividing a by b.)

```
function $gcd(a,b)$;
  if $b=0$ then gcd $\leftarrow a$ else $gcd \leftarrow gcd(b, a \bmod b)$
endfunction.
```

Try this version on 2424 and 772 and a couple of other pairs of your choice.

(b) One can also write a recursive extended gcd algorithm that returns the linear combination guaranteed by the GCD identity. This procedure again assumes that both a and b are non-negative. When the initial call returns, g will be the gcd of a and b and $g = ax + by$.

```
procedure $extgcd(a,b,g,x,y)$;
  if $b=0$ then
    $g \leftarrow a$; $x \leftarrow 1$; $y \leftarrow 0$;
  else
    $extgcd(b, a \bmod b, g, x, y)$;
    $temp \leftarrow y$;
    $y \leftarrow x - \lfloor a/b \rfloor y$;
    $x \leftarrow temp$;
endprocedure.
```

(Here, $\lfloor x \rfloor$ is the *floor* function. That is, $\lfloor x \rfloor =$ the greatest integer less than or equal to x.) Try this procedure on 285 and 255, then 2424 and 772, and a pair of your choice.

18. (a) Show that in Euclid's Algorithm, the remainders are at least halved after two steps. That is, $r_{i+2} < 1/2\ r_i$.

(b) Use part a to find the maximum number of steps required for Euclid's Algorithm. (Figure this in terms of the maximum of a and b.)

19. Recall from Exercise 1.13 the definition of the binomial coefficient $\binom{n}{k}$. Suppose that p is a positive prime integer, and k is an integer with $1 \le k \le p-1$. Prove that p divides binomial coefficient $\binom{p}{k}$.

Chapter 3

Modular Arithmetic

In this chapter we look again at the content of the Division Theorem 2.1, only this time placing our primary interest on the remainders obtained. By adopting a slightly more abstract point of view, we will be able to obtain some new insight into the arithmetic of $\mathbb{Z}$.

3.1 Residue Classes

For any positive integer m and integer a, the **residue of** a **modulo** m is the remainder one obtains when dividing a by m in the Division Theorem. (We will frequently write 'mod m' for 'modulo m'.)

Example 3.1

The residue of 8 (mod 5) is 3. The residue of -22 (mod 6) is 2.

Of course, many integers have the same residue (mod m). Given an integer a, the set of all integers with the same residue (mod m) as a is called the **residue class (mod** m**) of** a and denoted $[a]_m$.

Example 3.2

For instance,

$$[3]_5 = \{\ldots, -12, -7, -2, 3, 8, \ldots\},$$

and

$$[-22]_6 .= \{\ldots, -22, -16, -10, -4, 2, \ldots\}.$$

If $[a]_m = [b]_m$ we say that a and b are **congruent modulo** m, and write $a \equiv b\,(\text{mod } n)$. We simplify this notation to $a \equiv b$, if it is clear what **modulus** m is being used.

Our intention in this chapter is to define addition and multiplication on these residue classes to give us interesting new number systems. Before doing this we will explore more about the classes themselves.

Notice that $[3]_5$ consists of the list of every fifth integer, which includes 3. That is,

$$[3]_5 = \{\ldots,\ 3+(-3)5,\ 3+(-2)5,\ 3+(-1)5,\ 3+(0)5,\ 3+(1)5, \ldots\}.$$

And similarly, $[-22]_6$ consists of the list of every sixth integer, which includes -22. Our first theorem asserts that this is always true.

Theorem 3.1 $[a]_m = \{a + km : k \in \mathbb{Z}\}$.

Proof: We must show that two infinite sets are in fact equal. Our strategy is to show that each of these sets is a subset of the other. For that purpose, suppose that

$$x \in \{a + km : k \in \mathbb{Z}\}.$$

Then $x = a + k_0 m$ for some $k_0 \in \mathbb{Z}$. Suppose the residue (mod m) of a is r. That is, when we divide a by m, we have remainder r. But then $a = qm + r$, where $0 \leq r < m$ and q is some integer. Then

$$x = a + k_0 m = qm + r + k_0 m = (k_0 + q)m + r.$$

But this means that the residue of x modulo m is r, and so $x \in [a]_m$. Thus,

$$\{a + km : k \in \mathbb{Z}\} \subseteq [a]_m.$$

Now let $x \in [a]_m$. In other words, x has the same residue (mod m) as a. Suppose that the common residue of x and a modulo m is r, and so $x = q_1 m + r$ and $a = q_2 m + r$. Then $r = a - q_2 m$ and so

$$x = q_1 m + a - q_2 m = (q_1 - q_2)m + a.$$

That is, $x \in \{a + km : k \in \mathbb{Z}\}$, proving the theorem. □

As our examples above suggest, this theorem says that elements in a given residue class (mod m) occur exactly once every m integers. So, if $x \in [a]_m$, the next larger element in $[a]_m$ is $x+m$. Hence, any m consecutive integers will contain exactly one element of $[a]_m$. Thus, there are exactly m residue classes (mod m), and we can choose representatives from each class simply by picking any set of m consecutive integers. For example, we could certainly choose the m integers $0, 1, 2, \cdots, m-1$ (which are of course exactly the possible remainders from division by m). Indeed, with this conventional and convenient choice of representatives we can specify the m distinct residue classes as $[0], [1], \ldots, [m-1]$. These m residue classes then **partition** the integers, meaning that each integer belongs to exactly one of these classes, and if distinct classes intersect, they are in fact equal. Alternatively, this means that

$$\mathbb{Z} = [0]\ \cup\ [1]\ \cup\ [2]\ \cup \cdots \cup\ [m-1],$$

and the sets in this union are disjoint from one another pairwise.

In particular, we have that

$$\begin{aligned}\mathbb{Z} &= [0]_4\ \cup\ [1]_4\ \cup\ [2]_4\ \cup\ [3]_4\\ &= \{\ldots,-4,0,4,8,\ldots\}\ \cup\ \{\ldots,-3,1,5,9,\ldots\}\ \cup\\ &\qquad \{\ldots,-2,2,6,10,\ldots\}\ \cup\ \{\ldots,-1,3,7,11,\ldots\}.\end{aligned}$$

The next theorem provides a very useful way of determining when two integers are in the same residue class. Indeed, we will use this characterization more often than the definition itself.

Theorem 3.2 *Two integers, x and y, have the same residue (mod m) if and only if $x - y = km$ for some integer k.*

Proof: First, suppose $x \equiv y \pmod{m}$. Then $x = k_1 m + r$, and $y = k_2 m + r$ for some integers k_1 and k_2 and $0 \leq r < m$. But then $x - y = (k_1 - k_2)m$.

Conversely, suppose $x-y = km$, for some integer k with $x = k_1 m + r_1$ and $y = k_2 m + r_2$, where $0 \leq r_1 < m$ and $0 \leq r_2 < m$. Then

$$km = x - y = (k_1 - k_2)m + r_1 - r_2,$$

which implies that $r_1 - r_2 = (k - k_1 + k_2)m$. Now, because r_1 and r_2 are both non-negative integers less than m, the distance between them is less than m. That is, $|r_1 - r_2| < m$. So, $-m < r_1 - r_2 < m$. But we have just shown that $r_1 - r_2$ is an integer multiple of m. Hence, that multiple is 0. Therefore, $r_1 - r_2 = 0$ or $r_1 = r_2$. □

Example 3.3

We have $[18]_7 = [-38]_7$ because $18 - (-38) = 56 = (7)(8)$.

We now consider the set of all residue classes modulo m. We denote this set by $\mathbb{Z}_m$. That is,

$$\mathbb{Z}_m = \{[0], [1], [2], \cdots, [m-1]\}.$$

Be careful to note that we are considering here a *set of sets*: Each element of the finite set $\mathbb{Z}_m$ is in fact an infinite set of the form $[k]$. While this construct seems abstract, you should take heart from the fact that for the most part, we can focus our attention on particular representatives of the residue classes, rather than on the entire set.

3.2 Arithmetic on the Residue Classes

We are now ready to define an 'arithmetic' on $\mathbb{Z}_m$ which is directly analogous to (and indeed inherited from) the arithmetic on $\mathbb{Z}$. By an 'arithmetic' we mean operations on $\mathbb{Z}_m$ that we call addition and multiplication.

To *add* two elements of $\mathbb{Z}_m$ (that is, two mod m residue classes) simply take a representative from each class. The sum of the two residue classes is defined to be the residue class of their sum. For instance, to add $[3]_5$ and $[4]_5$, we pick, say, $8 \in [3]_5$ and $4 \in [4]_5$. But $[8+4]_5 = [2]_5$, and so $[3]_5 + [4]_5 = [2]_5$. Note that any other choice of representatives would also yield $[2]_5$.

▷ **Quick Exercise.** Try some other representatives of these two residue classes, and see that the same sum is obtained. ◁

It is vitally important that this definition be *independent* of representatives chosen, for otherwise it would be ambiguous and consequently not of much use. We will shortly prove that this independence of representatives in fact holds. Before we do so, we first observe that we can define *multiplication* on $\mathbb{Z}_m$ in a similar way.

More succinctly, the definition of the operations on $\mathbb{Z}_m$ are:

$$[a]_m + [b]_m = [a+b]_m$$
$$[a]_m \cdot [b]_m = [a \cdot b]_m.$$

Thus, $[4]_5[3]_5 = [12]_5 = [2]_5$.

▷ **Quick Exercise.** Try some other representatives of these two residue classes, and see that the same product is obtained. ◁

We now check to see that these definitions are well defined. That is, if one picks different representatives from the residue classes, the result should be the same. You have seen that this worked in the example above for addition and multiplication in $\mathbb{Z}_5$ (at least for the representatives you tried).

Proof that operations are well defined: To show addition on $\mathbb{Z}_m$ is well defined, consider $[a]$ and $[b]$. We pick two representatives from $[a]$, say x and y, and two representatives from $[b]$, say r and s. Now we must show that $[x+r] = [y+s]$. But $x, y \in [a]$ implies $x - y = k_1 m$, for some integer k_1. Likewise, $r - s = k_2 m$, for some integer k_2. So,

$$x + r - (y+s) = x - y + r - s = (k_1 + k_2)m.$$

In other words, $[x+r] = [y+s]$, which is what we wanted to show.

The proof that multiplication on $\mathbb{Z}_m$ is also well defined is similar and is left as Exercise 3.9. □

We now have an 'arithmetic' defined on $\mathbb{Z}_m$. To avoid cumbersome notation, it is common to write the elements of $\mathbb{Z}_m$ as simply $0, 1, \ldots, m-1$ instead of $[0], [1], \ldots, [m-1]$. So, in $\mathbb{Z}_5$, $3 + 4 = 2$ and $2 + 3 = 0$. (Thus, $-2 = 3$ and $-3 = 2$.) Bear in mind that the arithmetic is really on residue classes. For the remainder of this chapter we will not omit the brackets, although later we often will.

Example 3.4

A first simple example of this arithmetic is in the case where $m = 2$. We then have only two residue classes. In fact, $[0]_2$ is precisely the set of even integers and $[1]_2$ is the set of odd integers. The addition and multiplication tables for $\mathbb{Z}_2$ are given below. The addition table may be simply interpreted as 'The sum of an even and an odd is odd, while the sum of two evens or two odds is even.' The multiplication table may be interpreted as 'The product of two integers is odd only when both integers are odd.'

+	[0]	[1]
[0]	[0]	[1]
[1]	[1]	[0]

·	[0]	[1]
[0]	[0]	[0]
[1]	[0]	[1]

addition and multiplication tables for $\mathbb{Z}_2$

3.3 Properties of Modular Arithmetic

It is illuminating to compare the arithmetic on $\mathbb{Z}_m$ with that on $\mathbb{Z}$. Later in the book (in Chapter 6) we will meet a common abstraction of arithmetic on $\mathbb{Z}$ and on $\mathbb{Z}_m$ that will enable us to pursue this general question in more detail. For now we intend only to suggest a few of the ideas we will meet more formally later.

Arithmetic in $\mathbb{Z}$ depends heavily on the existence of an **additive identity** or **zero**. Zero has the pleasant property in $\mathbb{Z}$ that $0 + n = n$, for all integers n. Note that in $\mathbb{Z}_m$ the residue class [0] plays the same role because $[0] + [n] = [0 + n] = [n]$.

Also, each integer n has an **additive inverse** $-n$ in $\mathbb{Z}$, an element which when added to n gives the additive identity 0. This is the formal basis for subtraction, which enables us to solve equations of the form $a+x = b$ in $\mathbb{Z}$ (by simply adding $-a$ to both sides). Notice that additive inverses are available in $\mathbb{Z}_m$ as well. For,

$$[k] + [m - k] = [k + m - k] = [m] = [0],$$

and so $[m-k] = [-k]$ serves as the additive inverse of $[k]$. Consequently, we can always solve equations of the form $[a] + X = [b]$, where here X is an unknown in $\mathbb{Z}_m$.

▷ **Quick Exercise.** Solve the equation $[7]_{12} + X = [4]_{12}$ in $\mathbb{Z}_{12}$, by using the appropriate additive inverse. ◁

We can conveniently summarize the additive arithmetic in $\mathbb{Z}_m$ for a particular m in addition tables. (We have addition tables for $m = 5$ and $m = 6$ below.) Note that these tables reflect the fact that every element of these sets has an additive inverse. (How?)

+	[0]	[1]	[2]	[3]	[4]
[0]	[0]	[1]	[2]	[3]	[4]
[1]	[1]	[2]	[3]	[4]	[0]
[2]	[2]	[3]	[4]	[0]	[1]
[3]	[3]	[4]	[0]	[1]	[2]
[4]	[4]	[0]	[1]	[2]	[3]

+	[0]	[1]	[2]	[3]	[4]	[5]
[0]	[0]	[1]	[2]	[3]	[4]	[5]
[1]	[1]	[2]	[3]	[4]	[5]	[0]
[2]	[2]	[3]	[4]	[5]	[0]	[1]
[3]	[3]	[4]	[5]	[0]	[1]	[2]
[4]	[4]	[5]	[0]	[1]	[2]	[3]
[5]	[5]	[0]	[1]	[2]	[3]	[4]

addition tables $\mathbb{Z}_5$ and $\mathbb{Z}_6$

What about multiplication? In $\mathbb{Z}$ the integer 1 serves as a **multiplicative identity**, because $1 \cdot n = n$ for all integers n, and clearly [1] plays the same role in $\mathbb{Z}_m$.

▷ **Quick Exercise.** Check this. ◁

A multiplicative inverse in $\mathbb{Z}_m$ may be defined analogously to the way we have defined an additive inverse: $[a]$ is the **multiplicative inverse** of $[n]$ if $[a][n] = [1]$. The disadvantage of $\mathbb{Z}$ as opposed to $\mathbb{Q}$ is that no elements have multiplicative inverses (except 1 and -1). The consequence is that many equations of the form $ax = b$ are *not* solvable in the integers. But what about in $\mathbb{Z}_m$? Consider the following multiplication tables for our examples $\mathbb{Z}_5$ and $\mathbb{Z}_6$.

·	[0]	[1]	[2]	[3]	[4]
[0]	[0]	[0]	[0]	[0]	[0]
[1]	[0]	[1]	[2]	[3]	[4]
[2]	[0]	[2]	[4]	[1]	[3]
[3]	[0]	[3]	[1]	[4]	[2]
[4]	[0]	[4]	[3]	[2]	[1]

·	[0]	[1]	[2]	[3]	[4]	[5]
[0]	[0]	[0]	[0]	[0]	[0]	[0]
[1]	[0]	[1]	[2]	[3]	[4]	[5]
[2]	[0]	[2]	[4]	[0]	[2]	[4]
[3]	[0]	[3]	[0]	[3]	[0]	[3]
[4]	[0]	[4]	[2]	[0]	[4]	[2]
[5]	[0]	[5]	[4]	[3]	[2]	[1]

multiplication tables $\mathbb{Z}_5$ and $\mathbb{Z}_6$

Notice the remarkable fact that in $\mathbb{Z}_5$, every element (other than [0]) has a multiplicative inverse. For example, the multiplicative inverse of [3] is [2], because $[3][2] = [1]$. Thus, to solve the equation $[3]X = [4]$ in $\mathbb{Z}_5$, we need only multiply both sides of the equation by the multiplicative inverse of [3] (which is [2]) to obtain

$$X = [2][3]X = [2][4] = [3].$$

On the other hand, [3] has no multiplicative inverse in $\mathbb{Z}_6$ and there is in fact no solution to the equation $[3]X = [2]$ in $\mathbb{Z}_6$.

▷ **Quick Exercise.** Solve the equation $[4]X = [10]$ in $\mathbb{Z}_{11}$. Then argue that this equation has no solution in $\mathbb{Z}_{12}$. ◁

We have gone far enough here to illustrate the fact that the arithmetic in $\mathbb{Z}_m$ shares similarities with those of $\mathbb{Z}$, but also has some real differences (which depend on the choice of m).

Historical Remarks

The great German mathematician Karl Friedrich Gauss (1777-1855) first introduced the idea of congruence modulo m into the study of integers, in his important book *Disquisitiones Arithmeticae.* Gauss made important contributions to almost all branches of mathematics and did important work in astronomy and physics as well, but number theory (the study of the mathematical properties of the integers) was his first love. The *Disquisitiones* was a landmark work, which systematized and extended the work on number theory done by Gauss's predecessors, Fermat and Euler. Gauss's introduction of the notion of congruence is a good example of the way in which an effective and efficient notation can revolutionize the way a mathematical subject is approached.

Chapter Summary

In this chapter we defined the *residue class* $[a]_m$ of a modulo m (for a positive integer m) and characterized the elements of such classes. We then considered the set $\mathbb{Z}_m$ of the m residue classes and defined an *arithmetic* on this set. We proved the following facts about this arithmetic:

- Addition and multiplication are well defined;
- $\mathbb{Z}_m$ has an additive identity $[0]$ and a multiplicative identity $[1]$;
- All elements in $\mathbb{Z}_m$ have additive inverses, but not all elements have multiplicative inverses.

Warm-up Exercises

a. Write out the three residue classes modulo 3 (as we did for $\mathbb{Z}_4$). Write out the addition and multiplication tables for $\mathbb{Z}_3$. Which elements of $\mathbb{Z}_3$ have multiplicative inverses?

b. Does $\{47, 100, -3, 29, -9\}$ contain a representative from every residue class of $\mathbb{Z}_5$? Does $\{-14, -21, -10, -3, -2\}$? Does $\{10, 21, 32, 43, 54\}$?

c. What is the additive inverse of $[13]$ in $\mathbb{Z}_{28}$?

d. What is the relationship between 'clock arithmetic' and modular arithmetic?

e. (a) What time is it 100 hours after 3 o'clock?

   (b) What day of the week is it 100 days after Monday?

f. Solve the following equations, or else argue that they have no solutions:

   (a) $[4] + X = [3]$, in $\mathbb{Z}_6$.

   (b) $[4]X = [3]$, in $\mathbb{Z}_6$.

   (c) $[4] + X = [3]$, in $\mathbb{Z}_9$.

   (d) $[4]X = [3]$, in $\mathbb{Z}_9$.

Exercises

1. Repeat Warm-up Exercise a for modulo 8.
2. Determine the elements in $\mathbb{Z}_{15}$ that have multiplicative inverses. Give an example of an equation of the form $[a]X = [b]$ ($[a] \neq [0]$) that has no solution in $\mathbb{Z}_{15}$.
3. In Exercise c you determined the additive inverse of $[13]$ in $\mathbb{Z}_{28}$. Now determine its multiplicative inverse.
4. Find an example in $\mathbb{Z}_6$ where $[a][b] = [a][c]$, but $[b] \neq [c]$. How is this example related to the existence of multiplicative inverses in $\mathbb{Z}_6$?

5. If $\gcd(a,m)=1$, then the GCD identity 2.4 guarantees that there exist integers u and v such that $1=au+mv$. Show that in this case, $[u]$ is the multiplicative inverse of $[a]$ in $\mathbb{Z}_m$.

6. Now use essentially the reverse of the argument from Exercise 5 to show that if $[a]$ has a multiplicative inverse in $\mathbb{Z}_m$ then $\gcd(a,m)=1$.

7. According to what you have shown in Exercises 5 and 6, which elements of $\mathbb{Z}_{24}$ have multiplicative inverses? What are the inverses for each of those elements? (The answer is somewhat surprising.)

8. Repeat the previous exercise for $\mathbb{Z}_{10}$. Give the multiplication table for those elements in $\mathbb{Z}_{10}$ that have multiplicative inverses and find an $[n]$ such that all these elements are powers of $[n]$.

9. Prove that the multiplication on $\mathbb{Z}_m$ as defined in the text is well defined, as claimed in Section 3.2.

10. Prove that if all non-zero elements of $\mathbb{Z}_m$ have multiplicative inverses, then multiplicative cancellation holds: that is, if $[a][b]=[a][c]$, then $[b]=[c]$.

11. Consider the following alternate definition of addition of residue classes in $\mathbb{Z}_m$, by defining the set

$$S=\{x+y : x\in[a], y\in[b]\}.$$

Prove that $S=[a]+[b]$ (as defined in Section 3.2); thus, we could have used the definition above to define addition in $\mathbb{Z}_m$.

12. By way of analogy with Exercise 11, one might try to define the multiplication of residue classes in $\mathbb{Z}_m$ by considering the set

$$M=\{xy : x\in[a], y\in[b]\}.$$

Prove that $M\subseteq[a][b]$. Then choose particular m,a,b to show by example that this containment can be proper (that is, $M\subset[a][b]$).

13. In the integers, the equation $x^2=a$ has a solution only when a is a positive perfect square or zero. For which $[a]$ does the equation $X^2=[a]$ have a solution in $\mathbb{Z}_7$? What about in $\mathbb{Z}_8$? What about in $\mathbb{Z}_9$?

14. Explain what $a\equiv b \pmod 1$ means.

Chapter 4

Polynomials with Rational Coefficients

In Chapter 2 we proved that every integer ($\neq 0,\pm1$) can be written as a product of irreducible integers, and this decomposition is essentially unique. These irreducible integers turn out to be those integers that we call primes. To summarize, in that chapter we proved the following important theorems:

- The Division Theorem for integers (Theorem 2.1),
- Euclid's Algorithm (which yields the gcd of two integers) (Theorem 2.3),
- The GCD identity that $\gcd(a,b)=ax+by$, for some integers x and y (Theorem 2.4),
- Each non-zero integer ($\neq\pm1$) is either irreducible or a product of irreducibles (Theorem 2.8),
- An integer p is irreducible if and only if p is prime (that is, if $p|ab$, then either $p|a$ or $p|b$) (Theorem 2.7), and
- Each non-zero integer ($\neq\pm1$) is uniquely (up to order and factors of -1) the product of primes (Theorem 2.9).

4.1 Polynomials

In this chapter we turn our attention to another algebraic structure with which you are familiar—the polynomials (with unknown x) with coefficients from the rational numbers $\mathbb{Q}$. In this chapter and the next we discover that this set of polynomials obeys theorems directly analogous to those we have listed above for the integers.

We denote the set of polynomials with rational coefficients by $\mathbb{Q}[x]$. Let's be careful to define exactly what we mean by a polynomial. A **polynomial** $f \in \mathbb{Q}[x]$ is an expression of the form

$$f = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n + \cdots$$

where $a_i \in \mathbb{Q}$, and all but finitely many of the a_i's are 0. We call the a_i's the **coefficients** of the polynomial. When we write down particular polynomials, we will simply omit a term if the coefficient happens to be zero. Thus, such expressions as $2 + x$, $\frac{4}{7} + 2x^2 - \frac{1}{2}x^3$, and 14 are all elements of $\mathbb{Q}[x]$. Henceforth, when we wish to write down a generic polynomial, we will usually be content with an expression of the form $f = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$. This means that we're assuming that $a_m = 0$, for all $m > n$. It may of course be the case that some of the a_m for $m \leq n$ are 0 too.

We say that two polynomials are **equal** if and only if their corresponding coefficients are equal. Thus, $2 + 0x - x^2, 2 - x^2 + 0x^3$, and $2 - x^2$ are all equal polynomials. The first two polynomials are simply less compact ways of writing the third.

For the most part we deal with polynomials with rational coefficients, but sometimes we wish to restrict our attention to those polynomials whose coefficients are integers; we denote this set by $\mathbb{Z}[x]$. Of course $\mathbb{Z}[x]$ is a proper subset of $\mathbb{Q}[x]$.

Note that x is a formal symbol, not a variable or an indeterminate element of $\mathbb{Q}$. This is probably different from the way you are used to thinking of a polynomial, which is as a function from $\mathbb{Q}$ to $\mathbb{Q}$ (or from $\mathbb{R}$ to $\mathbb{R}$). This is not how we think of them here—we think of a polynomial as a formal expression. In fact, if we consider polynomials with coefficients taken not from $\mathbb{Q}$ but some other number system, two of these new polynomials can be equal as functions but not as polynomials.

▷ **Quick Exercise.** Consider polynomials with coefficients from $\mathbb{Z}_2$—denoted by $\mathbb{Z}_2[x]$, naturally. Show that the three different polynomials $x^2 + x + 1$, $x^4 + x^3 + x^2 + x + 1$, and 1 are indeed the same function from $\mathbb{Z}_2$ to $\mathbb{Z}_2$. (Two functions are equal if they have the same value at all points in their domain.) ◁

We will nearly always think of polynomials in the formal sense. To emphasize this point of view, when we speak of a particular polynomial we will denote it by a letter like f, rather than writing $f(x)$. The one time we wish to consider a polynomial as a function in this chapter will be made explicit, and then we will refer to it as a **polynomial function**.

The **degree** of a polynomial is the largest exponent with corresponding non-zero coefficient. So, a polynomial of degree 0 means the polynomial can be considered an element of $\mathbb{Q}$ (sometimes called a **scalar**). Of course, the zero polynomial has no non-zero coefficients. To cover this special case conveniently, we say that its degree is $-\infty$. We will denote the degree of a polynomial f by $\deg(f)$.

4.2 The Algebra of Polynomials

We can add, subtract, and multiply polynomials in the ways with which you are already familiar: If $f = a_0 + a_1x + \cdots + a_nx^n$ and $g = b_0 + b_1x + \cdots + b_mx^m$ (let's suppose $n > m$), then

$$\begin{aligned} f + g = {} & (a_0 + b_0) + (a_1 + b_1)x + \cdots \\ & + (a_m + b_m)x^m + a_{m+1}x^{m+1} + \cdots + a_nx^n. \end{aligned}$$

The difference, $f - g$, is similarly defined. The definition of product is more difficult to write down abstractly; the following definition actually captures your previous experience in multiplying polynomials:

$$f \cdot g = a_0b_0 + (a_0b_1 + a_1b_0)x + \cdots + \sum_{i+j=k} (a_ib_j)x^k + \cdots + (a_nb_m)x^{n+m}.$$

That is, the coefficient of x^k is the sum of all the products of the coefficients of x^i in f with the coefficients of x^j in g where i and j sum to k.

Example 4.1

If $f = 3 + x^4 - 2x^5 + x^6 + 2x^7$ and $g = -1 + 3x + x^2 + 4x^6$, then the coefficient of x^6 in $f \cdot g$ is $3 \cdot 4 + 1 \cdot 1 + (-2) \cdot 3 + 1 \cdot (-1) = 6$.

How is degree affected when we add or multiply polynomials? Your previous experience with polynomials suggests the right answer, which is contained in the following theorem.

Theorem 4.1 *Let $f, g \in \mathbb{Q}[x]$. Then*

a. $\deg(fg) = \deg(f) + \deg(g)$, *where it is understood that $-\infty$ added to anything is $-\infty$.*

b. $\deg(f+g)$ *is less than or equal to the larger of the degrees of f and g.*

Proof: We prove part a first. We consider first the case where one of the polynomials is the zero polynomial. Now, it is evident that $0g = 0$, for any polynomial g. Thus,

$$-\infty = \deg(0) = \deg(0g) = -\infty + \deg(g),$$

as required.

We thus may as well assume that neither f nor g is the zero polynomial; suppose that $\deg(f) = n$ and $\deg(g) = m$. Then $f = a_nx^n + f_1$, where $a_n \neq 0$ and $\deg(f_1) < n$. Similarly, $g = b_mx^m + g_1$, where $b_m \neq 0$ and $\deg(g_1) < m$. By the definition of multiplication of polynomials, the coefficient on x^{n+m} is a_nb_m, and this is not zero because neither factor is. But all remaining terms in the product have smaller degree than $n+m$, and so

$$\deg(fg) = n + m = \deg(f) + \deg(g),$$

as required.

▷ **Quick Exercise.** You prove part b. Also show by example that the degree of a sum of polynomials can be strictly smaller than the larger of the degrees. *Hint:* Take two polynomials with the same degree. ◁ □

An important particular case of the first part of this theorem is this: If a product of two polynomials is the zero polynomial, then one of the factors is the zero polynomial.

▷ **Quick Exercise.** Prove this, using the theorem. ◁

4.3 The Analogy between $\mathbb{Z}$ and $\mathbb{Q}[x]$

We will now begin to prove the theorems analogous to those proved about natural numbers and integers and summarized above. (Actually, in this chapter, *you* will do some of the proving.) You should notice, as you proceed through this chapter and the next, that not only are the theorems similar, but so are the proofs. (You will probably even be able to anticipate some theorems.) This suggests that the integers share properties with $\mathbb{Q}[x]$ that give rise to these theorems—in particular, unique factorization. Later, we will be able to identify these properties and prove unique factorization in a more general setting. This process of generalization is indeed a common theme in mathematics —one sees that A and B both have property C. What is shared by A and B that forces property C on both? For now, we are content to consider the concrete example of $\mathbb{Q}[x]$ and try to build more insight before getting abstract.

Before starting, recall that the proof technique used for most of the important theorems for natural numbers is induction. When considering polynomials, we also frequently use induction, but on the *degree* of the polynomial. Note that since the set of degrees of polynomials is $\{-\infty, 0, 1, 2, \ldots\}$, which is well ordered, induction may be used.

We now start, as with the integers, with the Division Theorem.

Theorem 4.2 Division Theorem for $\mathbb{Q}[x]$ *Let $f, g \in \mathbb{Q}[x]$ with $f \neq 0$. Then there are unique polynomials q and r, with $\deg(r) < \deg(f)$, such that $g = fq + r$.*

Before proving this theorem, we look at an example.

The actual computation for producing q and r, for given polynomials f and g, is just long division. For example, let $f = x^2 + 2x - 1$ and $g = x^4 + x^2 - x + 2$.

$$\begin{array}{r|rcrcrcrcr}
 & x^2 & - & 2x & + & 6 & & & & \\ \cline{2-10}
x^2 + 2x - 1 & x^4 & + & & & x^2 & - & x & + & 2 \\
 & x^4 & + & 2x^3 & - & x^2 & & & & \\ \cline{2-6}
 & & & -2x^3 & + & 2x^2 & - & x & + & 2 \\
 & & & -2x^3 & - & 4x^2 & + & 2x & & \\ \cline{4-10}
 & & & & & 6x^2 & - & 3x & + & 2 \\
 & & & & & 6x^2 & + & 12x & - & 6 \\ \cline{6-10}
 & & & & & & - & 15x & + & 8
\end{array}$$

Hence, $q = x^2 - 2x + 6$ and $r = -15x + 8$. That is,

$$x^4 + x^2 - x + 2 = (x^2 + 2x - 1)(x^2 - 2x + 6) + (-15x + 8).$$

▷ **Quick Exercise.** Find q and r as guaranteed by the Division Theorem for $g = x^5 + x - 1$ and $f = x^2 + x$. ◁

Proof of the Division Theorem: We first prove the existence of q and r, using induction on the degree of g. The base case for induction in this proof is $\deg(g) < \deg(f)$. If this is the case, then $g = f \cdot 0 + g$. So, $q = 0$ and $r = g$ satisfy the requirements of the theorem.

We now assume that $f = a_0 + a_1x + \cdots + a_nx^n$ and $g = b_0 + b_1x + \cdots + b_mx^m$, and $m = \deg(g) \geq \deg(f) = n$. Our induction hypothesis says that we can find a quotient and remainder whenever the dividend has degree less than m.

Let $h = g - (b_m/a_n)x^{m-n}f$. This makes sense because $m \geq n$. Note that the largest non-zero coefficient of g has been eliminated in h, so $\deg(h) < \deg(g)$. Hence, by the induction hypothesis, $h = fq' + r$, where $\deg(r) < \deg(f)$. But then,

$$\begin{aligned} g &= fq' + r + (b_m/a_n)x^{m-n}f \\ &= f(q' + (b_m/a_n)x^{m-n}) + r. \end{aligned}$$

Thus, $q = q' + (b_m/a_n)x^{m-n}$ and r serve as the desired quotient and remainder.

Now we prove the uniqueness of q and r by supposing that $g = fq + r = fq' + r'$, where $\deg(r) < \deg(f)$ and $\deg(r') < \deg(f)$. We will show that $q = q'$ and $r = r'$.

So, because $fq + r = fq' + r'$, we have that $f(q - q') = r' - r$. Because $\deg(f) > \deg(r)$ and $\deg(f) > \deg(r')$, we have $\deg(f) > \deg(r' - r)$. But $\deg(f(q - q')) \geq \deg(f)$, unless $f(q - q') = 0$. Hence, $\deg(f(q - q')) > \deg(r' - r)$ unless both are the zero polynomial. But this must be the case, and so $f(q - q') = 0$, forcing $q - q' = 0$ and $r' - r = 0$, proving uniqueness. □

4.4 Factors of a Polynomial

We now make some definitions analogous to those we made for $\mathbb{Z}$. We say a polynomial f **divides** a polynomial g if $g = fq$ for some polynomial q. In this case we say that f is a **factor** of g, and write $f|g$. In the context of the Division Theorem, $f|g$ means that the remainder obtained is 0. For example, $(x^2 + 1)|(2x^3 - 3x^2 + 2x - 3)$, because

$$2x^3 - 3x^2 + 2x - 3 = (x^2 + 1)(2x - 3).$$

Notice that any polynomial divides the zero polynomial because $0 = (f)(0)$.

Suppose now that a is a non-zero constant polynomial, and $f = a_0 + a_1x + \cdots + a_nx^n$ is any other polynomial. Then a necessarily divides f because

$$f = (a)\left(\frac{a_0}{a} + \frac{a_1}{a}x + \frac{a_2}{a}x^2 + \cdots + \frac{a_n}{a}x^n\right).$$

In Exercise 4.11 you will prove that the converse of this statement is true.

4.5 Linear Factors

In practice, it may be *very* difficult to find all the factors of a given polynomial. However, the following theorem shows how to determine factors of the form $x - a$, where $a \in \mathbb{Q}$.

Note carefully that the next theorem and its corollary are the only times in this chapter where we think of a polynomial as a function. For $f \in \mathbb{Q}[x]$ and $a \in \mathbb{Q}$, we define $f(a)$ to be the result that ensues when we replace x in f by a, and then apply the ordinary operations of arithmetic in $\mathbb{Q}$ to simplify the result. Thus, if $f = \frac{1}{3}x^2 - 2x + 1$ and $a = 2$, then

$$f(2) = \frac{1}{3}(2)^2 - 2(2) + 1 = -\frac{5}{3}.$$

This definition obviously gives us a function $f(x)$ which is defined for all rational numbers.

Of particular interest to us is the case when $f(a) = 0$. We say $a \in \mathbb{Q}$ is a **root** of $f \in \mathbb{Q}[x]$ if $f(a) = 0$. Thus, $\frac{2}{3}$ is a root of $g = 3x^3 + 19x^2 - 11x - 2$, because $g\left(\frac{2}{3}\right) = 0$.

Theorem 4.3 Root Theorem *If f is a polynomial in $\mathbb{Q}[x]$ and $a \in \mathbb{Q}$, then $x - a$ divides f if and only if a is a root of f.*

Proof: If $x - a$ divides f, then $f = (x-a)q$, and so

$$f(a) = (a-a)q(a) = 0.$$

Conversely, suppose $f(a) = 0$. Using the Division Theorem 4.2, we write $f = (x-a)g + r$ where $\deg(r) < \deg(x-a) = 1$. But $\deg(r) < 1$ means $\deg(r) = 0$ or $-\infty$; that is, $r \in \mathbb{Q}$. Thus, when we view r as a function, it is a constant function. We might as well call this constant r. Hence, $f(a) = (a-a)q(a) + r$. But, the left-hand side is 0 while the right-hand side is r. Hence, $r = 0$, and so $x - a$ divides f. □

Example 4.2

Consider the polynomial

$$f = x^4 + 2x^3 + x^2 + x - 2.$$

We can conclude that f has a factor of $x + 2$ because $f(-2) = 0$. We need not go through the trouble of long division to verify the fact.

In general, evaluating $f(a)$ is a simpler operation than dividing f by $x - a$, although the latter does have the advantage of giving the factorization if indeed $x - a$ is a factor.

▷ **Quick Exercise.** Check to see whether $x + 2$ is a factor of the following polynomials: $x^3 - 4x$, $x^2 + 2x - 1$, and $x^{100} - 4x^{98} + x + 2$. ◁

Notice that the Root Theorem 4.3 allows us to determine when any given linear factor, $ax + b$ $(a \neq 0)$, divides a polynomial f, because $ax + b$ is a factor if and only if $x + \frac{b}{a}$ is a factor, and the Root Theorem says that this last is a factor if and only if $-\frac{b}{a}$ is a root.

▷ **Quick Exercise.** Let $f \in \mathbb{Q}[x]$ and $a \neq 0$. Prove that $ax + b$ divides f if and only if $x + \frac{b}{a}$ divides f. ◁

Corollary 4.4 *A non-zero polynomial f of degree n has at most n distinct roots in $\mathbb{Q}$.*

Proof: We proceed by induction on the degree of f. If f is non-zero and degree 0, then clearly f has no roots.

Suppose the corollary holds for polynomials with degree smaller than n, and f has degree $n > 0$. If f has no roots, the theorem is proved. So, assume f has at least one root, call it a. Then by the Root Theorem 4.3, $x - a$ divides f. That is, $f = (x-a) \cdot g$. But $\deg(g) = \deg(f) - 1$. And so, by the induction hypothesis, g has at most $n-1$ roots. Because any root of f other than a must also be a root of g, f can have no more than n roots altogether. □

4.6 Greatest Common Divisors

We now turn our attention to finding greatest common divisors of two polynomials, paralleling our development for the integers. Notice that we didn't say *the* greatest common divisor. We will find some differences from the integers here.

If f and g are two polynomials in $\mathbb{Q}[x]$ (not both zero), then a polynomial d in $\mathbb{Q}[x]$ is a **greatest common divisor** (gcd) of f and g if it satisfies the following two criteria:

1. d divides both f and g (d is a **common divisor**), and

2. if a polynomial e divides both f and g, then $\deg(e) \leq \deg(d)$; that is, d has largest degree among common divisors.

Notice that according to this definition $x + 1$ is a gcd for $x^2 - 1$ and $x^2 + 4x + 3$, but so are $-\frac{1}{2}x - \frac{1}{2}$, $-x - 1$, and $20x + 20$. This is unlike the integer case, where we have a unique gcd for any pair of integers (not both zero). In fact, for polynomials, we have infinitely many distinct gcds. For if a is any non-zero rational number and d is a gcd of f and g, then so is the polynomial ad. This follows because ad is also a common divisor of f and g, and has the same degree as d.

▷ **Quick Exercise.** Argue that $x + 1$ is indeed a gcd of $x^2 - 1$ and $x^2 + 4x + 3$ as are the other polynomials listed. ◁

As with the integers, Euclid's Algorithm, when applied to polynomials, yields a gcd. (As you might expect, the arithmetic is messier.)

Example 4.3

As an example of this, consider the polynomials

$$x^6 - x^5 + 7x^4 + 5x^2 + 9x - 21$$

and

$$2x^4 - x^3 + 5x^2 - 3x - 3.$$

By repeatedly applying the Division Theorem 4.2 we obtain the following:

$$\begin{aligned}
&(x^6 - x^5 + 7x^4 + 5x^2 + 9x - 21) = \\
&\qquad (2x^4 - x^3 + 5x^2 - 3x - 3)\left(\frac{x^2}{2} - \frac{x}{4} + \frac{17}{8}\right) \\
&\qquad\qquad + \left(\frac{39}{8}\right)(x^3 - x^2 + 3x - 3) \\
&(2x^4 - x^3 + 5x^2 - 3x - 3) = \\
&\qquad \left(\left(\frac{39}{8}\right)(x^3 - x^2 + 3x - 3)\right)\left(\left(\frac{8}{39}\right)(2x + 1)\right) + 0.
\end{aligned}$$

This means that $(39/8)(x^3 - x^2 + 3x - 3)$ is a gcd of the given polynomials; thus, $x^3 - x^2 + 3x - 3$ is one also.

▷ **Quick Exercise.** Apply Euclid's Algorithm to the polynomials

$$x^3 - 3x^2 + 5x - 3 \text{ and } x^4 - 2x^3 + 4x^2 - 2x + 3. \quad \triangleleft$$

We state formally in the next theorem that Euclid's Algorithm gives gcds in the polynomial case. You should notice that the proof of this theorem is nearly identical to the proof of the corresponding theorem (Theorem 2.3) for integers.

Theorem 4.5 Euclid's Algorithm produces a GCD
In Euclid's Algorithm for polynomials f and g, the last non-zero remainder is a greatest common divisor for f and g.

Proof: The proof of this fact for the integers depended on a lemma, which remains true in this context. Namely, we observe that if $f = gq + r$, then d is a gcd of f and g if and only if d is also a gcd of g and r. (As before, this follows because in fact every common divisor of f and g is also a common divisor of g and r, and vice versa.)

▷ **Quick Exercise.** Show that d is a gcd of f and g exactly if d is a gcd of g and r, following the proof of Lemma 2.2, if necessary. ◁

We proceed by induction on the number of steps required for Euclid's Algorithm to terminate. If Euclid's Algorithm takes one step, $f = gq$ and so $d = g$, which is clearly a gcd of f and g in this case.

So suppose Euclid's Algorithm takes n (> 1) steps to obtain d with the first two steps being

$$\begin{aligned} f &= gq_0 + r_0 \\ g &= r_0q_1 + r_1 \end{aligned}$$

Performing Euclid's Algorithm on g and r_0 also yields d; the process is exactly the last $n - 1$ steps of Euclid's Algorithm for f and g. But then by the induction hypothesis, d is a gcd of g and r_0, and so, by the lemma, d is a gcd of f and g. □

We observed above that if d is a gcd of f and g, then so is any non-zero scalar multiple of d. The converse is also true. That is, if d is any gcd of two polynomials, then all the gcds are simply scalar multiples of d. This follows from the next theorem.

Theorem 4.6 *If d is the gcd of f and g produced by Euclid's Algorithm and e is another common divisor of f and g, then e divides d.*

Proof: The proof of this theorem is nearly identical to the proof that Euclid's Algorithm produces a gcd. The only real addition is that we also check that e divides what it is supposed to. Again, we use induction on n, the number of steps in Euclid's Algorithm when obtaining d.

If Euclid's Algorithm takes one step, then $f = gq$, and so $d = g$. Clearly, if e divides f and $g(= d)$, then e divides d.

Now, suppose Euclid's Algorithm takes $n(> 1)$ steps to obtain d, with the first two steps being

$$\begin{aligned} f &= gq_0 + r_0 \\ g &= r_0q_1 + r_1 \end{aligned}$$

Because $r_0 = f - gq_0$, if e divides f and g, e also divides r_0. Now d is also the gcd of g and r_0 obtained from Euclid's Algorithm. The process is the last $n-1$ steps of Euclid's Algorithm for f and g. So, by the induction hypothesis, because e divides g and r_0, it also divides d. □

An important consequence of this theorem is that any two gcds of two given polynomials are just scalar multiples of one another.

▷ **Quick Exercise.** Show that if e and d are two gcds of polynomials f and g, then d and e are scalar multiples of one another. ◁

Finally, we have the GCD identity for $\mathbb{Q}[x]$.

Theorem 4.7 GCD identity for polynomials *If d is a gcd of polynomials f and g, then there exist polynomials a and b so that $d = af + bg$.*

Proof: The proof is Exercise 4.8 (or 4.9). □

Chapter Summary

In this chapter we introduced the set $\mathbb{Q}[x]$ of all polynomials with rational coefficients and described the arithmetic of this set. In direct analogy with $\mathbb{Z}$, we proved the *Division Theorem*, *Euclid's Algorithm*, and the *GCD identity*. In addition, we described the relationship between the roots of a polynomial and its linear factors; this relationship is called the *Root Theorem.*

Warm-up Exercises

a. Compute the sum, difference, and product of the polynomials

$$1 - 2x + x^3 - \frac{2}{3}x^4 \quad \text{and} \quad 2 + 2x^2 - \frac{3}{2}x^3$$

in $\mathbb{Q}[x]$.

b. Give the quotient and remainder when the polynomial $2 + 4x - x^3 + 3x^4$ is divided by $2x + 1$.

c. Give two polynomials f and g, where the degree of $f+g$ is strictly less than either the degree of f or the degree of g.

d. Use the Root Theorem 4.3 to answer the following for polynomials in $\mathbb{Q}[x]$. Does $x-2$ divide $x^5-4x^4-4x^3-x^2+4$? Does $x+1$ divide $x^6+2x^5+x^4-x^3+x$? Does $x+5$ divide $2x^3+10x^2-2x-10$? Does $2x-1$ divide $x^5+2x^4-3x^2+1$?

e. We started this chapter with a list of theorems true about $\mathbb{Z}$. For how many of them have we stated and proved an analogue for $\mathbb{Q}[x]$?

Exercises

1. Find the gcd of the polynomials $x^6+x^5-2x^4-x^3-x^2+2x$ and $3x^6+4x^5-3x^3-4x^2$ in $\mathbb{Q}[x]$, and express them as $af+bg$, for some polynomials a, b.

2. Divide the polynomial x^2-3x+2 by the polynomial $2x+1$, to obtain a quotient and remainder as guaranteed by the Division Theorem 4.2. Note that although x^2-3x+2 and $2x+1$ are elements of $\mathbb{Z}[x]$, the quotient and remainder are not. Argue that this means that there is no Division Theorem for $\mathbb{Z}[x]$.

3. By Corollary 4.4 we know that a third-degree polynomial in $\mathbb{Q}[x]$ has at most three roots. Give four examples of third-degree polynomials in $\mathbb{Q}[x]$ that have 0, 1, 2, and 3 roots, respectively; justify your assertions. (Recall that here a root must be a rational number!)

4. Your example in the previous exercise of a third-degree polynomial with exactly 2 roots had one **repeated root**; that is, a root a where $(x-a)^2$ is a factor of the polynomial. (Roots may have multiplicity greater than two, of course.) Why can't a third-degree polynomial in $\mathbb{Q}[x]$ have exactly 2 roots where neither is a multiple root?

5. Let n be an odd integer and consider the polynomial

$$\Phi_n = x^n + x^{n-1} + \cdots x + 1.$$

Use the Root Theorem 4.3 to argue that Φ_n has a linear factor. We call Φ_n a **cyclotomic** polynomial; see Exercise 5.17 for more information.

6. Suppose that $f \in \mathbb{Q}[x]$, $q \in \mathbb{Q}$, and $\deg(f) > 0$. Use the Root Theorem 4.3 to prove that the equation $f(x) = q$ has at most finitely many solutions.

7. Let $f \in \mathbb{Q}[x]$. Recall from Exercise 4 above that a rational number a is a repeated root of f if $(x-a)^2$ is a factor of f. Given $f = a_0 + a_1x + \cdots + a_nx^n$, we define the **formal derivative** of f as $f' = na_nx^{n-1} + (n-1)a_{n-1}x^{n-2} + \cdots + a_1$. Prove that a is a repeated root of f if and only if a is a root of both f and f'. Conclude that if f is irreducible, then f has no repeated roots.

8. Prove Theorem 4.7: the GCD identity for $\mathbb{Q}[x]$. Use Euclid's Algorithm 4.5, and the relationship we know between the gcd produced by the algorithm and an arbitrary gcd (Theorem 4.6).

9. One can also prove the GCD identity for $\mathbb{Q}[x]$ with an argument similar to the existential proof of the GCD identity for integers, found in Section 2.3. Try this approach.

10. We say that $p \in \mathbb{Q}[x]$ has a multiplicative inverse if there exists a $q \in \mathbb{Q}[x]$ such that $pq = 1$. Prove that $p \in \mathbb{Q}[x]$ has a multiplicative inverse if and only if $\deg(p) = 0$.

11. Suppose that $g \in \mathbb{Q}[x]$, and g divides all elements of $\mathbb{Q}[x]$. Prove that g is a non-zero constant polynomial.

12. Find two different polynomials in $\mathbb{Z}_3[x]$ that are equal as functions from $\mathbb{Z}_3$ to $\mathbb{Z}_3$.

13. Find a non-zero polynomial in $\mathbb{Z}_4[x]$ for which $f(a) = 0$, for all $a \in \mathbb{Z}_4$.

Chapter 5

Factorization of Polynomials

We have already seen the Fundamental Theorem of Arithmetic, which says that every integer (other than 0, 1, and -1) can be uniquely factored into primes. We wish to come up with a corresponding theorem for the set $\mathbb{Q}[x]$ of polynomials with rational coefficients.

5.1 Factoring Polynomials

We note first that uniqueness of factorization cannot be as nice for polynomials as for integers because any factorization in $\mathbb{Q}[x]$ can be adjusted by factoring out scalars. The following example shows what we mean:

$$\begin{aligned} x^2 - 4 &= (x+2)(x-2) \\ &= \left(\frac{1}{2}x + 1\right)(2x-4) \\ &= (2x+4)\left(\frac{1}{2}x - 1\right), \end{aligned}$$

and so on.

But there is a close connection, after all, between the factors $x+2$ and $\frac{1}{2}x + 1$. Namely, they differ by only a scalar multiple. In fact, we will obtain uniqueness of factorization for polynomials, up to scalar multiples. We now head toward this result. The first order of business is to define irreducible polynomials in a way analogous to our definition of irreducible integers.

A polynomial p is **irreducible** if

a. p is of degree greater than zero, and

b. whenever $p = fg$, then either f or g has degree zero.

In other words, an irreducible is a non-scalar polynomial, whose only factorizations involve scalar factors. We are thus regarding such factorizations as $x + \frac{1}{2} = \frac{1}{2}(2x+1)$ as trivial, just as we regard factorizations such as $3 = (-1)(-3)$ as trivial in the integer case. So, a *non-trivial factorization* of a polynomial is one that has at least two non-scalar factors. We say that a polynomial is **reducible** if it does have a non-trivial factorization. Thus, $x^4 + 2x^2 + 1$ is reducible.

▷ **Quick Exercise.** Why is $x^4 + 2x^2 + 1$ reducible? ◁

Which polynomials in $\mathbb{Q}[x]$ are irreducible? One immediate consequence of the definition is that all polynomials of degree one are irreducible because if p is of degree 1 and $p = fg$, then exactly one of f and g has degree 0, and is consequently a scalar.

Are there any others? Consider the polynomial $f = x^2 + 2$. If this polynomial had a non-trivial factorization $x^2 + 2 = gh$, then both g and h would be degree one factors. But then the Root Theorem tells us that $x^2 + 2$ would have a rational root. But $f(q) = q^2 + 2 \geq 2$ for all rational numbers q, and so f can have no roots.

Consider next the polynomial $x^2 - 2$. By the same reasoning, if this polynomial were reducible, it would have a root, and so there would exist a rational number q so that $q^2 = 2$. There is no such rational number, as you probably know. The fact that $\sqrt{2}$ is an irrational number is a very famous theorem, first proved by the ancient Greeks. You will prove it (and more) in Exercise 5.13 (see also Exercise 2.14). Thus, $x^2 - 2$ is irreducible.

▷ **Quick Exercise.** Show that $x^4 + 2$ is irreducible in $\mathbb{Q}[x]$, taking your lead from the discussion of $x^2 + 2$ above. ◁

▷ **Quick Exercise.** Show that $x^3 - 2$ is irreducible in $\mathbb{Q}[x]$. *Hint:* If $x^3 - 2$ were reducible, then it would have a linear factor. You may assume (see Exercise 5.13) that there is no rational number r so that $r^3 = 2$. ◁

These examples suggest that there are many irreducible polynomials in $\mathbb{Q}[x]$, and indeed there are, of arbitrarily high degree. (See Exercise 5.12.) It would be difficult to describe them all, however.

Note that we are interested only in factors which belong to $\mathbb{Q}[x]$ (for the time being, at least). Thus, $x^2 - 2$ is irreducible in $\mathbb{Q}[x]$, even though we *can* factor it if we allow ourselves to use real numbers as coefficients:

$$x^2 - 2 = \left(x - \sqrt{2}\right)\left(x + \sqrt{2}\right).$$

In Chapter 9 we will discuss factorization of polynomials over the real numbers $\mathbb{R}$ and the complex numbers $\mathbb{C}$.

We now claim that every polynomial can be factored into irreducibles:

Theorem 5.1 *Any polynomial in $\mathbb{Q}[x]$ of degree greater than zero is either irreducible or the product of irreducibles.*

Proof: Prove this in a similar way to the proof for the corresponding theorem for the integers, Theorem 2.8. This proof is Exercise 5.1. □

5.2 Unique Factorization

The previous theorem is the first half of the unique factorization theorem that we want for $\mathbb{Q}[x]$. Recall that to obtain the uniqueness of factorization in $\mathbb{Z}$, we required the concept of a *prime* integer. We now make the analogous definition: a polynomial p (with degree bigger than 0) is **prime** if whenever p divides fg, then p divides f or p divides g.

We claim that $x-2$ is a prime polynomial. Suppose that $x-2$ divides the product fg. Then by the Root Theorem 4.3, 2 is a root of fg, and so $f(2)g(2) = 0$. But $f(2)$ and $g(2)$ are rational numbers, and the only way a product of rational numbers can be zero is if at least one of the factors is zero. That is, $f(2)$ (or $g(2)$) is zero, and so 2 is a root of f (or g). Notice that this argument could be modified to show that *any* degree one polynomial is prime.

▷ **Quick Exercise.** Modify the argument in the previous paragraph to prove that all degree one polynomials in $\mathbb{Q}[x]$ are prime. ◁

But we need not pursue any further examples because the concept of prime polynomial turns out to be equivalent to that of irreducible polynomial. This situation is what we discovered for $\mathbb{Z}$, and the proof is the same:

Theorem 5.2 *A polynomial in $\mathbb{Q}[x]$ is irreducible if and only if it is prime.*

Proof: Prove this, again taking your lead from the proof of Theorem 2.7, the analogous result for the integers. This proof is Exercise 5.2. □

Corollary 5.3 *If p is an irreducible polynomial that divides*

$$f_1 f_2 \cdots f_n,$$

then p divides one of the f_i.

Proof: Prove this. This proof is Exercise 5.3. □

So, for both $\mathbb{Z}$ and $\mathbb{Q}[x]$, primeness and irreducibility are equivalent. Be warned that we will eventually examine structures where this is not the case. It is thus important to keep these definitions straight.

Finally, we come to the unique factorization theorem for polynomials. It is similar to the unique factorization theorem for integers, except we must account for the fact that we can always factor out scalars, as noted above. Accordingly, we first make the following convenient definition.

Two polynomials f and g are called **associates** if there is a non-zero scalar a such that $f = ag$. For instance, $x^2 - 3$ and $2x^2 - 6$ are associates. Note that any two non-zero scalars are associates.

▷ **Quick Exercise.** Describe the set of all associates of the polynomial $x^2 - 3$ (there are infinitely many). ◁

Theorem 5.4 Unique Factorization Theorem for Polynomials *If h is a polynomial in $\mathbb{Q}[x]$, and*

$$h = f_1 f_2 \cdots f_n = g_1 g_2 \cdots g_m,$$

where the f_i and g_j are all irreducible, then $n = m$ and the g_j may be rearranged so that f_i and g_i are associates, for $i = 1, 2, \ldots n$.

Before we prove this theorem, let's look at an example. Consider the polynomial $h = 2x^3 + 3x^2 + 5x + 2$. Now h can be factored into irreducibles as $(2x + 1)(x^2 + x + 2)$ or as $(x + \frac{1}{2})(2x^2 + 2x + 4)$, or an infinite number of other ways. But, of course, $2x + 1$ and $x + \frac{1}{2}$ are associates, as are $x^2 + x + 2$ and $2x^2 + 2x + 4$.

▷ **Quick Exercise.** Verify that the factors of h given above are really irreducible. ◁

Proof: We use induction on n, the number of factors f_i of h. If $n = 1$, the theorem follows easily. (Right?) So, we assume $n > 1$. By the primeness of the irreducible f_1, f_1 divides one of the g_j (see the Corollary 5.3). By renumbering the g_j, if necessary, we may assume f_1 divides g_1. So, because g_1 is irreducible, $af_1 = g_1$, for some non-zero scalar a. That is, f_1 and g_1 are associates. Therefore, by dividing both sides by f_1, we have $f_2 f_3 \cdots f_n = a g_2 g_3 \cdots g_m$. (Because g_2 is irreducible, so is ag_2, and we consider ag_2 as an irreducible factor.) We now have two factorizations into irreducibles and the number of f_i factors is $n - 1$. So, by the induction hypothesis, $n - 1 = m - 1$ and, by renumbering the g_j if necessary, f_i and g_i are associates for $i = 2, 3, \ldots, n$. (Actually, we have that ag_2 is an associate of f_2. But then g_2 is also an associate of f_2.) This proves the theorem. □

Notice that the proof of this theorem is nearly identical to the proof of the corresponding Theorem 2.7 for $\mathbb{Z}$. We had only to handle the problem of associates. By now, you should have seen this similarity with *all* the theorems in this chapter and might be wondering the reason for it. We will eventually find a close connection between the integers and $\mathbb{Q}[x]$ that explains this similarity.

5.3 Polynomials with Integer Coefficients

We close this chapter by discussing the relationship between $\mathbb{Q}[x]$ and $\mathbb{Z}[x]$. You should first note that some of the theorems we have proved about $\mathbb{Q}[x]$ are *false* when we restrict ourselves to polynomials with integer coefficients. In particular, the Division Theorem is false for $\mathbb{Z}[x]$. To see this, merely try to divide $x^2 + 7x$ by $2x + 1$: the Division Theorem 4.2 for $\mathbb{Q}[x]$ provides us with the unique quotient $\frac{1}{2}x + \frac{11}{4}$ and remainder $-\frac{11}{4}$. These are not elements of $\mathbb{Z}[x]$, and so no such quotient/remainder pair can exist in $\mathbb{Z}[x]$. (See also Exercise 4.2).

Similarly, the GCD identity fails in $\mathbb{Z}[x]$. Consider the polynomials 2 and x; a gcd for these polynomials is 1. But we *cannot* write 1 as a linear combination of 2 and x in $\mathbb{Z}[x]$.

▷ **Quick Exercise.** Show that we cannot write 1 as a linear combination of 2 and x. *Hint*: Suppose that $1 = 2f + xg$, where $f, g \in \mathbb{Z}[x]$, and consider the constant term of the right-hand side of the equation. ◁

It is thus not surprising that we look to $\mathbb{Q}[x]$ (rather than $\mathbb{Z}[x]$) for our analogue to $\mathbb{Z}$. However, the news is not all bad. For it turns out that if a polynomial with integer coefficients can be non-trivially factored into polynomials with rational coefficients, then it can be factored into polynomials with integer coefficients; this result is known as Gauss's Lemma.

For example, consider $2x^2+3x-2$. Because $\frac{1}{2}$ is a root, we have that

$$2x^2+3x-2 = \left(x-\frac{1}{2}\right)(2x+4).$$

This is a factorization in $\mathbb{Q}[x]$. But by adjusting scalars, we can obtain the factorization

$$2x^2+3x-2 = (2x-1)(x+2)$$

in $\mathbb{Z}[x]$.

The idea of the proof below is essentially that of the example we've just given, although it is a bit messy to carry it out in general. Some readers may want to skip this proof on a first reading.

Theorem 5.5 Gauss's Lemma *If $f \in \mathbb{Z}[x]$ and f can be factored into a product of non-scalar polynomials in $\mathbb{Q}[x]$, then f can be factored into a product of non-scalar polynomials in $\mathbb{Z}[x]$; each factor in $\mathbb{Z}[x]$ is an associate of the corresponding factor in $\mathbb{Q}[x]$.*

Proof: Suppose $f = gh$, where $g, h \in \mathbb{Q}[x]$. We can assume that the coefficients of f have no common factors (other than ±1), or else we factor out the greatest common divisor of those coefficients, apply the following, and then multiply through by the gcd. So, suppose that

$$g = \frac{a_n}{b_n}x^n + \frac{a_{n-1}}{b_{n-1}}x^{n-1} + \cdots + \frac{a_0}{b_0}, \text{ and}$$
$$h = \frac{c_m}{d_m}x^m + \frac{c_{m-1}}{d_{m-1}}x^{m-1} + \cdots + \frac{c_0}{d_0}$$

where the a_i's, b_i's, c_i's, and d_i's are all integers and each fraction is in lowest terms. Multiply the first equation by the product B of the b_i's. We get

$$Bg = A'_n x^n + A'_{n-1}x^{n-1} + \cdots + A'_0.$$

Now divide by the greatest common divisor A of the A'_i's, yielding

$$(B/A)g = A_n x_n + A_{n-1}x^{n-1} + \cdots + A_0,$$

which is a polynomial in $\mathbb{Z}[x]$ whose coefficients have no common divisor (other than ±1).

Do likewise to the second equation, yielding

$$(D/C)h = C_m x^m + C_{m-1}x^{m-1} + \cdots + C_0,$$

which again is a polynomial whose coefficients have no common divisor (other than ±1).

Because $f = gh$, we get

$$BDf = (AC)((B/A)g)((D/C)h). \tag{5.1}$$

Because the coefficients of f have no common divisor, we see that BD is the gcd of the coefficients of BDf. Because AC is a common divisor of the coefficients of this polynomial (as we see by looking at the right side of equation 5.1) we have that $AC|BD$. Let $E = BD/AC$, which is an integer. We have

$$Ef = (A_n x^n + A_{n-1}x^{n-1} + \cdots + A_0)(C_m x^m + C_{m-1}x^{m-1} + \cdots + C_0). \tag{5.2}$$

We need only show that $E = \pm 1$ to be done, because then f has been written in the required form. If $E \neq \pm 1$, then let p be a prime factor of E. Remember that all the A_i's have no common factor and all the C_i's have no common factor. Therefore, there is a smallest i such that $p \nmid A_i$ and a smallest j such that $p \nmid C_j$. We now compute the coefficient of x^{i+j}. From the left side of equation 5.2 we see that this coefficient is divisible by E, and hence by p. From the right side, the coefficient is

$$\sum_{k=0}^{i-1} A_k C_{i+j-k} + \sum_{k=i+1}^{i+j} A_k C_{i+j-k} + A_i C_j.$$

In the first sum, each of A_k is divisible by p because those k are less than i. In the second sum, each C_{i+j-k} is divisible by p because those $i+j-k$ are less than j. But $A_i C_j$ is not divisible by p. Hence, the entire coefficient cannot be divisible by p, which is a contradiction. So it must be that $E = \pm 1$ as desired. □

We can rephrase Gauss's Lemma in the following way: To see whether a polynomial in $\mathbb{Z}[x]$ can be factored non-trivially in $\mathbb{Q}[x]$, we need only check to see if it can be factored non-trivially in $\mathbb{Z}[x]$. This latter is presumably an easier task.

For example, consider the case of cubic equations. If a cubic polynomial in $\mathbb{Z}[x]$ can be factored non-trivially in $\mathbb{Z}[x]$, one of the factors must be a degree one polynomial $ax + b$, where a and b are integers. This implies that the polynomial must have a root $-\frac{b}{a}$ in $\mathbb{Q}$. But we can limit the possibilities for a and b, which in turn limits which roots are possible. For a specific example of this, consider the polynomial $3x^3 + x + 1$. We wish to factor this into irreducibles in $\mathbb{Q}[x]$ or determine if the polynomial is itself irreducible. If $3x^3 + x + 1$ factors in $\mathbb{Q}[x]$, then, by Gauss's Lemma, it factors in $\mathbb{Z}[x]$ and in fact must have a linear factor in $\mathbb{Z}[x]$. The only possible such factors are of the form $(3x \pm 1)$ or $(x \pm 1)$, in order that the leading and trailing coefficients be correct. But this implies that the polynomial has a root of $\pm 1/3$ or ± 1, which, by inspection, is not the case. We conclude that $3x^3 + x + 1$ is irreducible in $\mathbb{Z}[x]$, and hence in $\mathbb{Q}[x]$.

The argument above can be generalized to obtain a valuable tool for factoring in $\mathbb{Z}[x]$, known as the *Rational Root Theorem.*

Theorem 5.6 The Rational Root Theorem *Suppose that $f = a_0 + a_1x + \cdots + a_nx^n$ is a polynomial in $\mathbb{Z}[x]$, and p/q is a rational root; that is, p and q are integers, $q \neq 0$, and $f(p/q) = 0$. We may as well assume also that* $\gcd(p, q) = 1$. *Then q divides the integer a_n, and p divides a_0.*

Proof: You will prove this in Exercise 5.6. □

▷ **Quick Exercise.** Determine which of the following polynomials has a rational root:

$$3x^4 + 5x^3 + 10, \quad 4x^4 + x^3 - 4x^2 + 7x + 2.$$

◁

▷ **Quick Exercise.** Is $3x^3+2x+1$ irreducible in $\mathbb{Q}[x]$? Is x^3-3x+4? Is $2x^3 + 7x^2 - 2x - 1$? ◁

Another important tool when considering factorization is the following sufficient condition for a polynomial being irreducible in $\mathbb{Z}[x]$.

Theorem 5.7 Eisenstein's Criterion *Suppose that $f \in \mathbb{Z}[x]$, and*

$$f = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n.$$

Let p be a prime integer, and suppose that

1. *p divides a_k, for $0 \leq k < n$,*
2. *p does not divide a_n, and*
3. *p^2 does not divide a_0.*

Then f is irreducible in $\mathbb{Z}[x]$.

Proof: Once again, you will prove this theorem in Exercise 5.16. □

Example 5.1

It is quite evident that the Eisenstein criterion implies that the polynomial $x^5 + 5x + 5$ is irreducible in $\mathbb{Z}[x]$, by using $p = 5$. But note that the criterion does not directly apply to the polynomial $x^5 + 5x + 4$. However, you will discover in Exercise 5.15 how to apply the criterion to show that this is in fact an irreducible polynomial.

Notice that although Eisenstein's criterion is phrased as a theorem about irreducibility in $\mathbb{Z}[x]$, Gauss's Lemma 5.5 implies that such a polynomial is also irreducible in $\mathbb{Q}[x]$.

Example 5.2

Consider the polynomial

$$f = \frac{1}{3}x^4 + \frac{2}{9}x^3 + 4x^2 - \frac{2}{5}x + 2 \in \mathbb{Q}[x].$$

We wish to conclude that f is irreducible in $\mathbb{Q}[x]$. By multiplying through by 45, we obtain the element

$$g = 45f = 15x^4 + 10x^3 + 140x^2 - 18x + 2 \in \mathbb{Z}[x].$$

Note that we could apply the Rational Root Theorem to g to conclude that it has no roots in $\mathbb{Q}$; it would be tedious to carry out the details. However, we can in fact apply Eisenstein's Criterion (with $p = 2$) to reach the strictly stronger conclusion that g is irreducible in $\mathbb{Z}[x]$. But then Gauss's Lemma says that g is in fact irreducible in $\mathbb{Q}[x]$. But then f is irreducible in $\mathbb{Q}[x]$ too.

Historical Remarks

In the last two chapters we have given a systematic account of the theory of $\mathbb{Q}[x]$, the set of polynomials with rational coefficients, giving due emphasis to its similarity to $\mathbb{Z}$. We have consequently emphasized a formal, algebraic approach to $\mathbb{Q}[x]$, which probably seems foreign to your previous experience with polynomials.

The pieces of this theory were put together over a number of centuries, beginning in earnest with the 17th-century French algebraists Descartes and Viète, and culminating in the work of Gauss. This development was relatively slow, primarily for two reasons. First of all, algebraic notation at the time of Viète was cumbersome and not standardized. Algebra flows much more easily when our notation is clear and efficient. Secondly, in the 17th century people had little agreement about the nature of numbers. Doubts were cast on the 'reality' and utility of irrational numbers, complex numbers, and even negative numbers. We'll discuss this issue more in Chapter 9. The process by which the number system which we use was standardized and widely accepted was long and difficult. The lesson from history is clear: It takes a lot of hard work to become comfortable with the elegant point of view of a modern mathematician!

Chapter Summary

In this chapter we considered factorization in $\mathbb{Q}[x]$ and proved that a polynomial is irreducible if and only if it is prime, and we used this fact to prove the *Unique Factorization Theorem for Polynomials.* We then proved *Gauss's Lemma*, which describes the relationship between factoring in $\mathbb{Z}[x]$ and in $\mathbb{Q}[x]$. We concluded by considering two important tools useful in factoring in $\mathbb{Z}[x]$, called the *Rational Root Theorem* and *Eisenstein's Criterion.*

Warm-up Exercises

a. Why is a linear polynomial in $\mathbb{Q}[x]$ always irreducible?

b. Why is a polynomial of the form $x^2 + a \in \mathbb{Q}[x]$, where $a > 0$, always irreducible?

c. Determine a factorization of $x^4 - 5x^2 + 4$ into irreducibles in $\mathbb{Q}[x]$.

d. Give several distinct factorizations of $x^4 - 5x^2 + 4$ into irreducibles in $\mathbb{Q}[x]$. Why don't these distinct factorizations violate the Unique Factorization Theorem 5.4?

e. We know that 7 is an irreducible integer, but is 7 an irreducible polynomial?

f. Pick your favorite polynomial. What are its associates in $\mathbb{Q}[x]$?

g. Factor $2x^3 + 7x^2 - 2x - 1$ completely into irreducibles in $\mathbb{Q}[x]$, using Gauss's Lemma and the Root Theorem. Adjust your factorization (if necessary) so that all factors belong to $\mathbb{Z}[x]$.

Exercises

1. Prove Theorem 5.1: A polynomial in $\mathbb{Q}[x]$ of degree greater than zero is either irreducible or the product of irreducibles.

2. Prove Theorem 5.2: A polynomial in $\mathbb{Q}[x]$ is irreducible if and only if it is prime.

3. Prove Corollary 5.3: If an irreducible polynomial in $\mathbb{Q}[x]$ divides a product $f_1 f_2 f_3 \cdots f_n$, then it divides one of the f_i.

4. Use Gauss's Lemma to determine which of the following are irreducible in $\mathbb{Q}[x]$:
$$4x^3 + x - 2, \quad 3x^3 - 6x^2 + x - 2, \quad x^3 + x^2 + x - 1.$$

5. Show that $x^4 + 2x^2 + 4$ is irreducible in $\mathbb{Q}[x]$.

6. Prove the Rational Root Theorem 5.6.

7. Use the Rational Root Theorem 5.6 to factor
$$2x^3 - 17x^2 - 10x + 9.$$

8. Use the Rational Root Theorem 5.6 to argue that
$$x^3 + x + 7$$
is irreducible over $\mathbb{Q}[x]$. Use elementary calculus to argue that this polynomial does have exactly one *real* root.

9. Use the Rational Root Theorem 5.6 (applied to $x^3 - 2$) to argue that $\sqrt[3]{2}$ is irrational.

10. Suppose that α is a real number (which might not be rational), and suppose that it is a root of a polynomial $p \in \mathbb{Q}[x]$; that is, $p(\alpha) = 0$. Suppose further that p is irreducible in $\mathbb{Q}[x]$. Prove that p has minimal degree in the set
$$\{f \in \mathbb{Q}[x] : f(\alpha) = 0 \text{ and } f \neq 0\}.$$

11. This is a continuation of Exercise 10. Suppose as above that α is a real number, $p \in \mathbb{Q}[x]$ is irreducible, and $p(\alpha) = 0$. Suppose also that $f \in \mathbb{Q}[x]$ with $f(\alpha) = 0$. Prove that p divides f.

12. Construct polynomials of arbitrarily large degree, which are irreducible in $\mathbb{Q}[x]$.

13. (a) Prove that the equation $a^2 = 2$ has no rational solutions; that is, prove that $\sqrt{2}$ is irrational. (This part is a repeat of Exercise 2.14.)

    (b) Generalize part a, by proving that $a^n = 2$ has no rational solutions, for all positive integers $n \geq 2$.

14. Let $f \in \mathbb{Z}[x]$ and n an integer. Let g be the polynomial defined by $g(x) = f(x+n)$. Prove that f is irreducible in $\mathbb{Z}[x]$ if and only if g is irreducible in $\mathbb{Z}[x]$.

15. (a) Apply Eisenstein's criterion 5.7 to check that the following polynomials are irreducible:
$$5x^3 - 6x^2 + 2x - 14 \text{ and } 4x^5 + 5x^3 - 15x + 20.$$

    (b) Make the substitution $x = y + 1$ to the polynomial $x^5 + 5x + 4$ that appears in Example 5.1. Show that the resulting polynomial is irreducible. Now conclude that the original polynomial is irreducible.

    (c) Use the same technique as in part b to find a substitution $x = y + m$ so you can conclude the polynomial
$$x^4 + 6x^3 + 12x^2 + 10x + 5$$
is irreducible.

    (d) Show that this technique works in general: Prove that if $f(x) \in \mathbb{Z}[x]$, then $f(x)$ is irreducible if and only if $f(y+m)$ is.

16. Prove Theorem 5.7 (Eisenstein's criterion).

17. Let p be a positive prime integer. Then the polynomial
$$\Phi_p = \frac{x^p - 1}{x - 1}$$
is called a **cyclotomic polynomial**.

    (a) Write out, in the usual form for a polynomial, the cyclotomic polynomials for the first three primes.

    (b) Prove that all cyclotomic polynomials Φ_p are irreducible over $\mathbb{Z}[x]$, using Eisenstein's criterion 5.7 and Exercise 15d for $m = 1$.

Section I in a Nutshell

This section examines the integers ($\mathbb{Z}$), the integers modulo m ($\mathbb{Z}_m$), and polynomials with rational coefficients ($\mathbb{Q}[x]$). These structures share many algebraic properties:

- Each has addition defined and the addition is commutative.
- Each has multiplicative defined and the multiplication is commutative.
- Each has an additive identity (0 for $\mathbb{Z}$, $[0]$ for $\mathbb{Z}_m$, and the zero polynomial for $\mathbb{Q}[x]$).
- Each has a multiplicative identity (1 for $\mathbb{Z}$, $[1]$ for $\mathbb{Z}_m$, and the polynomial 1 for $\mathbb{Q}[x]$).
- All elements have additive inverses, but not all elements have multiplicative identities.

Furthermore, $\mathbb{Z}$ and $\mathbb{Q}[x]$ have some notion of 'size'. The size of an integer is given by its absolute value, while the size of a polynomial is given by its degree.

This notion of size along with their similar algebraic properties, allow us to prove a series of parallel theorems for $\mathbb{Z}$ and $\mathbb{Q}[x]$:

(Theorem 2.1) *Division Theorem for* $\mathbb{Z}$: Let $a, b \in \mathbb{Z}$ with $a \neq 0$. Then there exist unique integers q and r with $0 \leq r < |a|$ such that $b = aq+r$.

(Theorem 4.2) *Division Theorem for* $\mathbb{Q}[x]$: Let $f, g \in \mathbb{Q}[x]$ with $f \neq 0$. Then there are unique polynomials q and r with $\deg(r) < \deg(f)$ such that $g = fq + r$.

These Division Theorems allow us to develop Euclid's Algorithm, which is a method to compute the gcd of two integers (Theorem 2.3) or polynomials (Theorem 4.5).

We then developed the notion of an irreducible integer and irreducible polynomial: An integer (polynomial) p is *irreducible* if $\neq \pm 1$ ($\deg(p) >$

0) and whenever $p = ab$, then either a or b is ± 1 (either a or b has degree 0). We were then able to prove the *Fundamental Theorem of Arithmetic*, and its analogue for polynomials: every integer (other than 0, 1 or -1) is irreducible or a product of irreducibles (Theorem 2.8). Similarly, any polynomial in $\mathbb{Q}[x]$ of degree greater than 0 is either irreducible or the product of irreducibles (Theorem 5.1). Both of these factorizations are unique:

(Theorem 2.9) If an integer $x = a_1 a_2 \cdots a_n = b_1 b_2 \cdots b_m$ where the a_i and b_j are all irreducible, then $n = m$ and the b_j may be rearranged so that $a_i = \pm b_i$ for $i = 1, 2, \ldots, m$.

(Theorem 5.4) If the polynomial $p = f_1 f_2 \cdots f_n = g_1 g_2 \cdots g_m$ where the f_i and g_j are all irreducible, then $n = m$ and the g_j may be rearranged so that f_i and g_j are associates (that is, non-zero scalar multiples of one another), for $i = 1, 2, \ldots, n$.

There is also a shared notion of primeness: An integer (polynomial) is *prime* if p is not 0, 1 or -1 ($\deg(p) > 0$) and whenever p divides ab, then either p divides a or p divides b. For $\mathbb{Z}$ and $\mathbb{Q}[x]$, the idea of primeness and irreducibility are one and the same: (Theorems 2.7 and 5.2). An integer (polynomial) is irreducible if and only if it is prime.

Finally, we examined the polynomials with integer coefficients, $\mathbb{Z}[x]$. Unfortunately, there is no Division Theorem for $\mathbb{Z}[x]$ and no GCD identity. However, $\mathbb{Z}[x]$ does have some interesting factorization properties:

(Theorem 5.5) *Gauss's Lemma* If $f \in \mathbb{Z}[x]$ and f can be factored into a product of non-scalar polynomials in $\mathbb{Q}[x]$, then f can be factored into a product of non-scalar polynomials in $\mathbb{Z}[x]$.

(Theorem 5.6) *Rational Root Theorem* Suppose $f = a_0 + a_1 x + \cdots + a_n x^n \in \mathbb{Z}[x]$ and $p, q \in \mathbb{Z}$ with $q \neq 0$, $\gcd(p, q) = 1$ and $f(p/q) = 0$. Then q divides a_n and p divides a_n.

(Theorem 5.7) *Eisenstein's Criterion* Suppose $f = a_0 + a_1 x + \cdots + a_n x^n \in \mathbb{Z}[x]$ and p is a prime where (1) p divides a_k for $0 \leq k < n$, (2) p does not divide a_n, and (3) p^2 divides a_0. Then f is irreducible in $\mathbb{Q}[x]$.

II

Rings, Domains, and Fields

Chapter 6

Rings

In the previous chapters we have examined several different algebraic objects: $\mathbb{Z}$, $\mathbb{Q}[x]$, and $\mathbb{Z}_m$ (for integers $m > 1$). You are also probably aquainted with the larger sets of real numbers, $\mathbb{R}$, and complex numbers, $\mathbb{C}$. In each of these cases we have a set of elements, which is equipped with two operations called *addition* and *multiplication*. Each of these is an example of an abstract concept called a *ring*. In this chapter we will give an abstract definition of this concept and look at some basic properties and examples.

6.1 Binary Operations

Before we can do this, we must understand better what we mean by an 'operation' defined on a set. A **binary operation** on a set S is a function $\circ : S \times S \mapsto S$, where

$$S \times S = \{(s,t) : s,t \in S\}$$

is the set of all ordered pairs with entries from S. Thus, $\circ$ is a function that takes ordered pairs of elements from S to elements of S.

If you think about it, that is exactly what an operation like addition (on $\mathbb{Z}$) does: It takes an ordered pair of elements (such as $(4,6)$), and assigns to that pair another element ($4 + 6 = 10$).

Because we wish to make this function $\circ$ look like our more familiar operations like addition or multiplication, we write the image of an element (s,t) in $S \times S$, under the function $\circ$ as $s \circ t$.

Thus, addition and multiplication on the set $\mathbb{Z}$ (or for that matter on $\mathbb{Z}_m$, $\mathbb{Q}[x]$, $\mathbb{Q}$, or even the natural numbers $\mathbb{N}$) are binary operations. Notice that although in our discussion above we were using the notation $\circ$ for our generic binary operation, we are perfectly happy to denote addition by + as usual. Subtraction is a binary operation on the first four of these sets but is *not* a binary operation on $\mathbb{N}$ because the function $-$ is not defined on such pairs as $(3, 4)$.

Similarly, division is not a binary operation on any of our sets considered so far because of zero: The function $\div$ has no image defined for such ordered pairs as $(1, 0)$. However, $\div$ is a binary operation on the sets $\mathbb{Q}^+$ of strictly positive rational numbers, and $\mathbb{Q}^*$ of non-zero rational numbers.

For a rather different example of a binary operation, consider any nonempty set S and define $s \circ t = s$. That is, $\circ$ assigns the first entry to any ordered pair it's given.

▷ **Quick Exercise.** Which of the following are binary operations?

1. Matrix multiplication, on the set of all 2×2 matrices.
2. $a \circ b = a + b + ab$, on the set $\mathbb{Z}$.
3. Dot product, on the set of all vectors in the plane.
4. Cross product, on the set of all vectors in space.
5. $A \cap B$, on the set of all subsets of $\{1, 2, 3, 4\}$.
6. $a \circ b = \sqrt{ab}$, on the set $\mathbb{R}$.
7. $a \circ b = \sqrt{ab}$, on the set $\mathbb{R}^+$, the set of all positive real numbers. ◁

Note the crucial importance of both the set on which the operation is defined, as well as the operation itself. A function can be a binary operation only if it gives a value for all possible ordered pairs in the set. We often say a set is **closed** under an operation if this is the case. Thus, $\mathbb{N}$ is closed under addition and multiplication but is not closed under subtraction or division.

You might already suspect that because the definition of binary operation is so general, we should reserve terms like 'addition' and 'subtraction' for very special binary operations which obey nice rules. This is precisely what we do when we define a ring.

6.2 Rings

A **ring** R is a set of elements on which two binary operations, addition $(+)$ and multiplication $(\cdot)$, are defined that satisfy the following properties. (The symbols a, b, and c represent any elements from R.)

(Rule 1) $a + b = b + a$

(Rule 2) $(a + b) + c = a + (b + c)$

(Rule 3) There exists an element 0 in R such that $a + 0 = a$

(Rule 4) For each element a in R, there exists an element x such that $a + x = 0$

(Rule 5) $(a \cdot b) \cdot c = a \cdot (b \cdot c)$

(Rule 6) $a \cdot (b + c) = a \cdot b + a \cdot c\ ,\ \ (b + c) \cdot a = b \cdot a + c \cdot a$

Let's introduce some terminology to describe these rules: Rule 1 says that addition is **commutative**, and Rules 2 and 5 say that addition and multiplication, respectively, are **associative**. Rule 3 says an **additive identity** (or **zero**) exists, and Rule 4 says that each element of the ring has an **additive inverse**. Finally, Rule 6 says that multiplication **distributes** over addition on the right and the left.

We will usually write ab instead of $a \cdot b$.

What are some examples of rings?

Example 6.1

The integers $\mathbb{Z}$, equipped with the usual addition and multiplication. The properties listed above should all be familiar facts about arithmetic in the integers.

Example 6.2

The rational numbers $\mathbb{Q}$, with the usual addition and multiplication. Once again, we rely on our previous experience with arithmetic to check that all these properties hold.

Example 6.3

$\mathbb{Z}_6$. In Chapter 3 we constructed what we mean by this set and defined operations + and $\cdot$ on it. Let's check that addition in $\mathbb{Z}_6$ is commutative: By the definition of addition in $\mathbb{Z}_6$, $[a]_6 + [b]_6 = [a + b]_6$. But because addition in $\mathbb{Z}$ is commutative, $[a + b]_6 = [b + a]_6$. And so (again by the definition of addition), $[a]_6 + [b]_6 = [b + a]_6 = [b]_6 + [a]_6$, as required.

The proof we just performed was admittedly tedious. Note that it could have been paraphrased as follows: Addition in $\mathbb{Z}_6$

is defined in terms of addition in $\mathbb{Z}$, and because addition in $\mathbb{Z}$ is commutative, so is addition in $\mathbb{Z}_6$.

▷ **Quick Exercise.** Check that the other five ring axioms are satisfied by $\mathbb{Z}_6$. Note that we have already discussed the existence of the additive identity and additive inverses for $\mathbb{Z}_6$ in Chapter 3. ◁

Example 6.4

$\mathbb{Z}_m$, for any integer $m > 1$, with the addition and multiplication defined in Chapter 3. The proofs in Example 6.3 certainly work for any m.

Example 6.5

The set $\{0\}$, where $0 + 0 = 0 \cdot 0 = 0$. This is the world's most boring ring, called the **zero ring**. Because our set has only a single element, and *any* computation we perform gives us 0, all six axioms must certainly hold.

Example 6.6

$\mathbb{Q}[x]$, with the addition and multiplication we defined in Chapter 4. When we add polynomials, we just add corresponding coefficients (x^2 terms added together, etc.). But then, because addition of rational numbers is commutative and associative, it follows that addition of polynomials is commutative and associative. The polynomial 0 clearly plays the role of the additive identity in $\mathbb{Q}[x]$, and we can obtain an additive inverse for any polynomial by changing the sign of every term (this amounts to just multiplying by -1).

Thus, to show that $\mathbb{Q}[x]$ is a ring, it remains only to prove that multiplication is associative, and that it distributes over addition. Because the multiplication of polynomials is difficult to describe formally (even though it is very familiar), formal proofs that Rules 5 and 6 are satisfied by $\mathbb{Q}[x]$ are exceedingly tedious. We consequently omit them here, and refer you to Exercises 6.21 and 6.22.

Example 6.7

$\mathbb{Z}[x]$, with the addition and multiplication as in Chapter 5. We first note that whenever we add or multiply two polynomials with integer coefficients, we get another one. That is, this set is closed under addition and multiplication. Because the rational polynomials satisfy Rules 1, 2, 5, and 6, it is quite evident that $\mathbb{Z}[x]$ does too because $\mathbb{Z}[x] \subseteq \mathbb{Q}[x]$. Rule 3 holds because the polynomial $0 \in \mathbb{Z}[x]$, and Rule 4 holds because multiplying a polynomial with integer coefficients by -1 gives more integer coefficients.

Example 6.8

The set of all even integers, which we abbreviate as $2\mathbb{Z}$, together with ordinary addition and multiplication.

▷ **Quick Exercise.** Use an argument modelled on that we used for Example 6.7 to show that $2\mathbb{Z}$ is a ring. ◁

Example 6.9

Let $\mathbb{Z} \times \mathbb{Z}$ be the set of ordered pairs with integer entries. That is,

$$\mathbb{Z} \times \mathbb{Z} = \{(a, b) : a, b \in \mathbb{Z}\}.$$

Define addition and multiplication *point-wise*; that is,

$$\begin{aligned}(n, m) + (r, s) &= (n + r, m + s)\\ (n, m) \cdot (r, s) &= (nr, ms).\end{aligned}$$

Then $\mathbb{Z} \times \mathbb{Z}$ is a commutative ring. You will verify the details in Exercise 6.13.

Example 6.10

Let R and S be arbitrary rings, and let $R \times S$ be the set of ordered pairs with first entry from R, and second entry from S. Then if we define addition and multipication point-wise (as in Example 6.9), we have created a new ring, called the **direct product** of the rings R and S. You will check the details, and further generalize this, in Exercise 6.15.

▷ **Quick Exercise.** Write out the addition and multiplication tables for the direct product ring $\mathbb{Z}_2 \times \mathbb{Z}_3$. ◁

It might also be worthwhile to provide a few examples of sets equipped with two operations, which are not rings:

Example 6.11

The set $\mathbb{N}$ of natural numbers, equipped with the usual addition and multiplication. This structure satisfies Rules 1, 2, 5, and 6, but Rules 3 and 4 are false.

▷ **Quick Exercise.** Check that these assertions are true. ◁

Example 6.12

The set $\mathbb{Z}$, with the usual addition, and the operation $\circ$ defined by $a \circ b = a$. This structure satisfies the first five rules. Furthermore, $\circ$ distributes over addition from the right, because

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

However, $\circ$ does not distribute over addition from the left. To see this, we need only provide an example:

$$2 \circ (3+4) = 2 \neq 2 + 2 = 2 \circ 3 + 2 \circ 4.$$

Thus, $\mathbb{Z}$ equipped with these operations is not a ring.

Let us now look a little more carefully at the rules determining a ring. Rules 1, 2, 5, and 6 specify that addition and multiplication are to satisfy certain nice properties (thus distinguishing them from general binary relations). These properties are universal statements applying to all elements of the ring. Thus, addition is required to be commutative and associative, multiplication is to be associative, and multiplication should distribute over addition. Note that we do not require that multiplication be commutative. If that is the case, we say that we have a **commutative ring**. The rings in Examples 6.1–6.9 are all commutative. Example 6.13 below is a non-commutative ring.

Rule 3 is quite different. It asserts that an element of a particular kind (the **additive identity**) exists in the ring. Without Rule 3, the empty set would qualify as a ring. Rule 4 specifies that additive inverses exist for each element we find in the ring, but it certainly does not require that a ring have any other elements than 0, because $0 + 0 = 0$ means that 0 is its own additive inverse. We already saw in Example 6.5 that a ring can consist of the additive identity only.

Example 6.13

Let $M_2(\mathbb{Z})$ be the set of 2×2 matrices with integer entries. We equip this set with the usual addition and multiplication of matrices. (In case you have not seen these operations before, or don't remember them well, we have relegated a discussion of them to Exercises 6.6 and 6.7.) Note that when we add or multiply two matrices with integer entries, we obtain another one, and so this set is closed under the operations. We claim that $M_2(\mathbb{Z})$ is a ring:

Rules 1 and 2 follow easily because they hold in $\mathbb{Z}$. The zero of $M_2(\mathbb{Z})$ is $\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$, which can easily be verified. The additive inverse of $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is $\begin{pmatrix} -a & -b \\ -c & -d \end{pmatrix}$. That multiplication is associative and that the distributive laws hold are left as Exercise 6.7.

But note that $M_2(\mathbb{Z})$ is not commutative. For example,

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 4 & 3 \end{pmatrix} \neq$$

$$\begin{pmatrix} 3 & 4 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}.$$

The previous example shows the relevance of the ring concept to students of linear algebra. Our final example of the chapter connects algebra to the study of calculus and analysis:

Example 6.14

Let $C[0,1]$ be the set of real-valued functions defined on the closed unit interval $[0,1] = \{x \in \mathbb{R} : 0 \leq x \leq 1\}$, which are continuous. Define the sum and product of two functions point-wise: $(f+g)(x) = f(x) + g(x)$ and $(fg)(x) = f(x)g(x)$. You can use theorems from calculus to show that this set is a commutative ring. (See Exercise 6.9.)

6.3 Arithmetic in a Ring

We can now begin to talk about the arithmetic in an arbitrary ring. The following theorem shows some of the simple arithmetic operations we're used to doing in $\mathbb{Z}$, which we can now perform in an arbitrary ring:

Theorem 6.1 *Suppose R is a ring, and $a, b, c \in R$.*

a. (Additive Cancellation) If $a + b = a + c$, then $b = c$.

b. (Solution of equations) The equation $a + x = b$ always has a unique solution in R.

c. (Uniqueness of additive inverse) Every element of R has exactly one additive inverse.

d. (Uniqueness of additive identity) There is only one element of R which satisfies the equations $z + a = a$, for all a; namely, the element 0.

Proof: We will proceed very carefully from the rules defining a ring, for the first of these proofs, and then argue less formally. The reader is invited to fill in the careful details.

$\triangleright$ **Quick Exercise.** Fill in the details for the proofs below of parts b, c, and d of the theorem. $\triangleleft$

(a): Suppose that R is a ring, $a, b, c \in R$, and $a+b = a+c$. By Rule 4, we know there exists an element $x \in R$ for which $a + x = 0$. By Rule 1 we know that $0 = a + x = x + a$. We can now add x to both sides of our given equation, to obtain $x + (a+b) = x + (a+c)$. By two applications of Rule 2, we then have that $(x+a)+b = (x+a)+c$. But then we have $0+b = 0+c$, and so by Rules 1 and 3, we conclude that $b = c$, as required.

(b): We will now proceed less formally than in the proof for (a). Note that $x = d + b$ will do the job, where d is the additive inverse of a. Suppose that e is some other solution to the equation. Then $a + e = a + (d + b)$, and so, by the Additive Cancellation property, $e = d + b$. Thus, our solution is unique.

(c): Suppose a has two additive inverses; say, x and y. Then $a+x = 0 = a+y$, and so, by the Additive Cancellation property, $x = y$.

(d): This is left as Exercise 6.2. □

Note well that although additive cancellation holds in any ring (by part a of Theorem 6.1), multiplicative cancellation can certainly fail. For instance, we have already seen that in $\mathbb{Z}_6$, $3 \cdot 2 = 3 \cdot 4$ $(= 0)$, but $2 \neq 4$. Because the axioms in the definition of a ring require less of multiplication, we should expect fewer nice properties for multiplication than for addition.

A very careful reader may still be concerned about the rigor of the proof we offered above for part (a). How can we justify adding x to both sides of the given equation? First of all, note that elements $x + (a + b)$ and $x + (a + c)$ exist in the ring, because the ring is closed under addition. The fact that they are equal follows, not from one of our Rules for rings, but rather from a property of equality, known as the *Substitution Rule for Equality*. This says that if $a = b$ (that is, if a and b are identical elements of the ring), then we can safely replace any appearances of a in an expression by b and get a new expression equal to the old. This is actually a rule of logic, rather than of ring theory, and in this book we have made no attempt to carefully axiomatize the rules of logic, since this would take us too far afield from algebra. In proofs you write you should feel free to use the ordinary properties of equality with which you have been long familiar, including the *Substitution Rule*. In particular, note that $a = a$, for all a (this is called *Reflexivity*). If $a = b$ then $b = a$ (this is called *Symmetry*). And if $a = b$ and $b = c$, then $a = c$ (this is called *Transitivity*).

6.4 Notational Conventions

Because additive inverses are unique, we can thus denote *the* additive inverse of element a by the unambiguous notation $-a$.

You should exercise some care in using this notation, however. We are *not* interpreting $-a$ as meaning the product of -1 and a. In an arbitrary ring we have no guarantee that there exist such elements as -1 or 1 (see Example 6.8).

We can now make sense of *subtraction* in an arbitrary ring, simply by interpreting $a - b$ as $a + (-b)$. In Exercise 6.5, you will show that subtraction in an arbitrary ring obeys rules like those found in $\mathbb{Z}$.

Here are some further handy notational conventions: For $n \in \mathbb{N}$ and ring element a, $na = a + a + \ldots + a$ and $a^n = a \cdot a \cdot \ldots \cdot a$, where a appears n times on the right-hand side of the equations. Thus, $3a$ is shorthand for $a + a + a$, and a^4 is shorthand for $a \cdot a \cdot a \cdot a$.

Care must be exercised in the use of these conventions. For example, in $2\mathbb{Z}$, we can write

$$2 + 2 + 2 = 3(2),$$

and interpret this as a calculation inside $2\mathbb{Z}$, even though there is no element 3 belonging to the ring.

▷ **Quick Exercise.** In the ring $M_2(\mathbb{Z})$, what are the elements

$$3\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}^3 ? \lhd$$

6.5 The Set of Integers is a Ring

In Chapters 1 and 2 we relied on your previous experience with the integers when dealing with arithmetic and admitted that this was a flaw if we were attempting to make a careful axiomatic development of the properties of the integers. We now have the language necessary to rectify this logical flaw by stating carefully what we've been assuming all along.

Axiom of Arithmetic *The integers $\mathbb{Z}$ under ordinary addition and multiplication is a ring.*

This axiom (together with the Well-ordering Principle) is all we need to prove what we have about the integers. In the future we will be showing that many other rings have 'integer-like' properties.

Historical Remarks

In this chapter we defined the abstract concept of ring, as a way of generalizing the arithmetic properties possessed by our particular examples $\mathbb{Z}$ and $\mathbb{Q}[x]$. Historically, this definition was a long time in

coming, but, in broad strokes, we are being accurate to the historical development in basing our definition on $\mathbb{Z}$ and $\mathbb{Q}[x]$. This is true because in the 19th century, the formal definition of ring (and its ensuing popularity in mathematical circles) grew out of two subjects, related respectively to $\mathbb{Z}$ and $\mathbb{Q}[x]$. The first subject is number theory. Such mathematicians as the Germans Ernst Kummer and Richard Dedekind discovered that number-theoretic questions about $\mathbb{Z}$ are related to such rings as $\mathbb{Z}[i]$ (described in Exercise 6.12). We will follow this topic further in ensuing chapters. The second subject is the geometry related to polynomial equations (especially in more than one variable). $\mathbb{Q}[x]$ is a starting point for this subject (called algebraic geometry), which we won't pursue in this book. The crucial historical figure here is another German, named David Hilbert. Hilbert was a great believer in the power of axiomization and abstraction, and his success profoundly influenced the course of 20th-century mathematics. If you are having some difficulty understanding the utility of such an abstract concept as ring, you should take solace in the fact that many of Hilbert's late 19th-century colleagues were also reluctant to follow him down the road of abstraction! But this road has many wonders, just around the corner.

Chapter Summary

In this chapter we defined what we mean by a *ring*: a set equipped with two operations called addition and multiplication, which satisfy certain natural axioms. We examined numerous examples of rings (including $\mathbb{Z}$ and $\mathbb{Q}[x]$) and began the study of the arithmetic of an arbitrary ring.

Warm-up Exercises

a. Explain why our definition of a binary operation guarantees that the set is closed under the operation.

b. Are the following binary operations?

(a) $a * b = 1$, on the set $\mathbb{Z}$.

(b) $a * b = a/b$, on the set $\mathbb{Q}$.

(c) $a * b = a + bi$, on the set $\mathbb{R}$.

c. Give examples of binary operations satisfying the following:

(a) A non-commutative binary operation.
(b) A non-associative binary operation.

d. Give examples of rings satisfying the following:

(a) A ring with finitely many elements.
(b) A non-commutative ring.

e. Are the following rings?

(a) $3\mathbb{Z}$, the set of all integers divisible by 3, together with ordinary addition and multiplication.
(b) The set of all irreducible integers, together with ordinary addition and multiplication.
(c) $\mathbb{R}$, with the operations of addition and division.
(d) The set $\mathbb{R}^*$ of non-zero real numbers, with the operations of multiplication, and the operation $a \circ b = 1$. *Note*: We are trying to use ordinary multiplication as the 'addition' in this set!
(e) The set of polynomials in $\mathbb{Q}[x]$, where the constant term is an integer, with the usual addition and multiplication of polynomials.
(f) The set of all matrices in $M_2(\mathbb{Z})$, whose lower left-hand entry is zero, with the usual matrix addition and multiplication.

f. Compute $4a$ and a^4 for the following elements a of the following rings:

(a) $1/2 \in \mathbb{Q}$.
(b) $2 \in \mathbb{Z}_8$.
(c) $2 \in \mathbb{Z}_{16}$.
(d) $2 \in \mathbb{Z}_3$.
(e) $1 + 3x^2 \in \mathbb{Q}[x]$.
(f) $\begin{pmatrix} 1 & 2 \\ -1 & 3 \end{pmatrix} \in M_2(\mathbb{Z})$.

Exercises

1. Show that in a ring, $0a = a0 = 0$.

2. Prove part d of Theorem 6.1: Show that in a ring the additive identity is unique, by supposing that both 0 and $0'$ satisfy Rule 3 and proving that $0 = 0'$.

3. Show that in a ring, $(-a)b = a(-b) = -(ab)$.

4. Show that in a ring, $(-a)(-b) = ab$.

5. Prove the following facts about subtraction in a ring R, where $a, b, c \in R$:

   (a) $a - a = 0$.
   (b) $a(b - c) = ab - ac$.
   (c) $(b - c)a = ba - ca$.

6. Given two matrices in $M_2(\mathbb{Z})$,
$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix},$$
define their **matrix sum** to be the matrix
$$A + B = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{pmatrix}.$$
Verify that this is a binary operation, which is associative, commutative, has an additive identity, and has additive inverses.

7. Given two matrices in $M_2(\mathbb{Z})$,
$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix},$$
define their **matrix product** to be the matrix
$$AB = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix}.$$
Verify that this is a binary operation, which is associative and distributes over matrix addition, but is not commutative.

8. We generalize Exercises 6 and 7: Let R be any commutative ring (other than the zero ring). Define $M_2(R)$ as the set of 2×2 matrices with entries from R. Show that $M_2(R)$ is a ring which is not commutative. (Note that for the most part the proofs in Exercises 6 and 7 lift over without change.)

9. Check that Example 6.14 is indeed a ring; that is, let $C[0,1]$ be the set of functions defined from the closed unit interval $[0,1]$ to the real numbers that are continuous. Define the sum and product of two functions point-wise: $(f+g)(x) = f(x)+g(x)$ and $(fg)(x) = f(x)g(x)$. Show that $C[0,1]$ is a commutative ring. (You may use theorems from calculus.)

10. Let $\mathcal{D}$ be the set of functions defined from the real numbers to the real numbers that are differentiable. Define addition and multiplication of functions point-wise, as in the previous exercise. Show that $\mathcal{D}$ is a commutative ring. (You may use theorems from calculus.)

11. Let $\mathbb{C}$ be the complex numbers. That is,

$$\mathbb{C} = \{a+bi : a, b \in \mathbb{R}\},$$

where i is the square root of -1 (that is, $i \cdot i = -1$). Here,

$$(a+bi)+(c+di) = (a+c)+(b+d)i$$

and

$$(a+bi)(c+di) = (ac-bd)+(ad+bc)i.$$

Show that $\mathbb{C}$ is a commutative ring.

12. Let

$$\mathbb{Z}[i] = \{a+bi \in \mathbb{C} : \ a, b \in \mathbb{Z}\}.$$

Show that $\mathbb{Z}[i]$ is a commutative ring (see Exercise 11). This is called the ring of **Gaussian integers**.

13. Verify that Example 6.9 is a ring. That is, let $\mathbb{Z} \times \mathbb{Z}$ be the set of ordered pairs with integer entries. That is,

$$\mathbb{Z} \times \mathbb{Z} = \{(a,b) : a, b \in \mathbb{Z}\}.$$

Define addition and multiplication coordinate-wise; that is,

$$\begin{aligned}(n,m)+(r,s) &= (n+r, m+s)\\ (n,m)\cdot(r,s) &= (nr, ms).\end{aligned}$$

Show that $\mathbb{Z} \times \mathbb{Z}$ is a commutative ring.

14. We generalize Exercise 13: Let $\mathbb{Z}^n$ be the set of ordered n-tuples with integer entries. Define addition and multiplication on $\mathbb{Z}^n$ coordinate-wise and show $\mathbb{Z}^n$ is a commutative ring. Similarly, define R^n for any ring R. Show R^n is a ring and is commutative if R is.

15. Verify that Example 6.10 is a ring. Namely, let R and S be arbitrary rings. Define addition and multiplication appropriately to make $R \times S$ a ring, where $R \times S$ is the set of ordered pairs with first entry from R and second entry from S. Now generalize this to the set $R_1 \times R_2 \times \cdots \times R_n$ of n-tuples with entries from the rings R_i. This new ring is called the **direct product** of the rings R_i.

16. Find an example in $M_2(\mathbb{Z})$ to show that $(a+b)^2$ is *not* necessarily equal to $a^2+2ab+b^2$. (Recall that $2ab = ab+ab$.) What is the correct expansion of $(a+b)^2$ for an arbitrary ring? What can you say if the ring is commutative?

17. (This exercise extends the discussion of Exercise 16.) Let R be a commutative ring and $a, b \in R$. Then prove the *binomial theorem* for R, by induction on n: Namely, show that

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

18. Suppose that $a \cdot a = a$ for every element a in a ring R. (Elements a in a ring where $a^2 = a$ are called **idempotent**.)

    (a) Show that $a = -a$.

    (b) Now show that R is commutative.

19. Let $S = \{(x_1, x_2, x_3, \ldots) : x_i \in \mathbb{R}\}$, the real-valued sequences. Define addition and multiplication on S coordinate-wise (see Exercises 13 and 14). Show that S is a commutative ring.

20. Let X be some arbitrary set, and $P(X)$ the set of all subsets of X. In Example 1.1 we proved that if X has n elements, then $P(X)$ has 2^n elements; we are here allowing the possibility that X (and hence $P(X)$) has *infinitely* many elements. Define operations on $P(X)$ as follows, where $a, b \in P(X)$:

$$a + b = (a \cup b) \backslash (a \cap b) \quad \text{and} \quad ab = a \cap b.$$

(Addition here is often called the **symmetric difference** of the two sets a, b.) Prove that $P(X)$ is a commutative ring. ($P(X)$ is called the **power set** for the set X.)

21. Prove that multiplication is associative in $\mathbb{Q}[x]$, as claimed in Example 6.6. Suppose that

$$f = a_0 + a_1 x + \cdots + a_n x^n,$$

$$g = b_0 + b_1 x + \cdots + b_n x^m,$$

and

$$h = c_0 + c_1 x + \cdots + c_n x^r.$$

Compute the x^k coefficient in $f(gh)$ and $(fg)h$. In each case these will be double summations. Argue that these double summations actually include the same terms and are therefore the same.

22. Prove that multiplication is distributive in $\mathbb{Q}[x]$, as claimed in Example 6.6.

23. Let R be any commutative ring. Let $R[x]$ be the collection of polynomials with coefficients from R. Show that $R[x]$ is a ring.

Chapter 7

Subrings and Unity

Consider Example 6.7: We discovered that it was relatively easy to show that $\mathbb{Z}[x]$ is a ring, because it is a subset of $\mathbb{Q}[x]$, which we had already shown is a ring. Because the operations in $\mathbb{Z}[x]$ are the same as in $\mathbb{Q}[x]$, we didn't have to check again the associative laws, the distributive laws, or that addition was commutative. Because the addition and multiplication of $\mathbb{Q}[x]$ have these properties, the addition and multiplication of $\mathbb{Z}[x]$ inherit them automatically. What did need to be checked was that addition and multiplication were closed in $\mathbb{Z}[x]$, that the additive identity of $\mathbb{Q}[x]$ was also in $\mathbb{Z}[x]$, and that the additive inverses of elements of $\mathbb{Z}[x]$ were also in $\mathbb{Z}[x]$. Similarly, in Example 6.8 you showed that $2\mathbb{Z}$ is a ring, taking advantage of the fact that $2\mathbb{Z} \subseteq \mathbb{Z}$.

7.1 Subrings

We generalize this situation: A subset S of a ring R is said to be a **subring** of R if S is itself a ring under the operations induced from R.

Example 7.1

$\mathbb{Z}[x]$ is a subring of $\mathbb{Q}[x]$.

Example 7.2

$2\mathbb{Z}$ is a subring of $\mathbb{Z}$.

Example 7.3

$\mathbb{Z}$ is a subring of $\mathbb{Q}$, which is in turn a subring of $\mathbb{R}$.

Example 7.4

The Gaussian integers $\mathbb{Z}[i]$ are a subring of $\mathbb{C}$. (See Exercises 6.11 and 6.12.)

Example 7.5

Let R be any ring. Then $\{0\}$ and R are always subrings of R. We call $\{0\}$ the **trivial** subring, and R the **improper** subring. All subrings other than R we call **proper** subrings.

In the general case, what exactly do we need to check to see that a subset of a ring is a subring? We need not go through all the work we did in showing that $\mathbb{Z}[x]$ is a subring of $\mathbb{Q}[x]$, as listed above. The following theorem provides a simpler answer. We will often use this theorem to check whether something is a ring (by considering it as a subset of a larger well-known ring):

Theorem 7.1 The Subring Theorem *A non-empty subset of a ring is a subring under the same operations if and only if it is closed under multiplication and subtraction.*

Proof: It is obvious that a subring is closed under multiplication and subtraction. For the converse, suppose that R is a ring and S a non-empty subset, which is closed under multiplication and subtraction. We wish to show that S is a ring. Now, because S is non-empty, we can then choose an element of it, which we call s. First note that because S is closed under subtraction, then $s - s = 0 \in S$. That is, the additive identity belongs to S (Rule 3). Next, suppose that $a \in S$. Because S is closed under subtraction, $-a = 0 - a \in S$. This means that S is closed under taking additive inverses (Rule 4). Now suppose that $a, b \in S$. Then we've just seen that $-b \in S$. But then

$$a + b = a - (-b) \in S,$$

and so S is closed under addition as well.

To show that S is a ring, it remains to show that addition is commutative, that addition and multiplication are associative, and that multiplication distributes over addition. But all these properties hold in R, and so are automatically inherited for S. □

The most important use of this theorem is to check that some set is in fact a ring, by viewing it as a subring of some larger previously known ring. We will see this principle illustrated repeatedly in the examples below.

Because commutativity is automatically inherited by a subset, it follows that a subring of a commutative ring is also commutative. Note that a subring of a non-commutative ring may be commutative, because the zero ring is a commutative subring of *any* ring. Example 7.9 is a more interesting example of this.

Example 7.6

Let $m\mathbb{Z} = \{mn : n \in \mathbb{Z}\}$, where m is an integer greater than 1. That is, $m\mathbb{Z}$ is the set of integer multiples of m. We have already seen this example in case $m = 2$ as Example 7.2. We claim that $m\mathbb{Z}$ is a subring of $\mathbb{Z}$. For if $ma, mb \in m\mathbb{Z}$, then

$$ma - mb = m(a - b) \in m\mathbb{Z}$$

and so $m\mathbb{Z}$ is closed under subtraction. Similarly,

$$(ma)(mb) = m(mab) \in m\mathbb{Z},$$

and so $m\mathbb{Z}$ is closed under multiplication.

Example 7.7

$6\mathbb{Z}$ is a subring of both $3\mathbb{Z}$ and $2\mathbb{Z}$.

▷ **Quick Exercise.** Check that $6\mathbb{Z}$ is a subring of $3\mathbb{Z}$ and $2\mathbb{Z}$, modelling your proof on that of Example 7.6. ◁

We generalize this example in Exercise 7.4.

Example 7.8

$\{0, 2, 4\}$ is a subring of $\mathbb{Z}_6$. It is easy to check directly that this set is closed under subtraction and multiplication. We extend this example in Exercise 7.c.

Example 7.9

Let $D_2(\mathbb{Z})$ be the set of all 2-by-2 matrices with entries from $\mathbb{Z}$ and with all entries off the main diagonal being zero. We call these the **diagonal matrices**. We claim that this is a subring of $M_2(\mathbb{Z})$.

▷ **Quick Exercise.** Check that the set of diagonal matrices is closed under subtraction and multiplication. ◁

Note that $D_2(\mathbb{Z})$ is in fact a commutative ring:

$$\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}\begin{pmatrix} c & 0 \\ 0 & d \end{pmatrix} = \begin{pmatrix} ac & 0 \\ 0 & bd \end{pmatrix} = \begin{pmatrix} c & 0 \\ 0 & d \end{pmatrix}\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}.$$

Thus, $D_2(\mathbb{Z})$ is a commutative subring (bigger than the zero ring) of a non-commutative ring.

Example 7.10

The direct product $\mathbb{Z} \times \mathbb{Q}$ is a subring of the ring $\mathbb{R} \times \mathbb{R}$. (See Example 6.10.)

Here are a few examples of subsets of rings which are *not* subrings. Note how many different ways a subset can fail to be a subring.

Example 7.11

Consider the set $\mathbb{Q}^+$ of all strictly positive elements from $\mathbb{Q}$. This is clearly not a subring, because the additive identity 0 does not belong to $\mathbb{Q}^+$.

Example 7.12

Consider the set $\mathbb{Q}^+ \cup \{0\} \subseteq \mathbb{Q}$. This set does include 0 and is in fact closed under multiplication. But it is *not* closed under subtraction and so is not a subring.

Example 7.13

Consider the set

$$\pi\mathbb{Z} = \{\pi n : n \in \mathbb{Z}\},$$

that is, the set of all integer multiples of the real number π. This is a subset of the ring $\mathbb{R}$ of all real numbers. It is closed under subtraction.

▷ **Quick Exercise.** Check that $\pi\mathbb{Z}$ is closed under subtraction. ◁

However, it is not closed under multiplication, for if it were, then π^2 would be an integer multiple of π. But then π itself would be an integer, which is false.

Example 7.14

Consider $\mathbb{Z}_4 = \{0, 1, 2, 3\}$. Is it a subring of $\mathbb{Z}$? Clearly not, because although our notation makes it *look* as if $\mathbb{Z}_4$ is a subset of $\mathbb{Z}$, this is not actually the case. Recall from Section 3.2, that when working with the rings $\mathbb{Z}_n$ we often drop the square brackets around the integers, if it is clear from context what we mean. In $\mathbb{Z}_4$, 3 means a residue class, and so is an infinite set of integers. This infinite set of integers is certainly not the same as the individual integer 3.

7.2 The Multiplicative Identity

We close this chapter by mentioning another topic, which further illustrates the different roles that addition and multiplication play in a ring. It is an important part of the definition of a ring that it have an additive identity or 0. We make no such assumption about a multiplicative identity; many rings do possess a multiplicative identity, however. We call an element u of a ring R a **unity** or **multiplicative identity** if $ua = au = a$ for all elements $a \in R$.

Example 7.15

Obviously, the integer 1 is the unity of $\mathbb{Z}$. Similarly, 1 is the unity for the rings $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$. In $\mathbb{Q}[x]$, the constant polynomial 1 is the unity. In the ring $M_2(\mathbb{Z})$, the unity is $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. In Chapter 3, we were careful to point out that the residue class [1] plays the role of multiplicative identity in $\mathbb{Z}_m$. On the other hand, the ring $2\mathbb{Z}$ has no unity, because in the integers $2a = 2$ holds exactly when $a = 1$, an element which $2\mathbb{Z}$ lacks. More generally, $m\mathbb{Z}$ lacks unity for all $m > 1$.

Note that for the particular examples we've discussed, we have spoken as if the unity element of a ring (if it exists) is unique. This is in fact the case and can be proved by a proof very similar to the one we used to show that the zero of a ring is unique. We leave this proof as Exercise 7.17. Because unity is unique, we can thus denote it unambiguously by 1.

There is a surprising difficulty with this, however. Consider the zero ring $\{0\}$. In this ring 0 is both the additive identity *and* the multiplicative identity, and so in this case $0 = 1$. In order to avoid this trivial case, we will henceforth reserve the term 'unity' for rings other than the zero ring. In Exercise 7.16, you will prove that in a ring with more than one element, the additive identity and multiplicative identity cannot be the same.

Chapter Summary

In this chapter we defined the notion of *subring* and proved that the subrings of a ring are exactly those non-empty subsets closed under subtraction and multiplication. We provided many examples of subrings.

We also defined the notion of *unity* (or multiplicative identity) and observed that many rings have unity, and many don't.

Warm-up Exercises

a. Give examples of the following:

(a) A non-empty subset of a ring, closed under subtraction, but not multiplication.

(b) A non-empty subset of a ring, closed under multiplication, but not subtraction.

b. Are the following subsets subrings?

(a) $\mathbb{Z} \subseteq \mathbb{Q}$.

(b) $\mathbb{Q}^* \subseteq \mathbb{Q}$; recall that $\mathbb{Q}^*$ is the set of non-zero rational numbers.

(c) $\mathbb{Q}^+ \subseteq \mathbb{Q}$; recall that $\mathbb{Q}^+$ is the set of strictly positive rational numbers.

(d) The set of irrational numbers, a subset of $\mathbb{R}$.

(e) $\{0, 1, 2, 3\} \subseteq \mathbb{Z}_8$.

(f) The linear polynomials, a subset of $\mathbb{Q}[x]$.

c. Find all the subrings of these rings: $\mathbb{Z}_5$, $\mathbb{Z}_6$, $\mathbb{Z}_7$, $\mathbb{Z}_{12}$.

d. Give examples of the following (or explain why they don't exist):

(a) A commutative subring of a non-commutative ring.

(b) A non-commutative subring of a commutative ring.

(c) A subring without unity, of a ring with unity. (See Exercise 22 for the converse possibility.)

(d) A ring (with more than one element) whose only subrings are itself, and the zero subring. *Hint:* Look at an earlier Warm-up Exercise.

e. What is the unity of the power set ring $P(X)$ considered in Exercise 6.20?

f. What is the unity of the ring $\mathbb{Z} \times \mathbb{Z}$? (See Example 6.9.) What about of $R \times S$, where R and S are rings with unity? (See Example 6.10.)

Exercises

1. Let $\mathbb{Z}[\sqrt{2}] = \{a + b\sqrt{2} : a, b \in \mathbb{Z}\}$. Show that $\mathbb{Z}[\sqrt{2}]$ is a commutative ring by showing it is a subring of $\mathbb{R}$.

2. We generalize Exercise 1: Let $\mathbb{Z}[\sqrt{n}] = \{a + b\sqrt{n} : a, b \in \mathbb{Z}\}$, where n is some fixed integer (positive or negative). Show $\mathbb{Z}[\sqrt{n}]$ is a commutative ring by showing it is a subring of $\mathbb{C}$.

3. Let
$$\alpha = \sqrt[3]{5}$$
and
$$\mathbb{Z}[\alpha] = \{a + b\alpha + c\alpha^2 : a, b, c \in \mathbb{Z}\} \subseteq \mathbb{R}.$$
Prove that $\mathbb{Z}[\alpha]$ is a subring of $\mathbb{R}$.

4. Show that $m\mathbb{Z}$ is a subring of $n\mathbb{Z}$ if and only if n divides m. (See Example 7.7.)

5. (a) Show that
$$4\mathbb{Z} \cap 6\mathbb{Z} = 12\mathbb{Z}.$$
(b) Let m and n be two positive integers. Show that
$$m\mathbb{Z} \cap n\mathbb{Z} = l\mathbb{Z},$$
where l is the least common multiple of m and n. (See Exercise 2.11.)

6. Let S be the set of all polynomials in $\mathbb{Q}[x]$ which have 0 as constant term (that is, polynomials of the form $a_1x+a_2x^2+\cdots a_nx^n$). Show that S is a subring of $\mathbb{Q}[x]$.

7. Let f be some polynomial with rational coefficients, with $\deg(f) > 0$, and let S be the set of all polynomials g in $\mathbb{Q}[x]$ for which f divides g. Show that S is a subring of $\mathbb{Q}[x]$. How is this exercise related to the previous exercise?

8. (a) Show that the set
$$\{(a, a) : a \in \mathbb{Z}\}$$
is a subring of $\mathbb{Z} \times \mathbb{Z}$.

(b) Now consider the set
$$\{(a, -a) : a \in \mathbb{Z}\}.$$
Show that this set is closed under subtraction, but not closed under multiplication, and so is *not* a subring of $\mathbb{Z} \times \mathbb{Z}$.

9. Show that the intersection of any two subrings of a ring is a subring.

10. Show by example that the union of any two subrings of a ring need *not* be a subring. *Hint:* You can certainly find such an example by working in $\mathbb{Z}$.

11. Let
$$\mathbb{Z}_{\langle 2\rangle} = \left\{q \in \mathbb{Q} : q = \frac{a}{b}, \quad a, b \in \mathbb{Z},\ b \text{ is odd}\right\}.$$
(a) Why is $\mathbb{Z}_{\langle 2\rangle}$ a proper subset of $\mathbb{Q}$?
(b) Show that $\mathbb{Z}_{\langle 2\rangle}$ is a subring of $\mathbb{Q}$.
(c) Show that
$$\left\{q \in \mathbb{Q} : q = \frac{a}{b}, \quad a, b \in \mathbb{Z},\ a \text{ is odd},\ b \neq 0\right\}$$
is *not* a subring of $\mathbb{Q}$.
(d) Show that
$$\left\{q \in \mathbb{Q} : q = \frac{a}{2^n}, \quad a \in \mathbb{Z},\ n = 0, 1, 2, \cdots\right\}$$
is a subring of $\mathbb{Q}$.

12. Let R be an arbitrary ring, and define
$$Z(R) = \{r \in R : rx = xr, \text{for all } x \in R\};$$
this subset is called the **center** of the ring R. Show that $Z(R)$ is a subring. What is $Z(R)$ if R is a commutative ring?

13. Find the center of $M_2(\mathbb{Z})$ (see the previous exercise).

14. Let R be a ring, and s a particular fixed element of R. Let
$$Z_s(R) = \{r \in R : rs = sr\}.$$

(a) Prove that $Z_s(R)$ is a subring of R.

(b) Recall the definition of $Z(R)$ from Exercise 12. Prove that

$$Z(R) = \cap \{Z_s(R) : s \in R\}.$$

15. (a) An element a of a ring is **nilpotent** if $a^n = 0$ for some positive integer n. Given a ring R, denote by $N(R)$ the set of all nilpotent elements of R. (This subring is called the **nilradical** of the ring.) If R is any commutative ring, show that $N(R)$ is a subring.

(b) Determine $N(\mathbb{Z}_{10})$, the nilradical of $\mathbb{Z}_{10}$.

(c) Determine $N(\mathbb{Z}_8)$, the nilradical of $\mathbb{Z}_8$.

16. Suppose that R is a ring with unity, and R has at least two elements. Prove that the additive identity of R is not equal to the multiplicative identity.

17. Show that if a ring has unity, it is unique.

18. (a) Let R be a ring, and consider the set $R \times \mathbb{Z}$ of all ordered pairs with entries from R and $\mathbb{Z}$. Equip this set with operations

$$(r, n) + (s, m) = (r + s, n + m)$$

and

$$(r, n)(s, m) = (rs + mr + ns, nm).$$

Prove that these operations make $R \times \mathbb{Z}$ a ring. (Note that this is *not* the same ring discussed in Example 6.10.)

(b) Show that $R \times \mathbb{Z}$ under these operations has unity, even if R does not.

(c) Show that $R \times \{0\}$ is a subring of the ring $R \times \mathbb{Z}$. Argue that this ring is 'essentially the same' as R. (*Note*: Later in the book we will make precise the notion of two rings which are 'essentially the same', by defining *ring isomorphism*.) This means that any ring without unity can essentially be found as a subring of a ring which has unity.

19. Consider the set

$$\left\{ \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} : a, b, c \in \mathbb{Z} \text{ and } a \text{ is even} \right\}.$$

Prove that this set is a subring of $M_2(\mathbb{Z})$. Does it have unity?

20. Some students wonder why we require that the addition in a ring be commutative; this exercise shows why. Suppose that R is a set with two operations $+$ and $\circ$, which satisfy the rules defining a ring, except for Rule 1; that is, we do not assume that the addition is commutative. Suppose that R also has a multiplicative identity 1. Then prove that the addition in R must in fact be commutative, and so R under the given operations is a ring.

21. Consider the ring S of all real-valued sequences, as discussed in Exercise 6.19. Let F be the set of all sequences $(x_1, x_2, x_3, \cdots)$ where at most finitely many of the entries x_i are non-zero. Prove that F is a subring of S. Does F have unity?

22. Let F be the ring of finitely non-zero real-valued sequences, considered in the previous exercise. Now let W consist of all sequences $(x_1, x_2, x_3 \cdots)$ where $0 = x_2 = x_3 = \cdots$. Show that W is a subring that has unity, even though the larger ring F does not.

23. Let R be a ring, and let

$$S = \{r \in R : r + r = 0\}.$$

Prove that S is a subring of R.

24. Generalize Exercise 23. That is, let R be a ring, and let n be a fixed positive integer. Let

$$S = \{r \in R : nr = 0\}.$$

Prove that S is a subring of R. (Recall that nr means to add r to itself n times; the integer n need not belong to R.)

25. Let R be a commutative ring with unity. An element $e \in R$ is **idempotent** if $e^2 = e$. Note that the elements 0 and 1 are idempotent. Throughout this problem assume that e is idempotent in R.

(a) Find a commutative ring with unity with at least one idempotent element; other than 0 and 1.

(b) Let $f = 1 - e$. Prove that f is idempotent, too.

(c) Let $Re = \{re : r \in R\}$. Prove that Re is a subring of R, and that e is unity for this subring.

(d) Prove that $Re \cap Rf = \{0\}$ (where f is the idempotent from part b).

(e) Prove that for all $r \in R$, $r = a+b$, where $a \in Re$ and $b \in Rf$.

Chapter 8

Integral Domains and Fields

Something surprising happens in the arithmetic in $\mathbb{Z}_6$: two non-zero elements 2 and 3 give 0 when multiplied together. This is something that never happens in $\mathbb{Z}$ (or $\mathbb{Q}$ or $\mathbb{R}$). In fact, this never happens in $\mathbb{Z}_5$.

$\triangleright$ **Quick Exercise.** Use the multiplication table for $\mathbb{Z}_5$ in Chapter 3 to check that the product of two non-zero elements in $\mathbb{Z}_5$ is always non-zero. $\triangleleft$

The fact that $2 \cdot 3 = 0$ in $\mathbb{Z}_6$ has undesirable consequences. For example, it means that $2 \cdot 3 = 2 \cdot 0$, and so we *cannot* cancel the 2 from each side of this equation.

8.1 Zero Divisors

We make a definition to explore this situation: Let R be a commutative ring. An element $a \neq 0$ is a **zero divisor** if there exists an element $b \in R$ such that $b \neq 0$ and $ab = 0$. Of course, then b is a zero divisor also.

Example 8.1

Thus, the elements 2 and 3 in $\mathbb{Z}_6$ are zero divisors because $2{\cdot}3 = 0$. For another example, consider the elements $(1,0)$ and $(0,1)$ in $\mathbb{Z} \times \mathbb{Z}$. They are zero divisors because $(1,0)(0,1) = (0,0)$. On the other hand, 2 is *not* a zero divisor in $\mathbb{Z}$ because the equation $2x = 0$ has only $x = 0$ as a solution.

$\triangleright$ **Quick Exercise.** Détermine the set of all zero divisors for the rings $\mathbb{Z}_6$, $\mathbb{Z}_5$, $\mathbb{Z} \times \mathbb{Z}$, and $\mathbb{Z}$. $\triangleleft$

You should have just concluded that the rings $\mathbb{Z}$ and $\mathbb{Z}_5$ have no zero divisors. This desirable property we highlight by a definition:

A commutative ring with unity that has no zero divisors is called an **integral domain**, or simply a **domain**.

Example 8.2

Thus, $\mathbb{Z}$ and $\mathbb{Z}_5$ are integral domains, as are such rings as $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$.

Example 8.3

What about the ring $\mathbb{Q}[x]$? We observed in Chapter 4 that the product of two non-zero polynomials is non-zero, because the degree of the product is the sum of the degrees (Theorem 4.1). Thus, $\mathbb{Q}[x]$ (and similarly $\mathbb{Z}[x]$) is a domain too.

Example 8.4

On the other hand, we have seen that the rings $\mathbb{Z} \times \mathbb{Z}$ and $\mathbb{Z}_6$ are *not* domains, simply because we have already exhibited the existence of zero divisors in them. In fact, if n is not prime, then $\mathbb{Z}_n$ is not a domain.

▷ **Quick Exercise.** Show that $\mathbb{Z}_n$ is not a domain when n is not prime by exhibiting zero divisors in each such ring. ◁

Notice that in the definition of domain we require both that it be commutative and have unity. These restrictions are standard in the subject, and we will consequently adhere to them. But note that this means that $2\mathbb{Z}$ is not a domain, even though it is commutative and has no zero divisors.

The arithmetic in integral domains is much simpler than in arbitrary commutative rings. An important example of this is the content of the following theorem: We can cancel multiplicatively in a domain. Of course, we already saw above that multiplicative cancellation fails in $\mathbb{Z}_6$.

Theorem 8.1 Multiplicative Cancellation *Suppose R is an integral domain and a, b, c are elements of R, with $a \neq 0$. If $ab = ac$, then $b = c$.*

Proof: Suppose that R is a domain, $a \neq 0$, and $ab = ac$. Then $ab - ac = 0$. But then $a(b - c) = 0$, and because R is a domain with $a \neq 0$, we must have $b - c = 0$, or $b = c$, as required. □

If a, b, c had been rational numbers in the previous proof, we would have been inclined to multiply both sides of the equation $ab = ac$ by $1/a$, the multiplicative inverse of a. However, the proof of the theorem holds even in the absence of multiplicative inverses (as is the case for most elements in $\mathbb{Z}$).

8.2 Units

Of course, when multiplicative inverses *do* exist, life is much simpler. Let us introduce some terminology to deal with this case. Suppose R is a ring with unity 1. Let a be any non-zero element of R. We say a is a **unit** if there is an element b of R such that $ab = ba = 1$. In this case, b is a **(multiplicative) inverse** of a. (Of course, b is also a unit with inverse a.)

First note that the unity 1 is always a unit, because $1 \cdot 1 = 1$. What other elements are units? In $\mathbb{Z}$, the units are just 1 and -1, because the only integer solutions of $ab = 1$ are ± 1. In $\mathbb{Q}$ and $\mathbb{R}$, *all* non-zero elements are units. In $\mathbb{Z}_6$ we have $5 \cdot 5 = 1$, and so 5 is a unit, as well as 1. Furthermore, there are no other elements a and b for which $ab = 1$ is true (see the multiplication table for $\mathbb{Z}_6$ in Chapter 3).

▷ **Quick Exercise.** Compute the units of these rings: $\mathbb{Z}_5$, $\mathbb{Z}_{12}$, $\mathbb{Z} \times \mathbb{Z}$, $\mathbb{R} \times \mathbb{R}$, $\mathbb{Q}[x]$. ◁

Note that the concept of multiplicative inverse makes perfectly good sense in non-commutative rings. For example, in the (non-commutative) ring of matrices $M_2(\mathbb{R})$, the elements

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{pmatrix}$$

are units, because their product (in either order) is the multiplicative identity. In Exercise 8.2 you will obtain all the units in this ring.

We now claim that multiplicative inverses, if they exist, are unique. To show this, suppose that a has multiplicative inverses b and c: then

$1 = ba = ca$. But now multiply through these equations on the right by b: We then have $b = bab = cab = c$ (where the last equation holds because $ab = 1$ also). But then $b = c$, as required. Consequently, we will denote the (unique) inverse of an element a (if it exists) by a^{-1}. This is of course consistent with the ordinary notation for multiplicative inverse which we use in $\mathbb{R}$.

We denote the set of units of a ring R by $U(R)$. Thus, $U(\mathbb{Z}) = \{1, -1\}$, $U(\mathbb{Q}) = \mathbb{Q}\backslash\{0\}$, and $U(\mathbb{Z}_6) = \{1, 5\}$.

The set $U(R)$ has some very nice properties, only some of which we can fully exploit right now. What we wish to observe immediately is that $U(R)$ is closed under multiplication.

▷ **Quick Exercise.** Check that this is true for $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{Z}_6$. ◁

To show that $U(R)$ is closed under multiplication, suppose that $a, b \in U(R)$. Then we claim ab is also a unit. But this is easy to see, because its inverse is just $b^{-1}a^{-1}$:

$$(ab)(b^{-1}a^{-1}) = a(bb^{-1})a^{-1} = aa^{-1} = 1,$$

and similarly for the product in the other order.

You might be surprised to see that taking the multiplicative inverse reverses the order of multiplication. But interpret a as putting on socks, and b as putting on shoes. To reverse the operation ab of putting on both socks and shoes, you must reverse the order: you take off shoes first, and so the inverse operation is $b^{-1}a^{-1}$. You will explore the importance of the order of the multiplication of the elements a^{-1} and b^{-1} in non-commutative rings in Exercise 8.3.

Students often confuse the unity of a ring with the concept of units, and you should take care to understand the definition. The unity of a ring is unique if it exists, and it is of course a unit (because it is its own multiplicative inverse). And a ring with units must necessarily have unity (because otherwise the concept of multiplicative inverse makes no sense). However, most rings have many units other than unity, as our examples above make clear.

8.3 Fields

Of course, rings like $\mathbb{Q}$ and $\mathbb{R}$ where *all* non-zero elements have multiplicative inverses seem particularly attractive, and we emphasize them by making a definition: A commutative ring with unity in which every non-zero element is a unit is called a **field**. Another way to look at the definition of a field is this: as a commutative ring in which one can always solve equations of the form $ax = b$ (when $a \neq 0$). The solution is, of course, $x = a^{-1}b$. In other words: In all rings we can add, subtract, and multiply, but in fields we can also *divide.*

Example 8.5

> Thus, $\mathbb{Q}$ and $\mathbb{R}$ are fields. By the Quick Exercise above, $\mathbb{Z}_5$ is also a field.

Every field is a domain. This follows immediately from the next theorem.

Theorem 8.2 *A field has no zero divisors.*

Proof: Suppose F is a field and $a \in F$, with $a \neq 0$. Now suppose that $ab = 0$. Then $0 = a^{-1} \cdot 0 = a^{-1}ab = b$. Thus, a is not a zero divisor. □

Note that the above proof actually shows that a unit of any ring cannot be a zero divisor. For example, in $\mathbb{Z}_6$, the units are 1 and 5, while the zero divisors are 2, 3, and 4.

8.4 The Field of Complex Numbers

Another important example of a field is the commutative ring $\mathbb{C}$ of complex numbers. (See Exercise 6.11.) To see that this ring is actually a field, we must compute the multiplicative inverse of the arbitrary non-zero complex number $\alpha = a + bi$, where a and b are real numbers (not both zero). We do this first by means of a bit of algebraic trickery.

The complex number $a - bi$ we call the **complex conjugate** of $a + bi$. We write it as $\bar{\alpha}$.

▷ **Quick Exercise.** Determine the complex conjugates of the following complex numbers:

$$1+i,\ 4-\frac{1}{2}i,\ 6i,\ 7 \ \triangleleft$$

It is easy to see that the product of a complex number with its conjugate is a real number:

$$\alpha\bar{\alpha} = (a+bi)(a-bi) = a^2-(bi)^2 = a^2+b^2.$$

But this means that

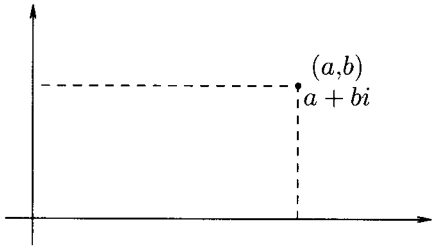
$$1 = (a+bi)(a-bi)\left(\frac{1}{a^2+b^2}\right) = (a+bi)\left(\frac{a}{a^2+b^2} - \frac{b}{a^2+b^2}i\right),$$

and so any non-zero complex number has a multiplicative inverse. Thus, $\mathbb{C}$ is a field. Note also that $\mathbb{R}$ is a subring of this field.

▷ **Quick Exercise.** Determine the multiplicative inverses of the following complex numbers:

$$1+i,\ 4-\frac{1}{2}i,\ 6i,\ 7 \ \triangleleft$$

There is a geometric approach to understanding these computations. Because a complex number $a+bi$ is determined by an ordered pair (a,b) of real numbers, it is only natural to associate with each complex number a point in the plane. The first coordinate gives us the **real part** of α, and the second coordinate gives us the **imaginary part** of α. We call the plane interpreted in this way the **complex plane**.



We can now talk about the **length** (or **modulus**) of a complex number. By the Pythagorean Theorem, this is evidently just

$$\sqrt{a^2+b^2} = \sqrt{\alpha\bar{\alpha}}.$$

We abbreviate this by $|\alpha|$. This generalizes the notion of *absolute value* for real numbers.

▷ **Quick Exercise.** Compute the moduli of the following complex numbers:

$$1+i,\ 4-\frac{1}{2}i,\ 6i,\ 7 \ \triangleleft$$

It is certainly true that the absolute value of a product of real numbers is the product of the absolute values. That is,

$$|ab| = |a||b|.$$

This generalizes to complex numbers:

Theorem 8.3 *For any complex numbers α and β,*

$$|\alpha\beta| = |\alpha||\beta|.$$

Proof: Suppose that $\alpha = a+bi$, and $\beta = c+di$ are complex numbers. Then

$$\begin{aligned}|\alpha\beta|^2 &= |(ac-bd)+(ad+bc)i|^2 = (ac-bd)^2+(ad+bc)^2\\ &= a^2c^2+b^2d^2+a^2d^2+b^2c^2 = (a^2+b^2)(c^2+d^2)\\ &= |\alpha|^2|\beta|^2.\end{aligned}$$

By taking the (positive) square root of both sides, we obtain what we wish. □

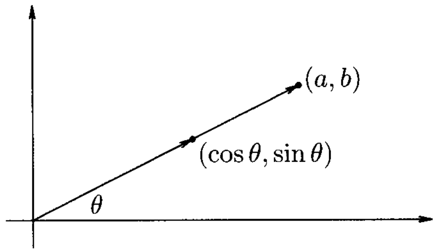
But a geometric understanding of complex multiplication extends even further than this. Note that the radial line from the origin to the point (a,b) makes an angle with the positive real axis. Let's call this angle θ, and always choose it in the interval $-\pi < \theta \leq \pi$. (We will leave this angle undefined for the complex number 0.) We call this angle the **argument** of α and write it as $\arg(\alpha)$.

▷ **Quick Exercise.** Compute the arguments of the following complex numbers (you may have to use a calculator to approximate the angle):

$$1+i,\ 4-\frac{1}{2}i,\ 6i,\ 7 \ \triangleleft$$

As the following diagram suggests, this means that we can now write any non-zero complex number as a product of its modulus (a positive real number), and a complex number of length 1, which can be written trigonometrically:

$$\begin{aligned}\alpha &= (\sqrt{a^2+b^2})\left(\frac{a}{\sqrt{a^2+b^2}}+\frac{b}{\sqrt{a^2+b^2}}i\right)\\ &= |\alpha|(\cos(\arg(\alpha))+i\sin(\arg(\alpha))).\end{aligned}$$



▷ **Quick Exercise.** Put together your results from the last two Quick Exercises to write $1+i$, $4-\frac{1}{2}i$, $6i$, 7 in this trigonometric form. ◁

We can now interpret multiplication in the field $\mathbb{C}$ as a geometric operation. Suppose that α and β are non-zero complex numbers, and we have factored them as

$$\alpha = |\alpha|(\cos\theta + i\sin\theta)$$

and

$$\beta = |\beta|(\cos\varphi + i\sin\varphi),$$

where θ and φ are, respectively, the arguments of α and β. To multiply these numbers, we first multiply the moduli to obtain the modulus of the product. Now let's deal with the trigonometric part. A surprising thing happens.

$$\begin{gathered}(\cos\theta + i\sin\theta)(\cos\varphi + i\sin\varphi) =\\ (\cos\theta\cos\varphi - \sin\varphi\sin\theta) + i(\cos\theta\sin\varphi + \cos\varphi\sin\theta) =\\ \cos(\theta+\varphi) + i\sin(\theta+\varphi).\end{gathered}$$

(We have used two familiar trigonometric identities.) This means that the argument of a product is the *sum* of the arguments (except that we may have to adjust the angle by 2π if $\theta+\varphi$ does not fall between $-\pi$ and π).

▷ **Quick Exercise.** Verify this statement about the argument of a product by computing the arguments of the products $(1+i)(6i)$ and $(-1+i)(1+\sqrt{3}i)$ in two ways. ◁

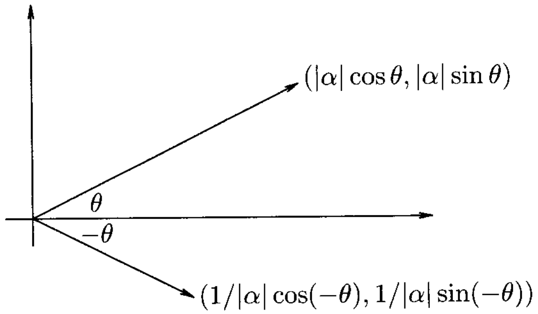
We record this statement precisely for future reference:

Theorem 8.4 DeMoivre's Theorem *Let θ and φ be two angles. Then*

$$\begin{gathered}(\cos\theta + i\sin\theta)(\cos\varphi + i\sin\varphi) =\\ \cos(\theta+\varphi) + i\sin(\theta+\varphi).\end{gathered}$$

This theorem is often written in a very compact way by making use of exponential notation. If we define $e^{i\theta} = \cos\theta + i\sin\theta$, then DeMoivre's Theorem takes the form $e^{i\theta}e^{i\varphi} = e^{i(\theta+\varphi)}$. Remarkably enough, this expression actually makes sense analytically, where $e = 2.71828\cdots$ is the base of the natural logarithm. In this book we will use this only as a formal shorthand for the more explicit expression involving the sine and cosine function. In Exercise 8.18 you will explore a derivation of the exponential form of DeMoivre's Theorem that depends on power series from calculus.

We can now interpret the computations of inverses in the field $\mathbb{C}$ geometrically. If we wish to compute the multiplicative inverse of the non-zero complex number $\alpha = |\alpha|(\cos\theta + i\sin\theta) = |\alpha|e^{i\theta}$, we need a complex number whose modulus is $1/|\alpha|$ (because the modulus of 1 is 1), and whose argument is $-\theta$ (because the argument of 1 is 0). What this means geometrically is this: Flip through the x-axis (which obtains the complex conjugate), and then adjust the length.


$(|\alpha|\cos\theta, |\alpha|\sin\theta)$
θ
$-\theta$
$(1/|\alpha|\cos(-\theta), 1/|\alpha|\sin(-\theta))$

In future chapters we will return often to the field of complex numbers, as a very important example, both practically and historically.

We have based our discussion of the field of complex numbers on your previous experience with the real numbers: The arithmetic of $\mathbb{C}$ comes entirely out of our understanding of the field properties of $\mathbb{R}$. Actually, it is possible to carefully prove that $\mathbb{Q}$ and $\mathbb{R}$ are fields, by basing their arithmetic on the arithmetic of the ring $\mathbb{Z}$. In this book we will not enquire into these details, and rely instead on your previous experience with these sets of numbers. We will continue to use $\mathbb{Q}$ and $\mathbb{R}$ as two of our most important examples of fields.

8.5 Finite Fields

To provide further examples of fields, we now turn our attention to the important collection of commutative rings, $\mathbb{Z}_n$. We will determine which of these are fields. From our examples above, we know that $\mathbb{Z}_5$ is a field while $\mathbb{Z}_6$ is not. These examples suggest the following theorem:

Theorem 8.5 *$\mathbb{Z}_n$ is a field if and only if n is prime.*

Proof: If n is not prime, then $n = xy$ for some $0 < x, y < n$. But then in $\mathbb{Z}_n$, $[x][y] = [0]$. So, $\mathbb{Z}_n$ could not be a field by Theorem 8.2.

Conversely, suppose n is prime and let $0 < x < n$. We need to establish the existence of $[x]^{-1}$. That is, we must find a y with $[x][y] = [1]$. Because n is prime, we know that $\gcd(n, x) = 1$. Then by the GCD identity 2.4, there exist r, s such that $rn + sx = 1$. But then $[s][x] = [1] - [rn] = [1]$, and so $[x]^{-1} = [s]$. □

The proof above has an interesting application: We can use it to compute multiplicative inverses in $\mathbb{Z}_p$, where p is prime.

Example 8.6

> Let's compute $[23]^{-1}$ in $\mathbb{Z}_{119}$. First apply Euclid's Algorithm to obtain the equation $1 = (6)(119) + (-31)(23)$. But then $[23]^{-1} = [-31] = [88]$.
>
> ▷ **Quick Exercise.** Check directly that $[23][88] = [1]$ in $\mathbb{Z}_{119}$. ◁

Thus, the general recipe for computing $[x]^{-1}$ in $\mathbb{Z}_p$ is this: Given x and p, apply Euclid's Algorithm to show that $\gcd(p, x) = 1$, and then work backward through the resulting equations to obtain r and s. Reducing s modulo p then gives the appropriate residue class. Notice that this method applies even if p is not prime, as long as $\gcd(p, x) = 1$. We can then conclude the following:

Theorem 8.6 *Let $0 < x < m$. Then $[x]$ is a unit in the ring $\mathbb{Z}_m$ if and only if $\gcd(x, m) = 1$.*

Proof: To show that $[x]$ has an inverse, merely repeat the argument above.

Conversely, if $\gcd(x, m) = d$ and $d \neq 1$, then $m = rd$ and $x = sd$, where r and s are integers with $m > r$, $s > 1$. But then $[x][r] = [sdr] = [sm] = [0]$. That is, $[x]$ is a zero divisor, and so cannot be a unit.

(Notice that if you have done Exercises 3.5 and 3.6, that you have already completed this proof!) □

For an alternative approach to computing $[x]^{-1}$ in $\mathbb{Z}_p$, we consider a beautiful and important theorem due to the 17th-century French mathematician Fermat.

Theorem 8.7 Fermat's Little Theorem *If p is prime and $0 < x < p$, then $x^{p-1} \equiv 1 (mod\ p)$. (Hence, in $\mathbb{Z}_p$, $[x]^{-1} = [x^{p-2}]$.)*

Example 8.7

> In $\mathbb{Z}_5$, this theorem asserts that $[3]^4 = [1]$. But then $[3][3]^3 = [1]$, and so $[3]^{-1} = [3]^3 = [27] = [2]$.

Proof: Suppose that p is prime and $0 < x < p$. Then $[x]$ is a non-zero element of the field $\mathbb{Z}_p$. Consider the set S of non-zero multiples of $[x]$ in $\mathbb{Z}_p$:

$$S = \{[x \cdot 1], [x \cdot 2], \ldots, [x \cdot (p-1)]\}.$$

Because a field has no zero divisors, each element of S is non-zero. Because a field satisfies multiplicative cancellation, no two of these elements are the same.

▷ **Quick Exercise.** Why? ◁

Thus, the set S consists of $p - 1$ distinct non-zero elements and so must consist of the set

$$\{[1], [2], \cdots, [p-1]\}$$

of all non-zero elements in $\mathbb{Z}_p$, in some order.

It might be helpful for you at this point in the proof to verify this fact in a particular case; see Exercise 8.8.

We now multiply all the elements of S together; this is the same as multiplying all the non-zero elements of $\mathbb{Z}_p$ together. We thus obtain the following equation in $\mathbb{Z}_p$:

$$[1][2]\cdots[p-1][x]^{p-1} = [1][2]\cdots[p-1].$$

But by multiplicative cancellation in the domain $\mathbb{Z}_p$, we may cancel the non-zero elements $[1],[2],\cdots,[p-1]$ from each side of this equation, leaving $[x]^{p-1}=[1]$. Or in other words, $x^{p-1}\equiv 1(\text{mod } p)$, as claimed. □

Example 8.8

As another example of this theorem, consider the element [3] in $\mathbb{Z}_7$. To compute $[3]^{-1}$ we do the following:

$$3^5 \equiv ((3^2)^2)(3) \equiv (9^2)(3) \equiv (2^2)(3) \equiv (4)(3) \equiv 12 \equiv 5.$$

Thus, $[3]^{-1}=[5]$; you can check directly that $[3][5]=[1]$.

We have now exhibited infinitely many distinct finite fields ($\mathbb{Z}_p$ for any prime p). It is natural to ask whether this is a complete list. We shall discover later that there are fields with finitely many elements which are not of the form $\mathbb{Z}_p$, but we will need to know a lot more about field theory. (See Chapter 46, and also Exercise 8.12.)

Note that we have *not* exhibited a finite domain that is not a field. There is good reason for this: The next theorem asserts that all finite domains are actually fields. Of course, the finiteness here is important; $\mathbb{Z}$ is an example of an infinite domain that is not a field.

Theorem 8.8 *All finite domains are fields.*

Proof: Suppose D is a finite domain with n elements. Then we can list the elements of D as $0, d_2, \cdots, d_n$. (Somewhere in this list must occur the multiplicative identity 1.) Suppose now that a is a *non-zero* element of D. It is then one of the d_i, where $i \geq 2$. Consider now the list of $n-1$ elements $ad_2, ad_3, \cdots, ad_n$. Because D is a domain, none of these elements is 0. Because multiplicative cancellation holds in a domain, no two of these elements are the same (because if $ad_i = ad_j$, then $d_i = d_j$ by cancellation). There are thus $n-1$ distinct elements in this list. This means the list consists of *all* the non-zero elements of D. But then $ad_i = 1$ for some i, because 1 is certainly one of the non-zero elements of D. This means that a is a unit. Hence, all non-zero elements of D are units, and so D is a field. □

You should think carefully about why the proof of the above theorem does not work when the domain in question has infinitely many elements. (See Exercise 8.10.)

Historical Remarks

Pierre Fermat was one of the most important mathematicians of the 17th century, even though he was an amateur. A lawyer by trade, he did most of his mathematical work in the evenings, as an intellectually stimulating recreation. Consequently, almost no mathematics under his name was published until after his death. He did engage actively in the mathematical life of his day, by corresponding with most of the great names in physics and mathematics of the era; such men as Huygens and Descartes were his correspondents. In this way, many of Fermat's results became well known to the mathematical community of the time. In the present day, mathematicians (and other scientists) communicate their results in widely available scholarly journals. In this way, scientific results can be easily used (and checked) by others. This sort of wide access to scientific knowledge was not yet available in the 17th century. The evolution toward that system began only with the founding of such scholarly societies as the Royal Society in Britain, near the end of the seventeenth century.

Fermat's Little Theorem is so called to distinguish it from his Great or Last Theorem. This asserts that there are no solutions in non-zero integers to the equation

$$x^n + y^n = z^n,$$

whenever $n > 2$. (Of course, there are many such solutions when $n = 2$; for example, $3^2+4^2=5^2$.) This assertion was found, without proof, in a margin of a book of Fermat's after his death. He noted that the margin was insufficiently large to contain his 'truly marvelous demonstration.' Because no mathematician for 350 years was able to prove this assertion, it seems highly unlikely that Fermat could prove it. Attempts to prove what might more properly be called Fermat's Last Conjecture have had an important impact on the development of abstract algebra. But note that if there had been a system of mathematical journals at that time, we might now have an answer to this famous problem, or at least know whether Fermat's proof was fallacious! In 1993 the British mathematician Andrew Wiles thrilled the mathematical world by announcing that he had proved Fermat's Last Theorem, basing his results on a large body of modern mathematical work. A gap in his proof was discovered as his work underwent the present-day scrutiny of peer review, but Wiles and his colleague Richard Taylor were able to fill in the gap, and complete the proof of a 350 year old conjecture.

Chapter Summary

In this chapter, we defined *integral domains* (commutative rings with unity without zero divisors) and *fields* (commutative rings in which all elements have multiplicative inverses).

We proved the following theorems about fields and domains:

- Multiplicative cancellation holds in integral domains.
- Every field is an integral domain.
- Every finite integral domain is a field.
- The ring $\mathbb{Z}_n$ is a field exactly if n is prime.
- Fermat's Little Theorem.

This last theorem gives a method for computing multiplicative inverses in $\mathbb{Z}_p$, for p prime.

Warm-up Exercises

a. Determine the units and the zero divisors in the following rings:

$$\mathbb{Z}\times\mathbb{Z},\quad \mathbb{Z}_{20},\quad \mathbb{Z}_4\times\mathbb{Z}_2,\quad \mathbb{Z}_{11},\quad \mathbb{Z}[x].$$

b. Suppose that a is a unit in a ring. Is $-a$ a unit? Why or why not?

c. Find two non-zero matrices A and B in $M_2(\mathbb{Z})$ so that $AB = 0$; that is, find some zero divisors.

d. Find a non-zero matrix A in $M_2(\mathbb{Z})$ so that $A^2 = 0$. Then A is a zero divisor. (A ring element a so that $a^n = 0$, for some positive integer n is called *nilpotent*. See Exercise 7.15.)

e. Suppose that D is a domain. Show that the direct product $D\times D$ is not a domain.

f. Compute the argument, modulus, and multiplicative inverse of the complex number $3-4i$.

g. What is the argument of a (non-zero) real number?

h. A complex number with modulus 1 is said to lie on the unit circle. Why?

i. Choose two complex numbers on the unit circle (see Exercise h). Why is their product also on the unit circle?

j. Consider the complex numbers $1-\sqrt{3}i$ and $2+2i$.

   (a) Compute their product twice: First do the arithmetic directly; then determine the arguments and moduli of these numbers, and compute the product using Theorems 8.3 and 8.4.

   (b) Compute their multiplicative inverses twice: First do the arithmetic directly, and then use Theorems 8.3 and 8.4 instead.

k. Write out the following complex numbers in the form $a+bi$, giving exact values for a and b if possible, or by using a calculator if necessary:

$$e^{\pi i},\quad e^{\frac{\pi i}{4}},\quad e^{i},\quad 2e^{\frac{2\pi i}{3}}$$

l. Give examples of the following, or explain why they don't exist:

   (a) A finite field.
   (b) A finite field that isn't a domain.
   (c) A finite domain that isn't a field.
   (d) An infinite field.

(e) An infinite domain that isn't a field.

m. Does there exist an integer m for which $\mathbb{Z}_m$ is a domain, but not a field? Explain.

n. Use Euclid's Algorithm to compute the multiplicative inverse of $[2]$ in $\mathbb{Z}_9$.

o. Use Fermat's Little Theorem 8.7 to compute the multiplicative inverse of $[2]$ in $\mathbb{Z}_5$.

Exercises

1. Prove that if R is a commutative ring and $a \in R$ is a zero divisor, then ax is also a zero divisor or 0, for all $x \in R$.

2. Consider the set $M_2(\mathbb{R})$ of all 2×2 matrices with entries from the real numbers $\mathbb{R}$; by Exercise 6.8, we know this is a ring. Prove that the units of $M_2(\mathbb{R})$ are precisely those matrices

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

such that $ad - bc \neq 0$. In this case, you can find a formula for the multiplicative inverse of the matrix. We call $ad - bc$ the **determinant** of the matrix. And so the units of $M_2(\mathbb{R})$ are those matrices with non-zero determinant; we will explore the notion of determinant further in Example 27.4.

3. Find two non-commuting units A, B in $M_2(\mathbb{R})$, and check that $(AB)^{-1} = B^{-1}A^{-1}$ and $(AB)^{-1} \neq A^{-1}B^{-1}$.

4. Generalizing Example 7.9, consider the subring $D_2(\mathbb{R})$ of diagonal matrices with real entries. What are the units of $D_2(\mathbb{R})$?

5. Let $\alpha = e^i = \cos 1 + i \sin 1$. (Note that the argument here is 1 *radian.*) What is the modulus of α^n, where n is a positive integer? Prove that $\alpha^n \neq \alpha^m$, whenever n and m are positive integers and $n \neq m$.

6. Use Fermat's Little Theorem 8.7 to find $[6]^{-1}$ in $\mathbb{Z}_{19}$.

7. Use Euclid's Algorithm to find $[36]^{-1}$ in $\mathbb{Z}_{101}$.

8. Verify explicitly the key idea in the proof of Fermat's Little Theorem 8.7 for $x = 3$ and $p = 17$; that is, check that the set S consists of the 16 non-zero elements of $\mathbb{Z}_{17}$.

9. Suppose that $b \in R$, a non-commutative ring with unity. Suppose that $ab = bc = 1$; that is, b has a **right inverse** c and a **left inverse** a. Prove that $a = c$ and that b is a unit.

10. Try to apply the proof that every finite integral domain is a field to the integral domain $\mathbb{Z}$. Where does it go awry?

11. Let R be a commutative ring with unity. Suppose that n is the least positive integer for which we get 0 when we add 1 to itself n times; we then say R has **characteristic n**. If there exists no such n, we say that R has **characteristic 0**. For example, the characteristic of $\mathbb{Z}_5$ is 5 because $1 + 1 + 1 + 1 + 1 = 0$, whereas $1 + 1 + 1 + 1 \neq 0$. (Note that here we have suppressed '[' and ']'.)

    (a) Show that, if the characteristic of a commutative ring with unity R is n and a is *any* element of R, then $na = 0$. (Recall that $na = \underbrace{a + a + \cdots + a}_{n \text{ times}}$.)

    (b) What are the characteristics of $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{Z}_{17}$?

    (c) Prove that if a field F has characteristic n, where $n > 0$, then n is a prime integer.

12. Consider the commutative ring

$$F = \{0, 1, \alpha, 1 + \alpha\},$$

where 0 is the additive identity, 1 is the multiplicative identity, $x + x = 0$, for all $x \in F$, and $\alpha^2 = \alpha + 1$.

    (a) Write out explicitly the addition and multiplication tables for F.

    (b) Prove that F is a field.

    (c) Because F has four elements, you might expect that F would be the 'same' as the ring $\mathbb{Z}_4$. Show this is false, by computing the characteristics of F and $\mathbb{Z}_4$ (see the previous exercise).

13. Suppose that R is a commutative ring and a is a non-zero nilpotent element. (See Exercise 7.15; this means that $a^n = 0$ for some positive integer n.) Prove that $1 - a$ is a unit. *Hint:* You can actually obtain a formula for the inverse.

14. Prove that $\mathbb{Z}_m$ is the union of three mutually disjoint subsets: its zero divisors, its units, and $\{0\}$. Show by example that this is false for an arbitrary commutative ring.

15. Prove that if p is prime, then $[(p-1)!] = [-1]$ in $\mathbb{Z}_p$. *Note:* In number theory, this result is known as Wilson's Theorem.

16. Suppose that
$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_2(\mathbb{R})$$
is a non-zero element, which is not a unit. Show that A is actually a zero divisor. (You might want to look at Exercise 2.) What is the relationship of this result to Exercise 14?

17. Consider the ring $C[0,1]$. (See Example 6.14.)
    (a) What are the units in this ring? (You need a theorem from calculus to prove your answer is correct.)
    (b) Given $f \in C[0,1]$, define $Z(f) = \{x \in [0,1] : f(x) = 0\}$. Prove that if f, g are associates, then $Z(f) = Z(g)$.

18. Recall the Taylor series expansions centered at 0 for the three functions $\sin x, \cos x, e^x$. (These are also called the MacLaurin series for these functions.) In calculus we discover that these series converge to their functions absolutely for all real numbers x. Let's assume that these series also make sense for imaginary numbers ix . By replacing x in e^x by ix, and rearranging terms of the series, verify that $e^{ix} = \cos x + i \sin x$, for all real numbers x. (This verification can be made rigorous, using analytic techniques we will avoid here.)

Chapter 9

Polynomials over a Field

In this chapter we generalize what we've learned about $\mathbb{Q}[x]$, the ring of polynomials with rational coefficients, by replacing the rational numbers by entries from an arbitrary field. We will discover that the resulting rings behave very similarly to $\mathbb{Q}[x]$, and hence to the ring $\mathbb{Z}$ as well.

9.1 Polynomials with Coefficients from an Arbitrary Field

Suppose we consider polynomials with coefficients not from $\mathbb{Q}$, but from some arbitrary field F. We denote this set of polynomials by $F[x]$. Addition and multiplication are defined as in $\mathbb{Q}[x]$, but when coefficients are added or multiplied, it is done in F.

Example 9.1

Consider $\mathbb{Z}_2[x]$, the set of polynomials with coefficients from $\mathbb{Z}_2$. (So coefficients are either 0 or 1.) Consider the sum and product of $x^2 + x + 1$ and $x^2 + 1$:

$$(x^2 + x + 1) + (x^2 + 1) = (x^2 + x^2) + x + (1 + 1) = x;$$

and

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

Notice how various of the above terms disappear, because $1 + 1 = 0$ in the ring $\mathbb{Z}_2$ of coefficients. Similar care must be taken with other finite fields.

▷ **Quick Exercise.** Compute the sum, difference, and product of $x^2 + 2x + 1$ and $x^2 + x + 2$ in $\mathbb{Z}_3[x]$ (here, the coefficients consist of 0, 1, or 2). ◁

It is clear that $F[x]$ (like $\mathbb{Q}[x]$) is a commutative ring with unity; in fact, it is an integral domain. But these two rings have much more in common than that.

▷ **Quick Exercise.** Convince yourself that $F[x]$ is an integral domain. ◁

If you return to Chapters 4 and 5 and look at the theorems (and their proofs), you'll see that every theorem about $\mathbb{Q}[x]$ is valid if $\mathbb{Q}$ is replaced by a field F of your choice. The properties of $\mathbb{Q}$ that are important in those theorems are the field properties of $\mathbb{Q}$. As before, the theorem that is the driving force is the Division Theorem.

▷ **Quick Exercise.** Prove the Division Theorem for $F[x]$, by carefully re-reading the proof of this theorem for $\mathbb{Q}[x]$. You will see that multiplicative inverses for coefficients are required in this proof, but nothing else special about $\mathbb{Q}$. ◁

So, the Division Theorem, the Root Theorem, Euclid's Algorithm, the GCD identity, and the Unique Factorization Theorem hold for $\mathbb{R}[x]$, $\mathbb{C}[x]$, and $\mathbb{Z}_p[x]$ (for prime p), as well as $F[x]$ where F is any other field. Notice that to make sense of these theorems in this new and more general context, we must be careful also to define such terms as **irreducible**, **prime**, and **associate** here.

▷ **Quick Exercise.** Check that our definitions for these terms for $\mathbb{Q}[x]$ still make sense for $F[x]$. ◁

▷ **Quick Exercise.** State the Unique Factorization Theorem for $F[x]$, where F is an arbitrary field. ◁

It is perhaps just as well to pause here a moment, and make certain that you are aware of what we have just done. We have just asserted (and you have checked!) that the statement and proofs of the Division Theorem, the Root Theorem, Euclid's Algorithm, the GCD identity, and the Unique Factorization Theorem generalize when $\mathbb{Q}$ is replaced by any field whatsoever. This is a striking example of the power of generalization and abstraction in algebra. We will devote the rest of this chapter to examining many examples of these theorems in action, using various fields (other than $\mathbb{Q}$) for the coefficients on our polynomials.

Example 9.2

Let's find a gcd of $x^3 + x + 1$ and $x^2 + 1$ in $\mathbb{Z}_2[x]$, using Euclid's Algorithm:

$$\begin{aligned} x^3 + x + 1 &= (x^2 + 1)x + 1 \\ x^2 + 1 &= 1(x^2 + 1) + 0. \end{aligned}$$

Therefore, 1 is a gcd of $x^3 + x + 1$ and $x^2 + 1$. Now any other gcd of these two polynomials would be an associate of 1. That is, it would be a non-zero scalar multiple of 1. But in $\mathbb{Z}_2[x]$, the only non-zero scalar is 1 itself. Thus, in $\mathbb{Z}_2[x]$, 1 is the *unique* gcd of $x^3 + x + 1$ and $x^2 + 1$.

▷ **Quick Exercise.** Write 1 as a linear combination of x^3+x+1 and $x^2 + 1$ in $\mathbb{Z}_2[x]$. ◁

Example 9.3

Let's now compute a gcd of $x^5 + x^2 + 1$ and $x^2 + 2$ in $\mathbb{Z}_3[x]$:

$$\begin{aligned} x^5 + x^2 + 1 &= (x^2 + 2)(x^3 + x + 1) + (x + 2) \\ x^2 + 2 &= (x + 2)(x + 1) + 0. \end{aligned}$$

Therefore, $x + 2$ is a gcd of $x^5 + x^2 + 1$ and $x^2 + 2$ in $\mathbb{Z}_3[x]$.

▷ **Quick Exercise.** Divide $x^5 + x^2 + 1$ by $x + 2$ in $\mathbb{Z}_3[x]$. ◁

▷ **Quick Exercise.** List *all* the gcds of $x^5 + x^2 + 1$ and $x^2 + 2$ in $\mathbb{Z}_3[x]$. ◁

9.2 Polynomials with Complex Coefficients

Given a quadratic polynomial with real coefficients, you probably recall the **quadratic formula** that provides the roots for the polynomial;

you may recall that the formula is proved by means of an algebraic technique called *completing the square.*

To demonstate this, suppose that $f = ax^2+bx+c \in \mathbb{R}[x]$, with $a \neq 0$. To solve the equation $ax^2 + bx + c = 0$ we can do a little algebra to obtain $x^2 + \frac{b}{a}x = -\frac{c}{a}$. To make the left side of the equation a perfect square, we add the term $\left(\frac{b}{2a}\right)^2$ to both sides of the equation. With a little further algebraic simplification (see Exercise 9.1), we obtain the usual form for the quadratic formula:

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

The quantity $D = b^2 - 4ac$ is called the **discriminant** of f. Clearly the two roots are equal if $D = 0$, and we get two distinct real roots if $D > 0$.

But if $D < 0$, the quadratic formula gives two distinct complex roots, which are conjugates of one another. But by the Root Theorem these two roots also give a factorization of f (in $\mathbb{C}[x]$) into linear polynomials:

$$f = ax^2 + bx + c = a\left(x - \frac{-b + \sqrt{b^2 - 4ac}}{2a}\right)\left(x - \frac{-b - \sqrt{b^2 - 4ac}}{2a}\right).$$

But in this case it is now clear that f *cannot* be non-trivially factored in $\mathbb{R}[x]$. For if it did so factor, it would then also have real roots, making more than two roots for f in $\mathbb{C}$, which is impossible.

We have thus shown that if $f \in \mathbb{R}[x]$ is a quadratic polynomial with negative discriminant, then it is an irreducible element of $\mathbb{R}[x]$. But any such quadratic polynomial *can* be factored further in $\mathbb{C}[x]$.

Example 9.4

Let's factor the polynomial

$$f = x^3 - 7x^2 + 17x - 15 \in \mathbb{R}[x],$$

into irreducibles, in both $\mathbb{R}[x]$ and $\mathbb{C}[x]$. By the Root Theorem, $x - 3$ is a factor of f because $f(3) = 0$. We can then factor $x - 3$ out of f:

$$f = (x - 3)(x^2 - 4x + 5).$$

Because $x-3$ is linear, it is irreducible. Now consider x^2-4x+5. Its discriminant is $D = -4 < 0$, and so $x^2 - 4x + 5$ is the other irreducible factor of f in $\mathbb{R}[x]$. But if we apply the quadratic formula we obtain two linear irreducible factors in $\mathbb{C}[x]$:

$$f = (x - 3)(x - 2 + i)(x - 2 - i).$$

The example we've just examined, showing that more factorization is possible in $\mathbb{C}[x]$ than in $\mathbb{R}[x]$, is a particular example of an important general fact about $\mathbb{C}[x]$: *Every* non-constant polynomial in $\mathbb{C}[x]$ can be factored into linear factors. This fact is known as the *Fundamental Theorem of Algebra*, a very important theorem first proved rigorously by Gauss. We won't prove this theorem here. The easiest proof of the Fundamental Theorem of Algebra uses complex analysis and is accessible only to those who have taken an introductory course in this subject. We will state the Fundamental Theorem in an (apparently weaker) form, and then prove the statement we've made above as a corollary.

Theorem 9.1 Fundamental Theorem of Algebra *Every non-constant polynomial in $\mathbb{C}[x]$ has a root in $\mathbb{C}$.*

Proof: We omit this proof, and refer the reader to any introductory text in complex analysis. □

Corollary 9.2 *Every non-constant polynomial with degree n in $\mathbb{C}[x]$ is linear or can be factored as a product of n linear factors. Thus, the irreducibles in $\mathbb{C}[x]$ consist exactly of the linear polynomials.*

Proof: Let f be a polynomial in $\mathbb{C}[x]$. The essence of this proof is to apply the Fundamental Theorem of Algebra to f and its factors, over and over again. We make this precise by means of an induction argument on n, the degree of f. If $n = 1$, then the polynomial is itself a linear polynomial. Now suppose $n > 1$. By the Fundamental Theorem of Algebra, f has a root α. By the Root Theorem, $x - \alpha$ is a factor of f, and so $f = (x - \alpha)g$, where $\deg(g) = n - 1$. By the induction hypothesis, g is either linear or a product of linear factors, and thus f itself is a product of linear factors.

We've already observed that linear polynomials are *always* irreducible. Any polynomial in $\mathbb{C}[x]$ of higher degree can be factored and so is not irreducible. □

One important observation needs to be made about the Fundamental Theorem of Algebra: It merely asserts the *existence* of the linear factors. Neither in the statement of the theorem nor in its (omitted) proof is an effective method provided for actually finding them. We will return to this subject in the final section of this book, on what is called *Galois Theory.*

9.3 Irreducibles in $\mathbb{R}[x]$

We now use the Fundamental Theorem of Algebra to find all irreducible polynomials in $\mathbb{R}[x]$. We have already proved half of the following theorem:

Theorem 9.3 *The irreducible polynomials in $\mathbb{R}[x]$ are the linear polynomials, and those quadratic polynomials with negative discriminant.*

Proof: We know already that linear polynomials are irreducible. And in our discussion of the quadratic formula in Section 9.2 above we have argued already that if a quadratic polynomial has a negative discriminant, then it is irreducible in $\mathbb{R}[x]$. We now will argue that linear polynomials and quadratics with negative discriminant are the *only* irreducibles in $\mathbb{R}[x]$.

Suppose that f is a non-linear irreducible in $\mathbb{R}[x]$. Then f can have no real roots (else it would have a linear factor, by the Root Theorem). But f has complex roots, by the Fundamental Theorem of Algebra. Let α be a complex root of f. Suppose that $\alpha = s + ti$, where s and t are real and $t \neq 0$. We now use $\bar{\alpha} = s - ti$, the complex conjugate of α which we discussed in the previous chapter. We define

$$\begin{aligned} g = (x-\alpha)(x-\bar{\alpha}) &= (x-(s+ti))(x-(s-ti)) \\ &= x^2 - 2sx + (s^2+t^2), \end{aligned}$$

which is a polynomial in $\mathbb{R}[x]$. We now apply the Division Theorem for $\mathbb{R}[x]$. By this theorem, $f = gq + r$, where f, g, q, and r all have real coefficients. Now think of these polynomials as being in $\mathbb{C}[x]$. So,

$$0 = f(\alpha) = g(\alpha)q(\alpha) + r(\alpha).$$

Because $g(\alpha) = 0$, we have $r(\alpha) = 0$. But $\deg(r) < 2$ and so $r = cx + d$ with $c, d \in \mathbb{R}$. But $r(\alpha) = 0$; that is, $c\alpha + d = 0$. But α is not real, and therefore we must have that $c = d = 0$. Hence, $f = gq$, and so f is not irreducible unless q is a scalar. But then f is of degree two with negative discriminant, which proves what we wanted. □

9.4 Extraction of Square Roots in $\mathbb{C}$

We now know that we can use the quadratic formula to factor *any* quadratic polynomial in $\mathbb{R}[x]$, over the complex numbers. But what about quadratic polynomials in $\mathbb{C}[x]$? We know by the Fundamental Theorem of Algebra that such polynomials factor and thus must factor into two linear factors. Surely the quadratic formula should still work. But we must extract a square root when using the quadratic formula, and if the original quadratic polynomial belongs to $\mathbb{C}[x]$, this means extracting the square root of an arbitrary complex number.

Example 9.5

Consider the quadratic equation

$$x^2 - ix - (1+i) = 0.$$

If we proceed with the quadratic formula, we obtain the following:

$$x = \frac{i \pm \sqrt{3+4i}}{2}.$$

▷ **Quick Exercise.** Check the above computation. ◁

Unfortunately, we have a complex number under the radical sign. If we are lucky enough to observe that $(2+i)^2 = 3+4i$, we can then obtain the two roots $x = -1$ and $x = 1+i$.

▷ **Quick Exercise.** Check our arithmetic, and verify that the given complex numbers are roots of the original equation. ◁

In the previous example we were lucky. But how can we extract square roots of arbitrary complex numbers? We make use of the geometric representation for complex numbers we discussed in the previous chapter.

Suppose that $\alpha = a+bi \in \mathbb{C}$. We wish to find roots to the polynomial $x^2 - \alpha$. From the previous chapter, we can factor α as

$$\alpha = |\alpha|(\cos\theta + i\sin\theta) = |\alpha|e^{i\theta},$$

where θ is the argument of α. We wish to find a complex number whose square is this. But because the modulus of a product is the product of the moduli (Theorem 8.3), the modulus of a square root of α must be $\sqrt{|\alpha|}$. And one argument that will surely work is $\theta/2$. Thus,

$$\beta = \sqrt[4]{a^2+b^2}(\cos(\theta/2) + i\sin(\theta/2))$$

is a square root of the complex number α, and of course $-\beta$ is another. Because the polynomial $x^2 - \alpha$ can have at most two roots, these must be the *only* roots it has; these roots are distinct unless $\alpha = 0$. Thus, every non-zero complex number has exactly two square roots.

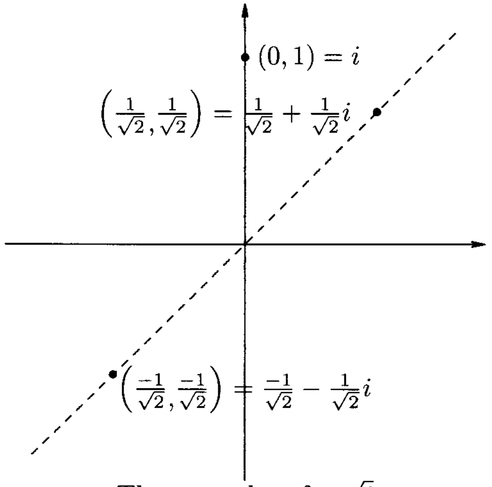
Example 9.6

Let's compute the square roots of i. Its modulus is 1, and so its square roots will also have modulus 1. Its argument is $\pi/2$, and so one square root must be

$$\cos(\pi/4) + i\sin(\pi/4) = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}i$$

and the negative of this is the other.

▷ **Quick Exercise.** Check this explicitly by squaring both these complex numbers. ◁


$(0,1) = i$
$\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}i$
$\left(\frac{-1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right) = \frac{-1}{\sqrt{2}} - \frac{1}{\sqrt{2}}i$

The two values for $\sqrt{i}$

▷ **Quick Exercise.** Apply this method to compute the square roots of $1+\sqrt{3}i$. ◁

In Exercise 9.23, you will extend this method to compute the nth roots of an arbitrary complex number.

Historical Remarks

Naturally, given a particular polynomial with real coefficients it may be *very* difficult to come up with the factorization into irreducibles, particularly if the degree of the polynomial is large. Of course, the quadratic formula allows us to find such a factorization for all polynomials of degree two, and there exist progressively more complicated cubic and quartic formulas, which allow us to factor all polynomials of degree 3 and degree 4. (See Exercise 9.12 for the degree 3 case, and Exercise 9.20 for degree 4.) However, a difficult theorem due to the 19th-century Norwegian mathematician Abel states that there are polynomials of degree 5 or higher for which *no* explicit determination of the roots is possible, using the ordinary operations of addition, subtraction, multiplication, division, and extraction of roots. We will encounter this theorem in the last chapter in this book.

The Fundamental Theorem of Algebra was one of Karl Gauss's favorites; he returned to it several times over the course of his career, proving it by different means. He had his first proof in hand before his twentieth year. The theorem had been conjectured (and believed) by such predecessors of Gauss as Lambert and Legendre. One of the most important obstacles preventing a proof before the time of Gauss was that there still remained doubts in the minds of most mathematicians as to the status of complex numbers. You may yourself when first encountering complex numbers have doubted they were as 'real' as real numbers; it is precisely this attitude which is reflected in the terms 'real' and 'imaginary'. Gauss was among the first mathematicians to use complex numbers with confidence and rigor. He made full use of the geometric interpretation we have discussed. In part, this reduction of complex arithmetic to geometry gave some reassurance to those who doubted the possibility of such calculations. Such doubts (which had been the rule in mathematical circles since the 16th century) were forgotten in the generation of mathematicians after Gauss.

Chapter Summary

In this chapter we discussed how the ring $F[x]$ of polynomials with coefficients from a field F has properties analogous to those of $\mathbb{Q}[x]$. In particular, this means that the *Division Theorem*, *Euclid's Algorithm*, the *GCD Identity*, the *Unique Factorization Theorem*, and the *Root Theorem* hold for $F[x]$.

We also considered the important examples $\mathbb{C}[x]$ and $\mathbb{R}[x]$. We stated the *Fundamental Theorem of Algebra* and inferred from it that the irreducible polynomials in $\mathbb{R}[x]$ are exactly the linear polynomials, and the quadratic polynomials with negative discriminant.

Warm-up Exercises

a. Calculate the quotient and remainder for the following, in various rings $F[x]$:

   (a) $x^3 + 4x + 1$ by $x + 2$, in $\mathbb{Z}_5[x]$.

   (b) $x^3 + 4x + 1$ by $2x + 1$, in $\mathbb{Z}_5[x]$.

   (c) $x^3 - (2 + i)x^2 + 5$ by $ix - 1$, in $\mathbb{C}[x]$.

   (d) $x^4 - 2x^3 + \frac{1}{3}$ by $\pi x + 1$, in $\mathbb{R}[x]$.

b. Use the Root Theorem to check for roots of $x^4 + 4$ in $\mathbb{Z}_5[x]$. Use your result to completely factor this polynomial.

c. Why are $x + i$ and $(1 + i)x + (-1 + i)$ associates in $\mathbb{C}[x]$?

d. List all associates of $2x^2 + 3x + 3$ in $\mathbb{Z}_5[x]$. Is it clearer how to factor this polynomial, if you consider one of its associates?

e. Why is $x^2 + 2$ irreducible in $\mathbb{Z}_5[x]$?

f. Factor $x^3 - 2$ into irreducibles in

$$\mathbb{Q}[x],\ \mathbb{R}[x],\ \mathbb{C}[x].$$

   Repeat this problem for $x^2 + \pi$.

g. Extract the square roots of $-3 + \sqrt{3}i$ in $\mathbb{C}$.

h. Let F be a field. Could the ring $F[x]$ be a field? Why or why not?

Exercises

1. Use the method of completing the square to complete the proof of the quadratic formula for finding roots of polynomials of degree two in $\mathbb{R}[x]$, as begun at the beginning of Section 9.2.

2. (a) Why does every non-zero complex number have exactly two square roots?

   (b) Given part a, check that the proof of the quadratic formula obtained in Exercise 9.1 still holds in $\mathbb{C}[x]$.

   (c) Use the quadratic formula to compute the roots of the polynomials

$$x^2 - (3 + 2i)x + (1 + 3i) \text{ and } x^2 - (1 + 3i)x + (-2 + 2i).$$

3. Give examples of two different polynomials in $\mathbb{Z}_5[x]$ that are identical as functions over $\mathbb{Z}_5$. This shows that equality of polynomials in $F[x]$ cannot be thought of as equality of the corresponding polynomial *functions*. (See the Quick Exercise in Section 4.1 for the $F = \mathbb{Z}_2$ case, and Exercise 4.12 for the case $F = \mathbb{Z}_3$.)

4. Consider the polynomial $f = x^3 + 3x^2 + 2x \in \mathbb{Z}_6[x]$. Show that this polynomial has more than three roots in $\mathbb{Z}_6$. Why doesn't this contradict the Root Theorem?

5. Find a gcd of $x^3 + 4x^2 + 4x + 9$ and $x^2 + x - 2$ in $\mathbb{Q}[x]$; in $\mathbb{R}[x]$; in $\mathbb{C}[x]$.

6. Find a gcd of $x^3 + x^2 + x$ and $x^2 + x + 1$ in $\mathbb{Z}_3[x]$; in $\mathbb{Z}_5[x]$; in $\mathbb{Z}_{11}[x]$.

7. Write the gcd you found in Exercise 5 as a linear combination of the two polynomials involved.

8. Show that if f is a polynomial with real coefficients and $\alpha = s+ti$ is a root of f in $\mathbb{C}$, then so is $\bar{\alpha} = s - ti$.

9. Use Exercise 8 to construct a polynomial in $\mathbb{R}[x]$ with roots $\frac{1}{2}$, i, $2-i$.

10. Factor $x^4 + x^3 + 2x^2 + 1$ into irreducibles in $\mathbb{Z}_3[x]$. Be sure to prove that your factors are in fact irreducible.

11. Determine all the irreducible elements of $\mathbb{Z}_2[x]$, with degree less than or equal to 4.

12. In this exercise, we describe the cubic formula for factoring an arbitrary polynomial of degree 3 in $\mathbb{R}[x]$. This version of the formula is called the *Cardano-Tartaglia* formula, after two 16th-century Italian mathematicians involved in its discovery. Consider the polynomial $f = x^3 + ax^2 + bx + c \in \mathbb{R}[x]$ (by dividing by the lead coefficient if necessary, we have assumed without loss of generality that it is 1).

    (a) Show that the change of variables $x = y - \frac{1}{3}a$ changes f into a cubic polynomial that lacks a square term; that is, a polynomial of the form $g = f(y - \frac{1}{3}a) = y^3 + py + q = 0$. *Note:* This process is called *depressing the conic.* Clearly we can solve $f = 0$ for x if and only if we can solve $g = 0$ for y.

    (b) Find explicit solutions u, v to the pair of simultaneous equations

    $$\begin{aligned} v^3 - u^3 &= q \\ uv &= \frac{1}{3}p. \end{aligned}$$

    Hint: These equations reduce to a *quadratic* equation in u^3 or v^3.

    (c) Prove the identity

    $$(u - v)^3 + 3uv(u - v) + (v^3 - u^3) = 0$$

    and use it to show that $y = u - v$ is a solution to the cubic equation $y^3 + py + q = 0$.

    (d) Let $D = q^2 + \frac{4p^3}{27}$. (This is called the *discriminant* of the conic.) Conclude that

    $$y = \sqrt[3]{\frac{-q + \sqrt{D}}{2}} - \sqrt[3]{\frac{q + \sqrt{D}}{2}}$$

    is a root for $g = 0$. (This is just $u - v$.)

13. In Exercise 12, there is an apparent ambiguity arising from the plus or minus when extracting the square root of D to obtain values for u^3 and v^3. However, show that we obtain the same value for the root $u - v$, regardless of which choice is made.

14. Apply the Cardano-Tartaglia formula to find a root of the cubic equation

    $$x^3 = 6x + 9.$$

    Then factor $x^3 - 6x - 9$, and use the quadratic formula to obtain the remaining two roots.

15. Suppose as in Exercise 12 that $g = y^3 + px + q$ is a cubic polynomial with real coefficients, and $y = u - v$ is the root given by the Cardano-Tartaglia formula. Suppose that $D > 0$. (Thus u and v are real numbers.) Let $\zeta = e^{\frac{2\pi}{3}}$ be a cube root of unity (called the primitive cube root of unity in Exercise 25 below). Argue that the other two distinct roots of $g = 0$ are the complex conjugates $u\zeta - v\zeta^2$ and $u\zeta^2 - v\zeta$. *Note:* Be sure and check both that these are roots and that they are necessarily distinct.

16. Apply Exercise 15 to obtain all three of the roots for the cubic $y^3 + 3y + 1$.

17. An interesting and surprising conclusion one can draw from Exercise 15 is that if the discriminant $D > 0$, then the cubic polynomial $y^3 + px + q \in \mathbb{R}[x]$ necessarily has exactly one real root, and a conjugate pair of complex roots. In this exercise you will use elementary calculus to verify this fact again:

    (a) Consider the function $g(y) = y^3 + px + q$. Suppose that $p > 0$. Compute the derivative $g'(y)$, and use it to argue that g has exactly one real root, and consequently two complex roots.

    (b) Suppose now that $p = 0$. Then conclude that $q \neq 0$. In this simple case, what are the roots of g?

    (c) Now suppose that $p < 0$. Compute the two roots of $g'(y) = 0$. Argue that the values of g at these two roots are both positive (using the assumption that $D > 0$). Why does this mean that g has exactly one real root?

18. (a) Find a cubic polynomial whose roots are $1 + \sqrt{3}, 1 - \sqrt{3}, -3$. *Hint:* Use the Root Theorem.

(b) Apply the Cardano-Tartaglia cubic formula derived in the previous exercise to solve the cubic obtained in part a.

(c) The answer you have obtained should be one of the roots you started with, but it does not *appear* to be. Can you explain this?

(d) Obtain the other two roots for this polynomial, by using the same strategy as in Exercise 15. Note that we must use complex arithmetic to obtain these three real numbers!

19. Exercise 18 is a particular example of what is called the *irreducible case* for a real cubic. Show that in case $D < 0$, we obtain a real root for the polynomial $g = y^3 + px + q$ by an appropriate choice for u and v.

20. In this problem we will explore Ferrari's approach to solving the general quartic equation. Consider the arbitrary quartic

$$f = x^4 + a_3x^3 + a_2x^2 + a_1x + a_0 \in \mathbb{R}[x].$$

(a) Find a linear change of variables $y = x + m$ so as to *depress* the quartic – that is, to eliminate the cubic term, as we eliminated the quadratic term in Exercise 12.

(b) We may by part a assume that our quartic equation is of the form $x^4 = px^2 + qx + r$, where $p, q, r \in \mathbb{R}$. Add the term $2bx^2 + b^2$ to both sides of this equation. Clearly this makes the left-hand side of the equation a perfect square. We would like to choose b so that the right-hand side is also a perfect square. Obtain an equation for b (in terms of p, q, r) that makes this true. The equation you obtain should be a cubic equation in b. Explain why a (real) solution to this cubic will always lead to a solution to the quartic equation. How do you then get all four solutions?

21. Carry out Ferrari's method for solving quartic equations, for the equation $x^4 = -x^2 - 4x + 3$.

22. Use a calculator and our algorithm to compute (approximations of) the square roots of $5 + 11.6i$ in $\mathbb{C}$.

23. Suppose that the complex number $\alpha = a + bi$ has been factored as

$$\alpha = \sqrt{a^2 + b^2}(\cos\theta + i\sin\theta) = |\alpha|e^{i\theta},$$

as we did in the text when computing the square roots of α.

(a) Show that

$$\beta_k = \sqrt[2n]{a^2 + b^2}\left(\cos\left(\frac{\theta + 2\pi k}{n}\right) + i\sin\left(\frac{\theta + 2\pi k}{n}\right)\right),$$

for $k = 0, 1, 2, \cdots n - 1$, are all nth roots of α.

(b) Show that these are all distinct roots.

(c) Why is this the *complete* list of nth roots of α?

24. What are the five complex fifth roots of 1? What are the five complex fifth roots of $1 + i$?

25. We generalize the previous exercise. Let p be a positive prime integer, and consider the cyclotomic polynomial

$$\Phi_p = \frac{x^p - 1}{x - 1} = x^{p-1} + x^{p-2} + \cdots + x + 1$$

first encountered in Exercise 5.17. Clearly its roots (together with 1) are exactly the p distinct pth roots of 1; these are called the pth **roots of unity**. If we set

$$\zeta = \cos(2\pi/p) + i\sin(2\pi/p) = e^{\frac{2\pi i}{p}},$$

show that the pth roots of unity are precisely

$$1,\ \zeta,\ \zeta^2,\ \cdots\zeta^{p-1}.$$

We call ζ the **primitive** pth root of unity.

26. A field F is said to be **algebraically closed** if every polynomial $f \in F[x]$ with $\deg(f) \geq 1$ has a root in F; we can rephrase this definition roughly by saying that a field is algebraically closed if it satisfies the Fundamental Theorem of Algebra. Thus, $\mathbb{C}$ is algebraically closed, while $\mathbb{R}$ and $\mathbb{Q}$ are not. Show that for every prime p, the field $\mathbb{Z}_p$ is not algebraically closed.

27. Show that the field in Exercise 8.12 is not algebraically closed. (See the previous exercise for a definition.)

28. Show that, if F is a field with infinitely many elements, then $f(x) = g(x)$ for all $x \in F$ implies that $f = g$ as polynomials. (We have already seen that this is not the case if F is a finite field. For example, consider $x^2 + x + 1$ and 1 in $\mathbb{Z}_2[x]$.)

Section II in a Nutshell

This section defines three important algebraic structures: rings, integral domains and fields.

Well-known objects ($\mathbb{Z}$, $\mathbb{Q}[x]$, $\mathbb{Z}_m$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$) share many algebraic properties. These properties define an abstract object called a *ring*:

A *ring* R is a set of elements on which two binary operations, addition ($+$) and multiplication ($\cdot$), are defined that satisfy the following properties for all $a, b, c \in R$:

1. (*Addition is commutative*) $a + b = b + a$
2. (*Addition is associative*) $(a + b) + c = (a + (b + c)$
3. (*Additive identity exists*) There exists an element 0 in R such that $a + 0 = a$.
4. (*Additive inverses exist*) For each element a in R, there exists an element x such that $a + x = 0$.
5. (*Multiplication is associative*) $(a \cdot b) \cdot c = a \cdot (b \cdot c)$
6. (*Multiplication distributes over addition*) $a \cdot (b + c) = a \cdot b + a \cdot c$, $(b + c) \cdot a = b \cdot a + c \cdot a$

Note that the multiplication in a ring need not be commutative. A ring where multiplication is commutative is called a *commutative ring*, naturally. All the examples of rings we've listed above are commutative rings. An example of a non-commutative ring is $M_2(\mathbb{Z})$, the collection of 2×2 matrices with integer entries.

Addition in rings has some useful properties:

(Theorem 6.1) Suppose R is a ring and $a, b \in R$.

1. (*Additive Cancellation*) If $a + b = a + c$, then $b = c$.
2. (*Solution of equations*) The equation $a + x = b$ always has a unique solution in R.

3. (*Uniqueness of additive inverses*) Every element of R has exactly one additive inverse.

4. (*Uniqueness of additive identity*) There is only one element of R that satisfies the equations $z + a = a$, for all a; namely, the element 0.

A *subring* S of ring R is subset of R that is itself a ring under the operations induced from R. To determine whether a subset of a ring is a subring, the *Subring Theorem* asserts that it is not necessary to verify all the rules that define a ring:

(Theorem 7.1) A non-empty subset of a ring is a subring under the same operations if and only if it is closed under multiplication and subtraction.

Some rings R have a *unity*, or *multiplicative identity*; that is, an element $u \in R$ where $au = ua = a$ for all $a \in R$. The number 1 is the unity in $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$. The scalar polynomial 1 is the unity in $\mathbb{Q}[x]$ and $\mathbb{Z}[x]$. And the matrix $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ is the unity in $M_2(\mathbb{Z})$. The residue class [1] is the unity in $\mathbb{Z}_m$. But $2\mathbb{Z}$ has no unity. If the unity exists, it is unique.

An element $A \neq 0$ is a *zero divisor* if there is an element $b \neq 0$ with $ab = 0$. A commutative ring with unity is an *integral domain* if it has no zero divisors. $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, $\mathbb{Q}[x]$ and $\mathbb{Z}[x]$ are all integral domains. $\mathbb{Z}_n$ is an integral domain if and only if n is prime. $M_2(\mathbb{Z})$ is not an integral domain.

Integral domains have the nice property of multiplicative cancellation:

(Theorem 8.1) If R is an integral domain and $a, b, c \in R$ with $a \neq 0$, then $ab = ac$ implies that $b = c$.

If R is a ring with unity 1, then an element $a \in R$ is a *unit* if there exists $b \in R$ such that $ab = 1$. In this case, b is said to be the *multiplicative inverse* of a. If all the non-zero elements of a commutative ring with unity are units, then we say the ring is a *field*. The rings $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are all fields but $\mathbb{Z}$, $\mathbb{Q}[x]$ and $\mathbb{Z}[x]$ are not fields. All fields are integral domains (Theorem 8.2). $\mathbb{Z}_p$ is a field, for prime p (Theorem 8.5). Indeed, all finite integral domains are fields (Theorem 8.8).

The field $\mathbb{Z}_p$ is the setting for the important Fermat's Little Theorem:

(Theorem 8.7) If p is prime and $0 < x < p$, then $x^{p-1} \equiv 1 (\text{mod } p)$. (Thus, $x^{p-2} \equiv x^{-1}$ in $\mathbb{Z}_p$.)

Finally, consider $F[x]$, the polynomials over an arbitrary field F. $F[x]$ is an integral domain. The Division Theorem, the Root Theorem, Euclid's Algorithm, the GCD identity and Unique Factorization all hold for $F[x]$. A particularly important example is $F = \mathbb{C}$, the field of complex numbers. $\mathbb{C}[x]$ satisfies the Fundamental Theorem of Algebra 9.1: every non-constant polynomial in $\mathbb{C}$ has a root.

III

Unique Factorization

Chapter 10

Associates and Irreducibles

In this chapter we begin to lay the groundwork necessary to explore those domains for which a unique factorization theorem is true. We wish to obtain a common generalization of $\mathbb{Z}$ and $\mathbb{Q}[x]$.

In both $\mathbb{Z}$ and $\mathbb{Q}[x]$, we ended up factoring elements into *irreducibles.* Roughly speaking, this means elements that admit no 'non-trivial' factorizations. But what do we mean by 'non-trivial'? In the case of the integers we disregarded factors of ± 1, while in the ring of polynomials with coefficients from a field we disregarded scalar factors (that is, polynomials of degree 0). In either case, these trivial factors amount exactly to those elements of the domain that have multiplicative inverses; that is, the *units.*

10.1 Associates

Recall from Chapter 5 that two polynomials that are scalar multiples of one another are called *associates.* This idea can be generalized: Two elements a and b of a commutative ring with unity are **associates** if there exists a unit u such that $a = ub$. Thus, if we speak intuitively, two elements are associates if they differ by only a 'trivial' factor.

Example 10.1

The set of units of $\mathbb{Z}$ consists of exactly $\{1, -1\}$, and so two elements are associates if they differ only by a factor of ± 1.

Example 10.2

The set of units of $\mathbb{Q}[x]$ consists of the non-zero constant polynomials (which we can identify with the non-zero rational numbers) and two polynomials are associates if they differ only by a (non-zero) rational multiple. Of course, this is exactly the definition we made of 'associate' in Section 5.2, and so our general definition is consistent.

Example 10.3

In $\mathbb{Z}_{12}$, the units are $\{1, 5, 7, 11\}$; 4 and 8, for example, are associates, because $5 \cdot 4 = 8$.

▷ **Quick Exercise.** What are the other associates of 4 in $\mathbb{Z}_{12}$? What are the associates of 4 in $\mathbb{Z}_{20}$? ◁

Example 10.4

In a field, *all* non-zero elements are units, and so in this case all (non-zero) elements are associates of one another.

10.2 Irreducibles

We now define what we mean by an irreducible element for an arbitrary commutative ring, generalizing our earlier notions for $\mathbb{Z}$ and $\mathbb{Q}[x]$. A non-zero element p of a commutative ring R is **irreducible** if

(1) it is a non-unit, and

(2) if whenever $p = ab$, then (exactly) one of a and b is a unit.

Example 10.5

The irreducibles in $\mathbb{Z}$ are exactly the prime numbers (and their negatives).

Example 10.6

The irreducibles in $\mathbb{R}[x]$ are exactly the linear polynomials, together with the quadratics with negative discriminant. (See Theorem 9.3.)

Example 10.7

The irreducibles in $\mathbb{Q}[x]$ include all linear polynomials, in addition to such polynomials as $x^2 + 1$, $x^3 - 2$, and others.

In all these cases, our general definition coincides with our earlier definitions of irreducible. Let's consider some examples of irreducibles in other commutative rings.

Example 10.8

Consider first the case of a field. Here, *all* non-zero elements are units, and so in a field there are no irreducible elements.

Example 10.9

What are irreducible elements in $\mathbb{Z}[x]$? Note that although $3x+6$ is irreducible in $\mathbb{Q}[x]$, it is not irreducible in $\mathbb{Z}[x]$. This is because 3 is a non-unit in $\mathbb{Z}[x]$, and so the factorization $3x+6 = 3(x+2)$ is non-trivial. Thus, the coefficients of an irreducible polynomial in $\mathbb{Z}[x]$ can have no common integer factor (other than ± 1). Such polynomials in $\mathbb{Z}[x]$ are called **primitive**. Gauss's Lemma 5.5 asserts that primitive polynomials (with degree greater than zero) of $\mathbb{Z}[x]$ are irreducible if and only if they are irreducible in $\mathbb{Q}[x]$. And what about polynomials of degree zero in $\mathbb{Z}[x]$ (that is, the integers)? Any factorization of such a polynomial would also be factorization in $\mathbb{Z}$. Thus, 3 (and any other prime integer, or its negative) will be an irreducible element of $\mathbb{Z}[x]$.

Example 10.10

What are the irreducible elements of $\mathbb{Z}_4$? In this ring, 1 and 3 are units, and so the only possible irreducible is 2. But the only factorizations of 2 in $\mathbb{Z}_4$ are $2 = 1 \cdot 2$ and $2 = 3 \cdot 2$, which are both trivial; thus, 2 is irreducible. Note in this case that 2 has no associates other than itself.

Although the examples just described are illustrative, we need further examples to really understand the idea of irreducible in general, and we will investigate such examples in the remainder of this chapter.

10.3 Quadratic Extensions of the Integers

We now describe a large class of rings of complex numbers, which are called quadratic extensions of $\mathbb{Z}$. Let n be an integer (positive or negative), other than 1. Suppose further that n has no non-trivial factors that are perfect squares. We call such an integer **square-free.** Thus, -21 is square-free, while 18 is not. We now define a subset of $\mathbb{C}$ as follows:

$$\mathbb{Z}[\sqrt{n}] = \{a + b\sqrt{n} : a, b \in \mathbb{Z}\},$$

which we call a **quadratic extension** of $\mathbb{Z}$. Our restriction to n being square-free is merely to ensure that no rational simplification is necessary under the radical sign. In particular, if n is square-free, we know that $\sqrt{n}$ is an irrational number (see Exercise 10.16). Notice that if n is positive, the quadratic extension consists of a set of real numbers.

We have met these sets before in the exercises. (See Exercise 6.12, and Exercises 7.1 and 7.2.) But in case you have not done these exercises, we now review the proof that these sets are subrings of $\mathbb{C}$. For a fixed n, we equip the set $\mathbb{Z}[\sqrt{n}]$ with the usual addition and multiplication inherited from $\mathbb{C}$. It is an easy matter to see that this set is non-empty, and closed under subtraction and multiplication:

$$(a + b\sqrt{n}) - (c + d\sqrt{n}) = (a - b) + (c - d)\sqrt{n},$$

and

$$(a + b\sqrt{n})(c + d\sqrt{n}) = (ac + nbd) + (ad + bc)\sqrt{n}.$$

This means by the Subring Theorem 7.1 that $\mathbb{Z}[\sqrt{n}]$ is a subring of $\mathbb{C}$, and so is a commutative ring (with unity), for each square-free integer n.

The case where $n = -1$ is of particular importance; this is called the ring of **Gaussian integers** and is usually written $\mathbb{Z}[i]$ (as we noted in Exercise 6.12).

What are the units and irreducibles of these rings? This general question actually turns out to be a fairly sophisticated inquiry from number theory, and we cannot give a general answer here. We will make some progress on this question, particularly in special cases.

10.4 Units in Quadratic Extensions

Let's see first what the question regarding units amounts to. If $a+b\sqrt{n}$ is a unit, then we have an equation of the form

$$(a + b\sqrt{n})(c + d\sqrt{n}) = (ac + nbd) + (ad + bc)\sqrt{n} = 1.$$

This means that we must find all simultaneous solutions a, b, c, d to the equations $ac+nbd = 1$ and $ad+bc = 0$, where a, b, c, d are *integers* (recall that here n is fixed). Equations where we require integer solutions are called **Diophantine equations.** Study of such equations makes up an important branch of number theory. In Exercise 10.1, you will try to solve this pair of equations directly in the particular case of the Gaussian integers (that is, where $n = -1$).

However, in the long run we will be better off by recalling our geometric representation for complex numbers introduced in Chapter 8. Let us view the Gaussian integers as points in the *complex plane.*

Recall that for a complex number $\rho = r+si$, we computed its modulus (or length) as $|\rho| = \sqrt{r^2 + s^2}$. Actually, for our purposes here it is rather less messy to consider the square of the length of a complex number $r+si$. We call that quantity $L(r+si)$. Thus, $L(r+si) = r^2+s^2$. Notice that we can view $L(\rho)$ as just $\rho\bar{\rho} = (r + si)(r - si)$, where $\bar{\rho} = r - si$ is the complex conjugate of ρ.

What is important for us is that the modulus function and, hence, the function L *preserve multiplication.* By this we mean that $L(\rho\tau) = L(\rho)L(\tau)$, for any pair of complex numbers ρ, τ. If you don't recall this fact, it would be well worth reviewing the proof of Theorem 8.3 now.

The fact that this function preserves multiplication means that we can translate certain algebraic questions about $\mathbb{Z}[i]$ into questions in the integers $\mathbb{Z}$, where they are presumably easier to answer.

For example, suppose that $\alpha \in \mathbb{Z}[i]$ is a unit. Then there exists another Gaussian integer β for which $\alpha\beta = 1$. But then we have that

$$L(\alpha)L(\beta) = L(\alpha\beta) = L(1) = 1.$$

Since $L(\alpha)$ and $L(\beta)$ are positive integers, this means that $L(\alpha) = L(\beta) = 1$. Hence, if a Gaussian integer α is a unit, then $L(a + bi) = a^2 + b^2 = 1$. It is then easy to check in $\mathbb{Z}$ that the only solutions to this equation are $a = \pm 1$ and $b = 0$, or vice versa. That is, the units of

$\mathbb{Z}[i]$ consist precisely of $\{1, -1, i, -i\}$, the four points on the unit circle in the complex plane at angles $0, \pi/2, \pi, 3\pi/2$. Notice that we have arrived at this conclusion much more easily than by directly solving the appropriate Diophantine equations, as in Exercise 10.1.

What we would really like to do is to equip $\mathbb{Z}[\sqrt{n}]$ with a function playing the same role as L does for the Gaussian integers, where n is any square-free integer. The function we propose to use is the following. Let $N : \mathbb{Z}[\sqrt{n}] \mapsto \mathbb{N} \cup \{0\}$ be defined by setting

$$N(a + b\sqrt{n}) = |(a + b\sqrt{n})(a - b\sqrt{n})| = |a^2 - nb^2|.$$

We call $N(\alpha)$ the **norm** of α.

Notice that if n is negative (as in the case of the Gaussian integers), this is merely the square of the ordinary complex modulus, and is automatically non-negative; consequently, the absolute value sign in the definition is superfluous. However, if n is positive, we need the absolute value, or else we might not get non-negative integer values. And note that as long as n is square-free, $N(\alpha) = 0$ only if $\alpha = 0$ (otherwise, $\sqrt{n}$ would be rational, which contradicts Exercise 10.16).

In order to use the function N as we did L for the Gaussian integers, we need to show that it preserves multiplication:

Theorem 10.1 *Let n be a square-free integer, and let $\alpha, \beta \in \mathbb{Z}[\sqrt{n}]$. Then $N(\alpha\beta) = N(\alpha)N(\beta)$.*

Proof: This proof is similar in flavor to that given for Theorem 8.3 and is left to the reader; see Exercise 10.2. □

We now show that the norm is useful for identifying units, in any of the rings $\mathbb{Z}[\sqrt{n}]$:

Theorem 10.2 *Let n be a square-free integer, and let $\alpha \in \mathbb{Z}[\sqrt{n}]$. Then α is a unit if and only if $N(\alpha) = 1$.*

Proof: Suppose that n is a square-free integer, and $\alpha \in \mathbb{Z}[\sqrt{n}]$ is a unit. Then there exists β with $\alpha\beta = 1$. But then application of the function N gives us $N(\alpha)N(\beta) = 1$. Because all values of the norm function are non-negative integers, we must have that $N(\alpha) = 1$, as required.

Conversely, suppose that $N(\alpha) = 1$. But if $\alpha = a + b\sqrt{n}$, this means that $(a + b\sqrt{n})(a - b\sqrt{n}) = 1$ or -1, which means that $a - b\sqrt{n}$ (or $-a + b\sqrt{n}$) is the required multiplicative inverse. □

▷ **Quick Exercise.** Notice that $5^2 - 6(2)^2 = 1$. What units does this provide for the ring $\mathbb{Z}[\sqrt{6}]$? ◁

Let us for the moment restrict ourselves to the question of explicitly determining the units in quadratic extensions of the form $\mathbb{Z}[\sqrt{n}]$, where n is negative. If in this case $\alpha = a + b\sqrt{n}$ is a unit, then $1 = N(\alpha) = a^2 - nb^2$. If $n = -1$, this is the case of the Gaussian integers, and we get $\{1, -1, i, -i\}$ as the units, exactly as before. But if $n < -1$, then there are no solutions to this equation if $b > 0$. And so the only units in this case are ±1.

▷ **Quick Exercise.** Check for yourself the solutions to the equation $1 = a^2 - nb^2$, when $n < -1$. ◁

Once we know the units for a ring, we know which elements are associates of which. For example, what are the associates of $3 + i$ in $\mathbb{Z}[i]$? They are $(1)(3+i) = 3+i, (-1)(3+i) = -3-i, (i)(3+i) = -1+3i$, and $(-i)(3 + i) = 1 - 3i$.

▷ **Quick Exercise.** What are the associates of $a + bi$ in $\mathbb{Z}[i]$? ◁

Example 10.11

What are the associates of a given element in $\mathbb{Z}[\sqrt{-5}]$? The only units in this ring are $\pm 1 + 0\sqrt{-5} = \pm 1$. Thus, the only associates of an element in $\mathbb{Z}[\sqrt{-5}]$ are the element itself and its negative.

Let's now turn to the case $\mathbb{Z}[\sqrt{n}]$, for $n > 0$. If we wish to find the units of $\mathbb{Z}[\sqrt{n}]$, when n is a positive square-free integer, we must solve the Diophantine equation $|a^2 - nb^2| = 1$. In the Quick Exercise after Theorem 10.2 we actually gave a solution to this equation for $n = 6$, which then gave us units in $\mathbb{Z}[\sqrt{6}]$.

Unfortunately however, this is in general a rather more subtle question than the corresponding question when $n < 0$. We are looking for integers a, b satisfying either $a^2 = nb^2 + 1$, or else $nb^2 = a^2 + 1$, where n is a constant positive square-free integer. Such Diophantine equations have a long history and are known as *Pell's equations.* The description of how such equations are solved in general is a fascinating piece of mathematics that is unfortunately beyond the scope of our text.

Example 10.12

For $n = 2$, observe that $a = 1$ and $b = 1$ does in fact give a solution, because $|1^2 - 2(1)^2| = 1$. Thus, $1 + 1\sqrt{2}$ is a unit for $\mathbb{Z}[\sqrt{2}]$.

We have already observed that in any ring the set of units is closed under multiplication, and so this means that $(1 + \sqrt{2})^2$ must also be a unit in $\mathbb{Z}[\sqrt{2}]$, and indeed so is $(1 + \sqrt{2})^n$, for any positive integer n. Furthermore, each of these units is distinct: since $1 + \sqrt{2} > 1$, these numbers form a list of strictly increasing real numbers.

Notice that we could also have concluded that the positive powers of α are units by observing that if $N(\alpha) = 1$, then $N(\alpha^n) = (N(\alpha))^n = 1^n = 1$. But this even makes sense for negative exponents. After all, because $1 + \sqrt{2}$ is a unit, $(1 + \sqrt{2})^{-1} = -1 + \sqrt{2}$ is a unit of $\mathbb{Z}[\sqrt{2}]$, and so all elements of the form $(1 + \sqrt{2})^{-n} = ((1 + \sqrt{2})^{-1})^n = (-1 + \sqrt{2})^n$ are units too.

▷ **Quick Exercise.** Why is $(1 + \sqrt{2})^{-1} = -1 + \sqrt{2}$? ◁

But -1 is a unit too (it is its own multiplicative inverse), and so we obtain the following lists of units of $\mathbb{Z}[\sqrt{2}]$:

$$1, \quad 1+\sqrt{2}, \quad (1+\sqrt{2})^2 = 3+2\sqrt{2}, \quad (1+\sqrt{2})^3 = 7+5\sqrt{2}, \cdots,$$
$$-1, \quad -1-\sqrt{2}, \quad -3-2\sqrt{2}, \quad -7-5\sqrt{2}, \cdots,$$
$$(1+\sqrt{2})^{-1} = -1+\sqrt{2}, \quad (1+\sqrt{2})^{-2} = -3+2\sqrt{2},$$
$$(1+\sqrt{2})^{-3} = -7+5\sqrt{2}, \cdots,$$

and

$$1-\sqrt{2}, \quad 3-2\sqrt{2}, \quad 7-5\sqrt{2}, \cdots.$$

It turns out (though we will not prove it here) that this is a complete list.

This infinite list of units means that detecting whether or not two elements in $\mathbb{Z}[\sqrt{2}]$ are associates is not the trivial matter it is in $\mathbb{Z}$ (or even in $\mathbb{Z}[i]$). For example, $4 + \sqrt{2}$ and $8 - 5\sqrt{2}$ are associates, because $(4 + \sqrt{2})(3 - 2\sqrt{2}) = 8 - 5\sqrt{2}$, and $3 - 2\sqrt{2}$ is a unit.

▷ **Quick Exercise.** Calculate several distinct associates of $3 + 5\sqrt{2}$ in $\mathbb{Z}[\sqrt{2}]$. ◁

▷ **Quick Exercise.** Find several units of $\mathbb{Z}[\sqrt{3}]$ (other than ±1), and then compute several associates of $\sqrt{3}$. ◁

Notice that in any of these quadratic extension rings, it is obvious that if two elements are associates, then they have the same norm.

▷ **Quick Exercise.** Why do two elements that are associates have the same norm? ◁

It is tempting to conjecture that the converse of this statement is true, but you will demonstrate by explicit example in Exercise 10.14 two elements of such a ring, with the same norm, which are not associates.

10.5 Irreducibles in Quadratic Extensions

What about irreducibles for $\mathbb{Z}[\sqrt{n}]$ (for any square-free n)? First observe that if $N(\alpha)$ is an irreducible (integer), then certainly α itself is an irreducible. For if $\alpha = \beta\gamma$, then the norm of exactly one of β and γ must be 1 (because $N(\alpha)$ is a (positive) irreducible in $\mathbb{Z}$) and units are exactly those elements of $\mathbb{Z}[\sqrt{n}]$ which have norm one. We state this result for emphasis:

Theorem 10.3 *Let n be a square-free integer, and $\alpha \in \mathbb{Z}[\sqrt{n}]$. If $N(\alpha)$ is a prime integer, then α is irreducible in $\mathbb{Z}[\sqrt{n}]$.*

Example 10.13

Thus, $2+5\sqrt{-5}$ is irreducible in $\mathbb{Z}[\sqrt{-5}]$, because $N(2+5\sqrt{-5}) = 129$, which is prime (in $\mathbb{Z}$). And $1 + 2\sqrt{2}$ is irreducible in $\mathbb{Z}[\sqrt{2}]$, because its norm is 7.

Example 10.14

Notice that $1 + i$ is irreducible in the Gaussian integers. But because $2 = (1 + i)(1 - i)$, 2 (an irreducible in $\mathbb{Z}$) is *not* an irreducible in $\mathbb{Z}[i]$.

▷ **Quick Exercise.** Use Theorem 10.3 to find some irreducible elements in $\mathbb{Z}[\sqrt{6}]$ and in $\mathbb{Z}[\sqrt{-3}]$. ◁

Unfortunately, the converse of Theorem 10.3 is false. That is, there exist irreducibles that do not have prime norm.

Example 10.15

For example, we claim that $1+\sqrt{-5}$ (which has norm 6) is irreducible in $\mathbb{Z}[\sqrt{-5}]$. For if it had a non-trivial factorization, the factors would have to have norms of 2 and 3. But this would require integer solutions of the Diophantine equations $a^2+5b^2=2$ and $a^2+5b^2=3$. Obviously, no such solutions are possible.

Example 10.16

Similarly, we claim that 3 is irreducible in $\mathbb{Z}[\sqrt{2}]$, even though the norm of 3 is 9 (which is not prime in $\mathbb{Z}$). In this case we would need solutions of at least one of the Diophantine equations $a^2 = 2b^2+3$ or $2b^2 = a^2+3$. Suppose by way of contradiction that the first of these equations did in fact have a solution. Then a would have to be an odd integer, and so $a = 2k+1$. But then $(2k+1)^2 = 2b^2+3$, or after a little algebra, $2k^2+2k = b^2+1$. This means that b must be odd, and so $b = 2m+1$. Then $2k^2+2k = (2m+1)^2+1$, or after simplification, $k^2+k = 2m^2+2m+1$. But k^2+k, as the product of consecutive integers, is necessarily even, while $2m^2+2m+1$ is necessarily odd. This contradiction shows that no integer solution to the equation $a^2 = 2b^2+3$ is possible. We leave it to you to check (see Exercise 10.4) that there is no solution to the other equation either. This means there are no members of $\mathbb{Z}[\sqrt{2}]$ which have norm 3, and so 3 (with norm 9) has no non-trivial factorizations.

The fact that it required this excursion into number theory to prove that 3 is irreducible in $\mathbb{Z}[\sqrt{2}]$ might convince you that the general question of determining all irreducibles for quadratic extensions of $\mathbb{Z}$ is difficult. This is in fact the case, and we will not pursue the matter here.

Chapter Summary

In this chapter we introduced the concepts of *associate* and *irreducible*, for any commutative ring. We examined examples of these concepts in many previously encountered rings.

We then introduced the *quadratic extensions of the integers* and looked at units and irreducibles in such rings, making heavy use of the *norm function.*

Warm-up Exercises

a. Determine whether the following pairs of elements are associates:

(a) x and $3x$ in $\mathbb{Q}[x]$.

(b) x and $3x$ in $\mathbb{Z}[x]$.

(c) 4 and 2 in $\mathbb{Z}_{10}$.

(d) 4 and 2 in $\mathbb{Z}_8$.

(e) 4 and 2 in $\mathbb{Z}_7$.

(f) $4+\sqrt{3}$ and $11+6\sqrt{3}$ in $\mathbb{Z}[\sqrt{3}]$.

(g) $(2,-1)$ and $(4,1)$ in $\mathbb{Z}\times\mathbb{Z}$.

(h) $(2,-1)$ and $(4,1)$ in $\mathbb{Q}\times\mathbb{Z}$.

b. Discover which elements of $\mathbb{Z}_{15}$ are associates of which.

c. List all associates of x^2+3x+4 in $\mathbb{Z}_5[x]$.

d. What are the units in $\mathbb{Q}\times\mathbb{Q}$? What are the associates in this ring of $(1,2)$? of $(1,0)$?

e. Determine four distinct associates of $3+2\sqrt{2}$ in $\mathbb{Z}[\sqrt{2}]$.

f. Determine all associates of $5+i$ in $\mathbb{Z}[i]$.

g. Let n be any square-free integer. Why is n not irreducible in $\mathbb{Z}[\sqrt{n}]$?

h. Determine four distinct irreducibles in $\mathbb{Z}[\sqrt{2}]$. *Hint:* Look for elements with prime norm.

i. Do irreducible elements of $\mathbb{Z}[\sqrt{n}]$ necessarily have prime norm?

j. Determine which of the following elements are irreducible:

(a) $9+\sqrt{10}$ in $\mathbb{Z}[\sqrt{10}]$.

(b) $5+\sqrt{5}$ in $\mathbb{Z}[\sqrt{5}]$.

(c) $2x^2+4$ in $\mathbb{Z}[x]$.

(d) $2x^2+4$ in $\mathbb{Q}[x]$.

(e) x^2+4 in $\mathbb{Z}_7[x]$.

(f) $(1,3)$ in $\mathbb{Z} \times \mathbb{Z}$.

(g) $(0,3)$ in $\mathbb{Z} \times \mathbb{Z}$.

(h) $(1,-1)$ in $\mathbb{Z} \times \mathbb{Z}$.

(i) $(1,3)$ in $\mathbb{Z} \times \mathbb{Q}$.

Exercises

1. Find all simultaneous integer solutions to the Diophantine equations $ac - bd = 1, ad + bc = 0$ directly, by eliminating variables; interpret your solutions as determining all units in the Gaussian integers.

2. Prove Theorem 10.1. That is, let n be a square-free integer. As in the text, define $N(a + b\sqrt{n}) = |a^2 - nb^2|$. Prove that N preserves multiplication, that is, $N(\alpha\beta) = N(\alpha)N(\beta)$.

3. Suppose that n, m are distinct square-free integers. Prove that
$$\mathbb{Z}[\sqrt{n}] \cap \mathbb{Z}[\sqrt{m}] = \mathbb{Z}.$$
This is not true if at least one of the integers n and m is not square-free. Give an example to show this.

4. Prove that the Diophantine equation $2b^2 = a^2 + 3$ has no integer solutions, proceeding similarly as the problem $a^2 = 2b^2 + 3$ is handled in the text in Example 10.16.

5. Find infinitely many distinct units in $\mathbb{Z}[\sqrt{7}]$. Then list infinitely many associates of $\sqrt{7}$ in $\mathbb{Z}[\sqrt{7}]$.

6. Suppose that n is a square-free integer and $n > 0$. Prove that $\mathbb{Z}[\sqrt{-n}]$ has only finitely many units.

7. Show that there are no irreducible elements in $\mathbb{Z}_6$.

8. Show that 2 is irreducible in $\mathbb{Z}_8$, and that every non-unit in $\mathbb{Z}_8$ is irreducible or a product of irreducibles.

9. Determine all irreducible elements of $\mathbb{Z} \times \mathbb{Z}$.

10. Prove that if p is a prime in $\mathbb{Z}$ and p is congruent to 3 mod 4, then p is irreducible in $\mathbb{Z}[i]$.

11. Prove 2 is irreducible in $\mathbb{Z}[\sqrt{n}]$ for all square-free $n < -2$.

12. Find two distinct square-free integers n (with $n > 1$) for which 2 is not irreducible in $\mathbb{Z}[\sqrt{n}]$.

13. Suppose that p is a positive prime integer. Prove that $\sqrt{p}$ is irreducible in $\mathbb{Z}[\sqrt{p}]$. Show by example that this is false if p is not prime; in particular, consider $p = 6$.

14. Show that $11 + 6\sqrt{-5}$ and $16 + 3\sqrt{-5}$ are irreducible elements in $\mathbb{Z}[\sqrt{-5}]$, with the same norm. Show that these elements are not associates.

15. Suppose that R is a commutative ring with unity, and consider the ring $R[x]$ of polynomials with coefficients from R. (See Exercise 6.23.) Let r be an irreducible element of R. Prove that r is an irreducible element of $R[x]$.

16. Suppose that n is a square-free integer. Prove that $\sqrt{n}$ is irrational.

Chapter 11

Factorization and Ideals

In the last chapter we acquainted ourselves with the constituent pieces of a general theory of factorization for domains: namely, the irreducibles. In this chapter and the next we discuss what is required to obtain a factorization into irreducibles, and along the way we meet some vitally important concepts for all of ring theory. Conditions necessary to force such a factorization to be unique will be examined in Chapter 13.

You should now recall the proof that every integer can be factored into irreducibles (and the analogous proof for $\mathbb{Q}[x]$). Both of these proofs depend heavily on the fact that $\mathbb{N}$ is well ordered: by continuing to extract factors from a positive integer, we decrease its size, and we cannot continue this indefinitely. What more general context is possible?

11.1 Factorization for Quadratic Extensions

For the quadratic extensions of $\mathbb{Z}$ discussed in the last chapter we have the appropriate tool at hand: The norm function N provides a measure of size. In fact, it shouldn't be too surprising that this might work, because the norm function takes values precisely in the set of non-negative integers. Recall that for any square-free integer n,

$$N(a + b\sqrt{n}) = |a^2 - nb^2|,$$

and N preserves multiplication (Theorem 10.1). We have also shown that the units of $\mathbb{Z}[\sqrt{n}]$ are precisely those numbers with norm 1 (Theorem 10.2).

We can now prove the following:

Theorem 11.1 Factorization Theorem for Quadratic Extensions of $\mathbb{Z}$ *Let n be a square-free integer. Then every non-zero* *non-unit of* $\mathbb{Z}[\sqrt{n}]$ *is either irreducible or a product of irreducibles.*

Proof: Let $\alpha \neq 0$ be a non-unit of $\mathbb{Z}[\sqrt{n}]$. We proceed by induction on $N(\alpha)$. Note that $N(\alpha) \neq 1$, because α is a non-unit, and if $N(\alpha) = 2$, then α is itself irreducible.

▷ **Quick Exercise.** Show that if $\alpha \in \mathbb{Z}[\sqrt{n}]$ (n square-free) and $N(\alpha) = 2$, then α is irreducible. ◁

Now suppose the theorem holds true for all β with $N(\beta) < m$. If α is irreducible already, we are done. If not, then $\alpha = \beta\gamma$, where both factors are non-units. But because $N(\alpha) = N(\beta)N(\gamma)$ and $N(\beta) > 1$ and $N(\gamma) > 1$, we have that $N(\beta) < m$ and $N(\gamma) < m$. By the induction hypothesis both β and γ can be factored as a product of irreducibles and, thus, so can their product α. □

▷ **Quick Exercise.** Find irreducible elements in $\mathbb{Z}[i]$ and $\mathbb{Z}[\sqrt{-7}]$ that have norm 2. ◁

It is of great importance to note that we have neither claimed nor proved that the factorization into irreducibles provided by this theorem is unique. There is good reason for this: For suitable choice of n, such factorizations are *not* unique.

To see this, consider the following two factorizations of 6 in $\mathbb{Z}[\sqrt{-5}]$:

$$6 = (1 + \sqrt{-5}) \cdot (1 - \sqrt{-5}) = 2 \cdot 3.$$

We argued in Example 10.15 that $1+\sqrt{-5}$ is irreducible, and a similar argument applies for the other factors in the two given factorizations.

▷ **Quick Exercise.** Verify that $1 - \sqrt{-5}$, 2, and 3 are irreducible in $\mathbb{Z}[\sqrt{-5}]$, by considering their norms. ◁

Now, if a unique factorization theorem applied for $\mathbb{Z}[\sqrt{-5}]$, this would mean that these two factorizations would be the same, up to order and unit factors. But it is quite easy to see that 2 is not an associate of either $1+\sqrt{-5}$ or $1-\sqrt{-5}$, because associates must have the same norm. In Exercise 11.7, you will provide two essentially distinct factorizations of 8 in $\mathbb{Z}[\sqrt{-7}]$.

But before we inquire into this uniqueness question, we must first confess that we have really made little progress toward a general theory of factorization into irreducibles, because the quadratic extensions of $\mathbb{Z}$ remain a fairly special class of domains to discuss.

11.2 How Might Factorization Fail?

To make a more general attack on the factorization problem, let us think about how factorization into irreducibles in a domain R could fail. Suppose then that $0 \neq a_1 \in R$ is not irreducible and is not the product of irreducibles. Then there exists a factorization $a_1 = a_2b_2$, where neither a_2 nor b_2 is a unit. Furthermore, because a_1 cannot be factored into a product of irreducibles, this must be true of at least one of a_2 or b_2. To be specific, let's suppose that a_2 can't be so factored. But then a_2 can be factored as $a_2 = a_3b_3$, where neither a_3 nor b_3 is a unit, and where a_3 cannot be factored into a product of irreducibles. If we continue in this fashion, we obtain an infinite sequence of non-trivial factorizations:

$$a_1 = a_2b_2, \quad a_2 = a_3b_3, \quad a_3 = a_4b_4, \quad \cdots$$

where every element in sight is a non-unit.

Here is a slightly different way to think of this infinite sequence. Let

$$\langle a\rangle = \{b \in R : b = ax, \text{for some } x \in R\}.$$

That is, $\langle a\rangle$ consists of the multiples of a in R. Our sequence of factorizations clearly in part asserts that

$$\langle a_1\rangle \subseteq \langle a_2\rangle \subseteq \langle a_3\rangle \subseteq \cdots.$$

This string of inclusions really says something very simple: All multiples of a_1 are in turn multiples of a_2 (because a_1 is itself a multiple of a_2), all multiples of a_2 are in turn multiples of a_3, and so on.

But we claim further that these containments are proper. For if $\langle a\rangle = \langle b\rangle$, then a is a multiple of b, and b is a multiple of a. That is, $ax = b$ and $by = a$, for some $x, y \in R$. Therefore, $byx = b$, and so $yx = 1$ (because the cancellation law holds in domains). But this means that y and x are units, and so a and b are associates. We are assuming in our infinite sequence that none of the a_i's are associates, and so our infinite sequence of proper factorization leads to an infinite *ascending* sequence of sets of the form $\langle a\rangle$. If we could always factor into irreducibles, such a situation would be impossible.

Example 11.1

Let's consider an example of these ideas and this notation in the ring $\mathbb{Z}$. Consider the integer 360. By successively factoring 360, we might obtain the following chain of inclusions:

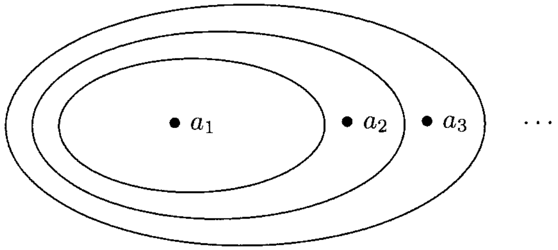
$$\langle 360\rangle \subset \langle 180\rangle \subset \langle 60\rangle \subset \langle 20\rangle \subset \langle 10\rangle \subset \langle 5\rangle.$$

We cannot continue any further in this example, precisely because 5 is irreducible and so admits no further non-trivial factorization. Of course, we know already that $\mathbb{Z}$ has a factorization theorem, and so we should have expected this chain of inclusions to halt.

The argument above about the potential lack of factorization will be important in the next chapter, and so we record its conclusion as a lemma:

Lemma 11.2 *Let R be a domain and $0 \neq a_1$ an element of R that is neither irreducible nor the product of irreducibles. Then there exist non-units $a_2, a_3, a_4, \cdots$, such that*

$$\langle a_1\rangle \subset \langle a_2\rangle \subset \langle a_3\rangle \cdots.$$



11.3 Ideals

What we need now is a further abstraction that will enable us to talk fruitfully about such subsets of a ring as $\langle a\rangle$. The required definition is the following: Let R be a commutative ring. An **ideal** of R is a non-empty subset I of R satisfying the following criteria:

1. I is closed under subtraction;

2. if $a \in I$ and $r \in R$, then $ar \in I$.

Thus, an ideal is a subring that satisfies the stronger multiplicative closure property (2). This property says that I 'absorbs' multiplication by *any* ring element. We remind ourselves of this by calling (2) the **multiplicative absorption property** of ideals.

We have made this definition only for commutative rings, because we deal primarily with such rings in this book. However, the definition above can be modified to account for the notion of ideal in the non-commutative case: Simply strengthen (2) to require that both ar and ra belong to I if $a \in I$.

We now consider examples of ideals:

Example 11.2

Every commutative ring (except the zero ring) has at least two ideals. The first is the **trivial** or **zero** ideal $\{0\}$. We've already noted that this is a subring, and the multiplicative absorption property holds, because $0a = 0$ for all a. The second is the **improper** ideal consisting of the entire ring. It is obviously closed under subtraction and multiplicative absorption. Note that we call any ideal which does not consist of the entire ring a **proper** ideal.

Example 11.3

Let's discuss some ideals of $\mathbb{Z}$. Consider the following subsets:

$$\langle 0\rangle = \{0\}, \quad \langle 1\rangle = \mathbb{Z}, \quad \langle 2\rangle = \{x \in \mathbb{Z} : x \text{ is even}\}, \quad \langle 3\rangle, \quad \langle 4\rangle, \cdots.$$

Each of these is a subring, and in fact, each satisfies the multiplicative absorption property. For example, to see that $\langle 2\rangle$ does, we need only observe that $\langle 2\rangle$ is the set of even integers: If we multiply an even integer by *any* integer, we obtain another even integer.

▷ **Quick Exercise.** Check that the set $\langle 3\rangle$ in $\mathbb{Z}$ satisfies the multiplicative absorption property. ◁

11.4 Principal Ideals

We can clearly generalize the observation in Example 11.3: Given a commutative ring R with $a \in R$, we claim that $\langle a \rangle$ is an ideal. We first show that it is closed under subtraction. For if $x, y \in \langle a \rangle$, then $x = ar$ and $y = as$, for some $r, s \in R$. But then $x - y = ar - as = a(r - s)$, which is clearly a multiple of a, and so an element of $\langle a \rangle$. To show that $\langle a \rangle$ is closed under multiplicative absorption, suppose that $x \in \langle a \rangle$ and $y \in R$. Then $x = ar$, and so $xy = ary = a(ry)$. This latter element is a multiple of a, and so belongs to $\langle a \rangle$. We call ideals of the form $\langle a \rangle$ **principal ideals**, and call a a **generator** for $\langle a \rangle$.

Example 11.4

Consider the principal ideal $\langle 3 \rangle$ in $\mathbb{Z}_{12}$. Clearly, this ideal consists of the elements $\{0, 3, 6, 9\}$.

Example 11.5

Consider the principal ideal $\langle \sqrt{5} \rangle$ in $\mathbb{Z}[\sqrt{5}]$. What does a typical element of this ideal look like? All such elements are multiples of $\sqrt{5}$, and this naturally leads one to think of such elements as $\sqrt{5}, -\sqrt{5}, 2\sqrt{5}$, and so forth. However, these are merely integer multiples of $\sqrt{5}$, and we must allow multiples of $\sqrt{5}$ by any element from the ring in question: namely, $\mathbb{Z}[\sqrt{5}]$. Thus, a typical element of this principal ideal is an element of the form $(a + b\sqrt{5})(\sqrt{5}) = 5b + a\sqrt{5}$. This means that

$$\langle \sqrt{5} \rangle = \{5r + s\sqrt{5} : r, s \in \mathbb{Z}\}.$$

That is, $\langle \sqrt{5} \rangle$ consists of all elements of $\mathbb{Z}[\sqrt{5}]$ whose "rational part" is divisible by 5.

Example 11.6

Consider the principal ideal $\langle x^2 \rangle$ in $\mathbb{Q}[x]$. This consists of all multiples of x^2; that is, those polynomials with x^2 as a factor. Equivalently, $\langle x^2 \rangle$ is the set of all polynomials whose constant and degree 1 coefficients are zero. More generally, $\langle f \rangle$ in $\mathbb{Q}[x]$ consists of all polynomials with f as a factor.

Example 11.7

Consider the principal ideal $\langle 2 \rangle$ in the ring $2\mathbb{Z}$ of even integers. In this example, when we form the multiples of 2, we are restricted only to even numbers. Thus,

$$\langle 2 \rangle = \{\cdots, 0, 4, 8, 12, \cdots\}.$$

Notice that in this case $2 \notin \langle 2 \rangle$. This situation arises because $2\mathbb{Z}$ lacks unity.

In case a commutative ring has unity, it is certainly always the case that $a \in \langle a \rangle$. And in fact, this gives us a more abstract but still useful way of describing the principal ideal $\langle a \rangle$ of an element a in a commutative ring R with unity: It is the smallest ideal of R containing a. This merely means that $\langle a \rangle$ is a subset of *any* ideal of R containing a. But this is obvious because if $a \in I$, where I is an ideal, then by the multiplicative absorption property for I, all multiples of a are elements of I.

Example 11.8

Consider now the ring $\mathbb{Q}$. We know that $\mathbb{Z}$ is a subring of $\mathbb{Q}$. Is it an ideal of $\mathbb{Q}$? Suppose it is; we now apply the multiplicative absorption property to the element 1 of $\mathbb{Z}$. Choose any rational number q. Then by multiplicative absorption, $(q)(1) = q$ would be an element of $\mathbb{Z}$, which is certainly not always true. Thus, $\mathbb{Z}$ is a subring of $\mathbb{Q}$ which is *not* an ideal.

▷ **Quick Exercise.** Is $\mathbb{Q}$ an ideal of the ring $\mathbb{R}$? ◁

The previous example makes clear that if R is a commutative ring with unity, then any ideal containing 1 must be the improper ideal R. In particular, this means that $\langle 1 \rangle = R$.

What other elements besides 1 generate the improper ideal? To answer this, suppose that $\langle r \rangle = R$, where R is a commutative ring with unity, and $r \in R$. This means that 1 is a multiple of r; but if $rs = 1$, then r is a unit. And conversely, if r is a unit, then $1 = rt$, for some $t \in R$. But then $x = r(tx)$, for any $x \in R$. Thus, all elements of R are multiples of r. We have thus shown that in a commutative ring with unity, $\langle r \rangle = R$ if and only if r is a unit. We record this fact for future reference in the theorem below:

Theorem 11.3 *Let R be a commutative ring with unity, with $r \in R$. Then $\langle r \rangle = R$ if and only if r is a unit.*

▷ **Quick Exercise.** Consider the unit 5 in $\mathbb{Z}_{12}$. Verify explicitly that $\langle 5 \rangle = \mathbb{Z}_{12}$. (That is, show by computation that every element of the ring is a multiple of 5.) ◁

Because all non-zero elements in a field are units, a field has only the two ideals that all commutative rings possess: namely, the zero ideal and the improper ideal. The converse is also true:

Corollary 11.4 *A commutative ring with unity is a field if and only if its only ideals are the trivial and improper ideals.*

Proof: We have already proved half of this corollary. For the converse, suppose that R is a commutative ring with unity whose only ideals are the trivial and improper ideals. Choose any non-zero element $r \in R$. Because r is non-zero, then $\langle r \rangle \neq \{0\}$, and so $\langle r \rangle = R$. But then $1 \in \langle r \rangle$, and so $1 = rs$, for some element s. But then r is a unit. Thus, all non-zero elements of R are units, and so R is a field. □

Let's return to our discussion of ideals in $\mathbb{Q}[x]$, in Example 11.6.

Example 11.9

Consider the ideal $\langle x \rangle$ in $\mathbb{Q}[x]$. Clearly, this consists of those polynomials with no constant term. But what about the ideal $\langle 3x \rangle$? We claim that $\langle x \rangle = \langle 3x \rangle$. It is clear that $\langle 3x \rangle \subseteq \langle x \rangle$, but the reverse inclusion holds too, because $x = \frac{1}{3}(3x)$, and so $x \in \langle 3x \rangle$.

This example raises a more general question: When do two elements of a commutative ring generate the same principal ideal? The previous example suggests the answer, at least in the case of domains. (The answer given in the next theorem for domains does not hold for all commutative rings.)

Theorem 11.5 *Let R be a domain, and $r, s \in R$. Then $\langle r \rangle = \langle s \rangle$ if and only if r and s are associates.*

Proof: Suppose first that r and s are associates. Then $r = su$, where u is a unit. Thus, r is a multiple of s, and so $r \in \langle s \rangle$. Hence, $\langle r \rangle \subseteq \langle s \rangle$. But $s = ru^{-1}$, and so similarly $\langle s \rangle \subseteq \langle r \rangle$.

For the converse, suppose that $\langle r \rangle = \langle s \rangle$. Then r is a multiple of s and vice versa; that is, $r = sx$ and $s = ry$, for some x and y. But then $r = ryx$, and so by the multiplicative cancellation property for domains, $1 = yx$. That is, y and x are units, and so r and s are associates. (Note that we used exactly this argument in our discussion of factorization above.) □

In Example 11.3 we listed as ideals of $\mathbb{Z}$ the principal ideals $\langle n \rangle$, for all non-negative integers n. But are there any ideals other than these? The next theorem asserts that there are not.

Theorem 11.6 *All ideals of $\mathbb{Z}$ are principal.*

Proof: Suppose that I is an ideal of $\mathbb{Z}$. We wish to show that I is principal. If I is the zero ideal this is already obvious; so suppose I has more elements than just 0. What element of I might serve as a generator for I? We answer this question by using the Well-ordering Principle: Choose the smallest positive element m of I.

▷ **Quick Exercise.** Why need I have *any* positive elements? ◁

We claim that $\langle m \rangle = I$. Because I is an ideal (and so satisfies the multiplicative absorption property), it is clear that $\langle m \rangle \subseteq I$. Suppose now that $b \in I$. We claim $b \in \langle m \rangle$; that is, we claim that b is a multiple of m. To check this, we will use the Division Theorem 2.1 to obtain a quotient q and a remainder r where $b = qm + r$. Obviously, we hope that the remainder is zero. But what do we know about r? Because $r = b - qm$, $r \in I$, using both of the defining properties of ideal. But we also know that $0 \leq r < m$, and because m is the *smallest* positive element of I, $r = 0$, as required. Thus for $\mathbb{Z}$, all ideals are principal. □

We are leaving it to you in Exercise 11.2 below to show that the same fact holds true for $\mathbb{Q}[x]$. By this time you should not be surprised by this analogy between $\mathbb{Z}$ and $\mathbb{Q}[x]$. The proof you will construct is similar, except you will choose m to be a non-zero polynomial with smallest *degree* in the ideal.

All the examples of ideals we've encountered so far have been principal ideals. Are there any *non-principal* ideals? Such ideals do exist (but not in $\mathbb{Z}$ or $\mathbb{Q}[x]$), and we will meet some in the next chapter.

Chapter Summary

In this chapter we proved the *Factorization Theorem for Quadratic Extensions* of $\mathbb{Z}$ and showed by example that such factorizations need not be unique.

We then discussed how factorization in a domain could fail. This led us to the notion of *ideal* and *principal ideal*. We showed that all ideals in $\mathbb{Z}$ are principal.

Warm-up Exercises

a. Do the two factorizations

$$5 = \sqrt{-5} \cdot -\sqrt{-5} = 1 \cdot 5$$

provide another example that factorization into irreducibles is not unique in $\mathbb{Z}[\sqrt{-5}]$? Why or why not?

b. Do the two factorizations

$$6 = 3 \cdot 2 = (-2 + 2\sqrt{2})(3 + 3\sqrt{2})$$

provide an example to show that factorization into irreducibles is not unique in $\mathbb{Z}[\sqrt{2}]$? Why or why not?

c. Are the following ideals?

(a) $\mathbb{Q}$ in $\mathbb{R}$.

(b) $\{0, 2, 4, 6\}$ in $\mathbb{Z}_8$.

(c) $\mathbb{Z}[x]$ in $\mathbb{Q}[x]$.

(d) $\mathbb{Z}$ in $\mathbb{Z}[i]$.

(e) $\{ni : n \in \mathbb{Z}\}$ in $\mathbb{Z}[i]$.

(f) $\mathbb{Z} \times \{0\}$ in $\mathbb{Z} \times \mathbb{Z}$.

(g) $\{5n + 5m\sqrt{2} : n, m \in \mathbb{Z}\}$ in $\mathbb{Z}[\sqrt{2}]$.

(h) The set of all polynomials with even degree (together with the zero polynomial), in $\mathbb{Q}[x]$.

d. Are the following equalities true or false?

(a) $\langle x \rangle = \langle 3x \rangle$, in $\mathbb{Z}[x]$.

(b) $\langle x \rangle = \langle 3x \rangle$, in $\mathbb{Q}[x]$.

(c) $\langle 1 + i \rangle = \mathbb{Z}[i]$.

(d) $\langle 2 \rangle = \langle 10 \rangle$, in $\mathbb{Z}_{14}$.

e. Suppose that R is a commutative ring with unity, and I is an ideal of R. If $1 \in I$, what can you say about I?

f. Give a nice description of the elements of the ideal $\langle \sqrt{7} \rangle$ in the ring $\mathbb{Z}[\sqrt{7}]$.

g. Have we given an example in this chapter of an ideal that is *not* principal?

h. Give examples of the following (or explain why they don't exist):

(a) A non-zero proper ideal of a finite ring.

(b) A non-zero proper ideal of $\mathbb{C}$.

(c) A non-zero proper ideal of $\mathbb{Z}[i]$.

(d) A non-zero proper ideal of $\mathbb{Z}[x]$.

Exercises

1. In Examples 7.1–7.10, determine in each case whether the subrings described there are ideals.

2. Prove that all ideals in $\mathbb{Q}[x]$ are principal, using a similar proof to that for $\mathbb{Z}$ (Theorem 11.6).

3. Consider the set

$$I = \{f \in \mathbb{Q}[x] : f(i) = 0\}.$$

(Here, i is the usual complex number.)

(a) Prove that I is an ideal.

(b) We know that I is a principal ideal (why?). Find a generator for I, and prove that it works. *Hint:* The generator should be an element in I with the smallest degree greater than 0.

4. Consider the set

$$I = \{f \in \mathbb{C}[x] : f(i) = 0\}.$$

Repeat Exercise 3 for this set.

5. Consider the set

$$I = \{f \in \mathbb{Q}[x] : f(3) = 0 \text{ and } f\left(\sqrt{3}\right) = 0\}.$$

Repeat Exercise 3 for this set.

6. Determine all the principal ideals of $\mathbb{Z}_{12}$, and draw a diagram describing their containment relations. Prove that $\mathbb{Z}_{12}$ has no other ideals.

7. Find two factorizations of 8 into irreducibles in $\mathbb{Z}[\sqrt{-7}]$ that are essentially distinct.

8. Consider the ring $\mathbb{Z}[\alpha]$, where $\alpha = \sqrt[3]{5}$, as described in Exercise 7.3. Describe the elements of the principal ideals $\langle \alpha \rangle$ and $\langle 2 \rangle$.

9. Consider

$$I = \left\{ \begin{pmatrix} a & 0 \\ b & 0 \end{pmatrix} \in M_2(\mathbb{Z}) : a, b \in \mathbb{Z} \right\}.$$

(a) Show that I is a subring of $M_2(\mathbb{Z})$.

(b) Show that I is *not* an ideal of $M_2(\mathbb{Z})$ (using the stronger definition of ideal used for non-commutative rings).

10. Consider the principal ideal $\langle 3 + \sqrt{5} \rangle$ in $\mathbb{Z}[\sqrt{5}]$. Prove that

$$\langle 3 + \sqrt{5} \rangle = \{c + d\sqrt{5} : 4|(c + d)\}.$$

11. Suppose that p, q are distinct prime integers in $\mathbb{Z}$. Prove that

$$\langle p \rangle \cap \langle q \rangle = \langle pq \rangle.$$

12. Let R be a commutative ring, and suppose that I and J are ideals. Prove that $I \cap J$ is an ideal. (Note the many examples of intersections of ideals provided in Exercise 11. Also compare this exercise to Exercise 7.9.) Now describe the ideal in Exercise 5 as an intersection of two proper ideals.

13. Suppose that R is a commutative ring, I and J are ideal of R, and $I \subseteq J$. Prove that I is an ideal of the ring J.

14. Let R be a commutative ring, with ideals I and J. Let

$$I + J = \{a + b : \ a \in I, \ b \in J\}.$$

Prove that $I + J$ is an ideal of R. For obvious reasons, we call the ideal $I + J$ the **sum** of the ideals I and J.

15. Suppose that R is a commutative ring with unity, and I and J are ideals. Define

$$I \cdot J = \left\{ \sum_{k=1}^{n} a_k b_k : \ a_k \in I, \ b_k \in J, \ n \in \mathbb{N} \right\}.$$

(That is, $I \cdot J$ consists of all possible finite sums of products of elements from I and J.)

(a) Prove that $I \cdot J$ is an ideal. We call the ideal $I \cdot J$ the **product** of the ideals I and J.

(b) Prove that $I \cdot J \subseteq I \cap J$.

(c) Prove that if $a, b \in R$, we have that $\langle a \rangle \cdot \langle b \rangle = \langle ab \rangle$.

(d) Show by example in $R = \mathbb{Z}$ that we can have $I \cdot J \subset I \cap J$.

16. Suppose that I, J, K are ideals in R, a commutative ring with unity. Prove that $I \cdot (J + K) = I \cdot J + I \cdot K$.

17. Let X be a set; consider the power set ring $P(X)$, described in Exercise 6.20. Pick a fixed subset a of X, and let

$$I = \{b \in P(X) : b \subseteq a\}.$$

(a) Prove that I is an ideal of $P(X)$.

(b) Prove that I is a principal ideal (find the generator!).

18. Prove that the nilradical of a commutative ring with unity is an ideal. (See Exercise 7.15, where you proved it is a subring.)

19. Suppose that R is a commutative ring with unity, and $r \in R$. Let $A(r) = \{s \in R : rs = 0\}$. (This set is called the **annihilator** of r.) Prove that $A(r)$ is an ideal.

20. Generalize Exercise 19: Suppose that R is a commutative ring with unity, and I is an ideal. Let
$$A(I) = \{s \in R : rs = 0, \text{ for all } r \in I\};$$
we call $A(I)$ the **annihilator** of I. Prove that $A(I)$ is an ideal.

21. Suppose that R and S are domains. Determine all possible ideals of the direct product $R \times S$, which occur as annihilators $A((r,s))$. *Hints:* See Exercise 19 for the definition of annihilator. In this problem there are only four cases.

22. Suppose that R is a commutative ring with unity, and e is an idempotent. In Exercise 7.25, we defined what this means and also considered the subring Re; we now know that this subring is the principal ideal $\langle e \rangle$. Prove that the annihilator $A(e)$ is a principal ideal. (See Exercise 19 for the definition of annihilator.)

23. Generalize Theorem 11.5: Suppose that R is a commutative ring with unity and $r, s \in R$, with r not a zero divisor. Prove that $\langle r \rangle = \langle s \rangle$ if and only if r, s are associates.

Chapter 12

Principal Ideal Domains

In the previous chapter we encountered the notions of ideal and principal ideal and proved that in $\mathbb{Z}$ (and in $\mathbb{Q}[x]$) all ideals are principal. In this chapter we will discover the close connection between this property and the Factorization Theorems we know are true for $\mathbb{Z}$ and $\mathbb{Q}[x]$.

12.1 Ideals that are not Principal

Before proceeding, we should first convince you that there do exist ideals that are *not* principal. It turns out that we can find such ideals in the domains $\mathbb{Z}[x]$ and $\mathbb{Z}[\sqrt{-5}]$.

To describe these examples we first introduce a little more general notation. Suppose that R is a commutative ring with unity and $a, b \in R$. We shall define $\langle a, b \rangle$ as
$$\{ax + by : x, y \in R\}.$$

We leave it as Exercise 12.1 below for you to prove that this is an ideal and is in fact the smallest ideal of R that contains both a and b. (In fact, this idea can be generalized to ideals generated by any finite number of elements; see Exercise 12.2.) Note that $\langle a, b \rangle$ is really just the set of all linear combinations of a and b (where the coefficients on a and b are allowed to be any ring elements).

Example 12.1

In the ring $\mathbb{Z}$, the ideal $\langle 12, 9 \rangle$ consists of the set of all linear combinations of 12 and 9. We know from our work about the integers that this is the set of all multiples of $\gcd(12, 9) = 3$, and so
$$\langle 12, 9 \rangle = \langle 3 \rangle.$$

This is a principal ideal, which shouldn't be too surprising, because we've already proved that *all* ideals of $\mathbb{Z}$ are principal.

Example 12.2

For a first example of a non-principal ideal, let's work in the ring $\mathbb{Z}[x]$. We claim that the ideal $\langle 2, x\rangle$ is not principal. To see this, we will first look more closely at arbitrary elements of this ideal. Any element of it is of the form $2f + xg$, where f and g are arbitrary polynomials from $\mathbb{Z}[x]$. Let's consider the constant term of this polynomial. The polynomial xg has no constant term, and so the constant term for $2f + xg$ is equal to the constant term of $2f$. This constant must be even. Thus, every element in $\langle 2, x\rangle$ has even constant term. But conversely, consider any polynomial in $\mathbb{Z}[x]$ with even constant term. We can write such a polynomial as $2n + xh$, where n is an integer and h is a polynomial. Thus, $\langle 2, x\rangle$ consists of all polynomials with even constant term. (Note of course that zero is an even integer.)

We now claim that $\langle 2, x\rangle$ is not a principal ideal. If it were, with generator f, then all polynomials in the ideal would be multiples of f. In particular, 2 and x would be multiples of f. Because 2 is a multiple of f, the degree of f must be zero. But the only zero-degree polynomials that have x as a multiple are ± 1. Neither of these could be f, because their constant terms are odd. Thus, $\langle 2, x\rangle$ is not a principal ideal.

Example 12.3

Now let's consider an example in the ring $\mathbb{Z}[\sqrt{-5}]$: We claim that $\langle 3, 1+\sqrt{-5}\rangle$ is an ideal that is not principal. To see this, we examine a typical element of this ideal; it is a linear combination of the elements 3 and $1 + \sqrt{-5}$. Such an element must be of the form

$$(a + b\sqrt{-5})3 + (c + d\sqrt{-5})(1 + \sqrt{-5}) =$$

$$(3a + c - 5d) + (3b + c + d)\sqrt{-5}.$$

(Note that in this linear combination we must choose as coefficients arbitrary elements $a + b\sqrt{-5}$ and $c + d\sqrt{-5}$ from the ring $\mathbb{Z}[\sqrt{-5}]$.)

What elements from $\mathbb{Z}[\sqrt{-5}]$ are of this form? Note that if we subtract the coefficients $3a + c - 5d$ and $3b + c + d$, we obtain $3(a - b) - 6d$, which is divisible by 3. We have thus shown that if $x + y\sqrt{-5} \in \langle 3, 1 + \sqrt{-5}\rangle$, then $x - y$ is divisible by 3. This means in particular that such elements as $5 + \sqrt{-5}$ and 1 are not elements of the ideal $\langle 3, 1 + \sqrt{-5}\rangle$. Thus, we have a proper ideal.

▷ **Quick Exercise.** Is $12 + 7\sqrt{-5}$ an element of $\langle 3, 1 + \sqrt{-5}\rangle$? ◁

But surprisingly this property actually characterizes elements of the ideal $\langle 3, 1 + \sqrt{-5}\rangle$. For if $x - y$ is divisible by 3, then $x - y = 3k$, and so

$$x + y\sqrt{-5} = x(1 + \sqrt{-5}) + (y - x)\sqrt{-5} =$$

$$x(1 + \sqrt{-5}) - k(\sqrt{-5})(3) \in \langle 3, 1 + \sqrt{-5}\rangle.$$

That is,

$$\langle 3, 1 + \sqrt{-5}\rangle = \{x + y\sqrt{-5} : 3 \text{ divides } (x - y)\}.$$

We now claim that $\langle 3, 1 + \sqrt{-5}\rangle$ is not a principal ideal. Suppose by way of contradiction that this ideal is generated by $\alpha = a + b\sqrt{-5}$. But then 3 and $1 + \sqrt{-5}$ are multiples of α, and so the norm $N(\alpha)$ divides both $N(3) = 9$ and $N(1 + \sqrt{-5}) = 6$. This means that $N(\alpha)$ is either 1 or 3. It can't be 1 (because then it would be a unit, which can't be an element of a proper ideal), and it can't be 3 because (as we observed in Example 10.15) the Diophantine equation $a^2 + 5b^2 = 3$ has no solutions. Thus, the ideal $\langle 3, 1 + \sqrt{-5}\rangle$ is a non-principal ideal in $\mathbb{Z}[\sqrt{-5}]$.

12.2 Principal Ideal Domains

The two examples above make the following definition more interesting. A domain is a **principal ideal domain** (or **PID**) if all its ideals are principal.

We have thus proved that $\mathbb{Z}$ and $\mathbb{Q}[x]$ are principal ideal domains, while $\mathbb{Z}[\sqrt{-5}]$ and $\mathbb{Z}[x]$ are not.

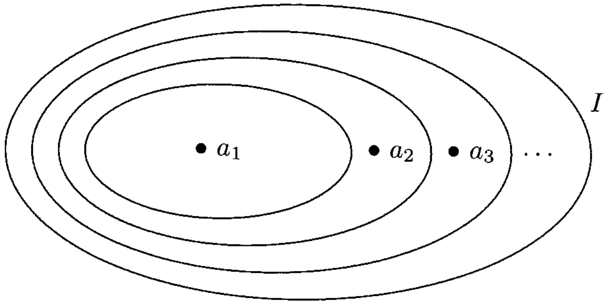
We now obtain the following elegant theorem that provides a sufficient condition for factorization into irreducibles. This returns us to the discussion of factorization near the beginning of Chapter 11.

Theorem 12.1 Factorization Theorem for PIDs *In a principal ideal domain, every non-zero non-unit is either irreducible or a product of irreducibles.*

Proof: Suppose by way of contradiction that R is a PID with at least one non-zero non-unit that is neither irreducible nor factorable as a product of irreducibles. By Lemma 11.2, we then have a properly ascending chain of principal ideals, where each a_i is a non-unit:

$$\langle a_1 \rangle \subset \langle a_2 \rangle \subset \langle a_3 \rangle \subset \cdots.$$

Consider now the set $I = \bigcup\{\langle a_i \rangle : i \in \mathbb{N}\}$. We claim that this is an ideal. First, we check that I is closed under subtraction. Given $x, y \in I$, there exist i and j such that $x \in \langle a_i \rangle$ and $y \in \langle a_j \rangle$. Suppose (without loss of generality) that j is the larger of i and j. Then $x, y \in \langle a_j \rangle$, and so $x - y \in \langle a_j \rangle \subseteq I$. Now I also satisfies the multiplicative absorption property: Suppose that $x \in I$ and $r \in R$. There exists i so that $x \in \langle a_i \rangle$, an ideal. So, $rx \in \langle a_i \rangle \subseteq I$. Thus, I is an ideal.



Because R is a PID, $I = \langle a \rangle$, for some $a \in R$. But I is the union of the $\langle a_j \rangle$'s, and so $a \in \langle a_j \rangle$, for some j; thus, $\langle a \rangle = \langle a_j \rangle$. But then

$$\langle a \rangle = \langle a_j \rangle = \langle a_{j+1} \rangle = \langle a_{j+2} \rangle = \cdots,$$

which is contrary to our assumption. This contradiction means that every non-unit can be factored into irreducibles. □

This abstract proof now gives an alternate approach to seeing that factorization holds for $\mathbb{Z}$ and for $\mathbb{Q}[x]$. The additional abstraction of this proof makes it more powerful. As we shall see, it applies to many more domains than our two familiar examples.

An important generalization of one of the ideas in the proof above is due to the great German mathematician Emmy Noether, who in the 1920s laid much of the important groundwork for axiomatic ring theory. She isolated as particularly important those commutative rings where every ascending sequence of ideals is finite. That is, if a set of ideals I_n is totally ordered under inclusion

$$I_1 \subseteq I_2 \quad \subseteq I_3 \subseteq \cdots \subseteq I_n \subseteq \cdots$$

then there must exist an integer j for which $I_j = I_{j+1} = \cdots$. We have seen in the proof above that PIDs satisfy this property. Such commutative rings are said to have the **ascending chain condition** (ACC for short) on ideals, and are called **Noetherian**.

Chapter Summary

In this chapter we discovered that not all ideals are principal; when all ideals of a domain are principal, we call it a *principal ideal domain*. We then proved the Factorization Theorem for all PIDs.

Warm-up Exercises

a. What is the ideal $\langle 35, 15 \rangle$ in $\mathbb{Z}$? What about $\langle 12, 20, 15 \rangle$?

b. What is the ideal $\langle 4, x \rangle$ in $\mathbb{Z}[x]$? What about $\langle 4, x^2 \rangle$?

c. What is the ideal $\langle x^2 - 3x + 2, x^2 - 2x + 1 \rangle$ in $\mathbb{Q}[x]$?

d. What is the ideal $\langle (1,0), (0,1) \rangle$ in $\mathbb{Z} \times \mathbb{Z}$? What about $\langle (1,1) \rangle$?

e. Why is a field always a PID, practically by default?

f. Give examples of the following (or explain why they don't exist):

   (a) A domain that is not a PID.
   (b) Elements $f, g \in \mathbb{Z}[x]$, so that $f \neq g$, but $\langle f \rangle = \langle g \rangle$.
   (c) Elements $f, g \in \mathbb{Q}[x]$, so that $\langle f, g \rangle = \langle f \rangle$ but $\langle f, g \rangle \neq \langle g \rangle$.

Exercises

1. Let R be a commutative ring with unity and $a, b \in R$. Prove that

$$\langle a, b \rangle = \{ax + by : x, y \in R\}$$

is an ideal; furthermore, show that it is the smallest ideal of R that contains a and b.

2. Let R be a commutative ring with unity and $a_1, a_2, \cdots, a_n \in R$. Prove that
$$\langle a_1, a_2, \cdots, a_n \rangle =$$
$$\{a_1x_1 + a_2x_2 + \cdots + a_nx_n : x_i \in R, i = 1, 2, \cdots, n\}$$
is an ideal; furthermore, show that it is the smallest ideal of R that contains all the a_i's. We call the a_i's the **generators** of the ideal; we say that the ideal is **finitely generated**. Note that each element of the ideal can be expressed as a *linear combination* of its generators.

3. Let n be a positive integer and consider the ideals
$$\langle x^n \rangle$$
in $\mathbb{Q}[x]$. Describe succinctly the elements of $\langle x^n \rangle$. What containment relations hold among these ideals? Explain why $\mathbb{Q}[x]$ is Noetherian. Explain why the ideals $\langle x^n \rangle$ do not contradict the assertion that $\mathbb{Q}[x]$ is Noetherian.

4. Consider the ideal $I = \langle 3, 1 - \sqrt{-5} \rangle$ in $\mathbb{Z}[\sqrt{-5}]$. Prove that
$$I = \{x + y\sqrt{-5} : 3 \text{ divides } (x + y)\}.$$
Show that I is not principal. (See Example 12.3.)

5. Let X be an arbitrary set, and consider the power set ring $P(X)$. (See Exercise 6.20, where we made $P(X)$ a ring, by equipping it with an addition (symmetric difference) and a multiplication (intersection).) In Exercise 11.17 we discussed a principal ideal of $P(X)$; that exercise is relevant to the present problem but not strictly necessary.

   (a) Let $a \in P(X)$. Describe the elements of the principal ideal $\langle a \rangle$.

   (b) Suppose that X has more than one element. Show that $P(X)$ is not a domain.

   (c) Suppose that X has infinitely many elements. Let
   $$I = \{a \in P(X) : a \text{ has finitely many elements}\}.$$
   Prove that I is an ideal of $P(X)$. Show that I is not a principal ideal.

6. Consider the domain $\mathbb{Z}[\sqrt{3}]$. Prove that in this ring,
$$\langle 1 + \sqrt{3} \rangle = \{x + y\sqrt{3} : x + y \text{ is an even integer}\}.$$

7. Consider the ring $\mathbb{Z}[\sqrt{2}]$. Prove that
$$\langle 3 + 8\sqrt{2}, 7 \rangle = \langle 3 + \sqrt{2} \rangle.$$

8. Consider the ring S of real-valued sequences considered in Exercise 6.19.

   (a) Let n be a fixed positive integer, and let
   $$I_n = \{\{s_k\} \in S : s_m = 0, \text{ for all } m > n\}.$$
   Prove that I_n is an ideal of S.

   (b) Use the ideals in part a to show that S is not a Noetherian ring.

   (c) Let
   $$\Sigma = \{\{s_n\} \in S : \text{at most finitely many } s_i \neq 0\};$$
   prove that Σ is an ideal of S. What is the relationship between Σ and the I_i's?

   (d) Prove that Σ is not finitely generated. Recall from Exercise 2 that by this we mean that
   $$\Sigma \neq \langle \vec{s}_1, \vec{s}_2, \cdots, \vec{s}_n \rangle,$$
   for any finite set of sequences $\vec{s}_i$.

9. Consider again the ring S of real-valued sequences. Let
$$B = \{\{s_n\} \in S : \text{there exists } M \in \mathbb{R} \text{ with } |s_n| \leq M, \text{ for all } n\}.$$
These are the **bounded** sequences. Note that M is not fixed in the definition of B; that is, different sequences may require different bounds. Prove that B is a subring, but is nonetheless *not* an ideal of S.

10. Let I be an ideal of $\mathbb{Z}[\sqrt{n}]$, where n is a square-free integer. Define
$$\bar{I} = \{a + b\sqrt{n} : a - b\sqrt{n} \in I\}.$$

(a) Prove that $\bar{I}$ is an ideal of $\mathbb{Z}[\sqrt{n}]$.

(b) Provide particular examples of such ideals, where $I \neq \bar{I}$, and where $I = \bar{I}$.

(c) Prove that I is a principal ideal if and only if $\bar{I}$ is a principal ideal.

11. In this exercise we generalize an argument used in the proof of Theorem 12.1. Let R be a commutative ring with unity, and suppose that I_n is a proper ideal for all positive integers n. Suppose further that

$$I_1 \subseteq I_2 \cdots \subseteq I_n \subseteq I_{n+1} \subseteq \cdots .$$

Let $I = \bigcup I_n$. Prove that I is a proper ideal.

12. Consider $\mathbb{Q}[x, y]$, the set of all polynomials with coefficients from $\mathbb{Q}$, in the two indeterminates x and y. In Exercise 6.23, you showed that the ring of polynomials $R[x]$ with coefficients in any given commutative ring makes sense. Applying this construction with coefficients from $\mathbb{Q}[x]$, where we then have to use a different symbol y for the new indeterminate, gives us the ring $\mathbb{Q}[x, y]$. It turns out (though we won't verify all details here) that addition and multiplication in this ring behave just as you would expect. Formally then, an element of $\mathbb{Q}[x, y]$ can be viewed as an element of the form

$$a_{0,0} + a_{1,0}x + a_{0,1}y + a_{2,0}x^2 + a_{1,1}xy + a_{0,2}y^2 + \cdots ,$$

where the $a_{i,j}$'s are rational numbers, and only finitely many of them are not zero.

(a) Provide a nice description of the elements in the ideal $\langle x, y \rangle$.

(b) Show that the ideal $\langle x, y \rangle$ is not principal, thus showing that $\mathbb{Q}[x, y]$ is not a PID.

Chapter 13

Primes and Unique Factorization

You should now recall the proof of the uniqueness of factorization into irreducibles for $\mathbb{Z}$ (or $\mathbb{Q}[x]$). Our intent in this chapter is to construct a more general context in which this proof is true.

13.1 Primes

This proof relies heavily on the fact that irreducible elements in $\mathbb{Z}$ (or $\mathbb{Q}[x]$) are in fact prime. It should not then be surprising that a general unique factorization theorem should rely on the same considerations. We now define prime elements in an arbitrary domain: A non-unit $p \neq 0$ of a domain R is a **prime** if, whenever p divides ab, then p divides a or b.

Example 13.1

Under this new definition, the prime integers remain prime in $\mathbb{Z}$. And the irreducible polynomials in $\mathbb{Q}[x]$ are prime also.

Example 13.2

Let's show directly that the element $\sqrt{3}$ is prime in the domain $\mathbb{Z}[\sqrt{3}]$. For that purpose, we must suppose that $\sqrt{3}$ divides a product $\alpha\beta$, where $\alpha, \beta \in \mathbb{Z}[\sqrt{3}]$. Now $\alpha = a + b\sqrt{3}$ and $\beta = c + d\sqrt{3}$, and

$$\alpha\beta = (ac + 3bd) + (ad + bc)\sqrt{3}.$$

If $\sqrt{3}$ divides this product, it must clearly divide the rational part $ac + 3bd$. But $\sqrt{3}$ obviously divides 3, and so $\sqrt{3}$ must divide ac, in $\mathbb{Z}[\sqrt{3}]$. But ac is an integer, and so the only way this can occur is if 3 actually divides ac. Now 3 is a prime *integer*, and so (without loss of generality) 3 divides a. But then $\sqrt{3}$ divides $a + b\sqrt{3}$, as required.

Note that an easy inductive proof (identical to that for $\mathbb{Z}$) shows that if p is prime and it divides a product of n terms, then p divides at least one of the factors. (See Exercise 2.5 and Exercise 5.3.)

Now examine the proof that for $\mathbb{Z}$, irreducible and prime elements are the same (Theorem 2.7). Notice that the proof that primeness implies irreducibility is general, while the converse depends on the GCD identity (a theorem which need not hold for arbitrary domains). We thus have the following:

Theorem 13.1 *In any domain, prime elements are irreducible.*

Proof: Check that the proof for $\mathbb{Z}$ (Theorem 2.7) holds in general. (This is Exercise 13.1.) □

Example 13.3

The converse of this theorem is in general false, and the factorization

$$6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5})$$

in $\mathbb{Z}[\sqrt{-5}]$ that we considered in Section 11.1 provides the counterexample, for 2 is irreducible, yet it divides $(1+\sqrt{-5})(1-\sqrt{-5})$ without dividing either factor and hence *is not* prime.

The previous example finally justifies our long-standing careful distinction between the concepts of irreducibility and primeness.

The following theorem now shows that uniqueness of factorization fails in a domain exactly when the concepts of irreducibility and primeness fail to coincide:

Theorem 13.2 *Consider a domain in which every non-zero non-unit is either an irreducible or a product of irreducibles. Then all irreducible elements are prime if and only if the factorization of non-units into irreducibles is unique, up to order and unit factors.*

Proof: Suppose that D is a domain with unique factorization, and p is an irreducible element. We wish to show that p is prime and toward that end assume that $p|ab$. Then $pc = ab$, for some $c \in D$. Now factor c, a, and b into irreducibles, thus obtaining the following equation:

$$pc_1c_2\dots c_k = a_1a_2\dots a_mb_1b_2\dots b_n.$$

By the uniqueness of the factorization of pc, p must be an associate of some a_i or b_j; that is, $p|a$ or $p|b$, as required.

For the other direction, check that the proof for $\mathbb{Q}[x]$ works in general. (This is Exercise 13.2; see Theorem 5.4.) □

Example 13.4

In Exercise 11.7 you obtained two essentially distinct factorizations of 8 in $\mathbb{Z}[\sqrt{-7}]$; namely, as

$$2 \cdot 2 \cdot 2 = (1 + \sqrt{-7})(1 - \sqrt{-7}).$$

Theorem 13.2 means that of necessity $\mathbb{Z}[\sqrt{-7}]$ must possess some element that is irreducible, but not prime. It is in fact easy in this case to identify a particular element that plays this role.

▷ **Quick Exercise.** What element is that? ◁

13.2 UFDs

We now introduce some convenient terminology: A **unique factorization domain** (or **UFD**) is a domain in which all non-zero non-units can be factored uniquely (up to units and order) into irreducibles. More concretely, uniqueness of factorization means that if a is a non-unit, and

$$a = a_1a_2\cdots a_n = b_1b_2\cdots b_m,$$

where the a_i and b_i are irreducibles, then $n = m$, and under some rearrangement of the b_i we find a_i and b_i are associates, for $i = 1, \cdots n$.

Let's now rephrase Theorem 13.2 in light of our new terminology: We showed that if factorization of non-units into irreducibles is always possible, then the domain is a UFD if and only if each irreducible element is prime. Following the model of our PIDs $\mathbb{Z}$ and $\mathbb{Q}[x]$, we'd now like to prove that *all* PIDs are UFDs. Because we already know that factorization occurs in PIDs, we can complete this proof by showing that in a PID all irreducible elements are prime. Our analysis of this question depends on discussing a bit more closely the principal ideals of a domain in general, and of a PID in particular.

13.3 Expressing Properties of Elements in Terms of Ideals

Let D be a domain with $a, b \in D$. We first note that divisibility can be phrased in terms of ideal containment. For a divides b if and only if $b = ax$, for some $x \in D$, and this is true if and only if $\langle b \rangle \subseteq \langle a \rangle$. Furthermore, this containment is proper if and only if $b = ax$, for some non-unit $x \in D$.

▷ **Quick Exercise.** Show that $\langle b \rangle \subset \langle a \rangle$ exactly if $b = ax$, for some non-unit $x \in D$. ◁

Example 13.5

Thus, in $\mathbb{Z}$ the ideal $\langle 6 \rangle$ is a proper subset of $\langle 2 \rangle$, because $6 = 2 \cdot 3$, and 3 is a non-unit.

Note that we already referred to a particular case of this result in Chapter 11: $\langle a \rangle = \langle b \rangle$ if and only if a and b are associates (Theorem 11.5).

Now suppose that $p \in D$ is an irreducible. Then if a divides p, we have that either a is a unit or a and p are associates. Thus, if p is an irreducible and $\langle p \rangle \subseteq \langle a \rangle$, then either $\langle a \rangle = D$ (if a is a unit), or else $\langle a \rangle = \langle p \rangle$ (if a and p are associates). This means that p is irreducible exactly if D has no proper principal ideals strictly larger than $\langle p \rangle$. In a PID all ideals are of course principal, and so in this case p is irreducible exactly if D has no proper ideals strictly larger than $\langle p \rangle$.

This latter property of the ideal $\langle p \rangle$ deserves its own definition: An ideal I of a ring R is **maximal** if the only ideal properly containing I is R. Using our new terminology, we have shown that in a PID, p is an irreducible element if and only if $\langle p \rangle$ is a maximal ideal.

Example 13.6

In $\mathbb{Z}$, there is no proper ideal larger than $\langle 3 \rangle$, because 3 is irreducible. But we can see this directly: Because $\mathbb{Z}$ is a PID, a larger ideal would have to be principal, and $\langle 3 \rangle \subseteq \langle a \rangle$ is true exactly if a divides 3. That is, a must be ± 1 (and so $\langle a \rangle = \mathbb{Z}$), or else a must be ± 3 (and so $\langle 3 \rangle = \langle a \rangle$).

You are forgiven for doubting the usefulness of translating statements about elements into statements about ideals. After all, an individual element seems rather more concrete and understandable than the more abstract notion of an ideal. However, it is one of the important lessons of abstract algebra that the abstract ideal is often *easier* to deal with than the concrete element. For example, the annoying and insignificant distinction between $x^2 + 1$ and $\frac{5}{2}x^2 + \frac{5}{2}$ in $\mathbb{Q}[x]$ disappears when we consider the ideal $\langle x^2 + 1 \rangle = \langle \frac{5}{2}x^2 + \frac{5}{2} \rangle$. We will later see more dramatic evidence of the simplicity and clarity that comes from discussing ideals rather than elements.

Let us record here for future reference the translations we have made up to now between statements about elements and statements about principal ideals:

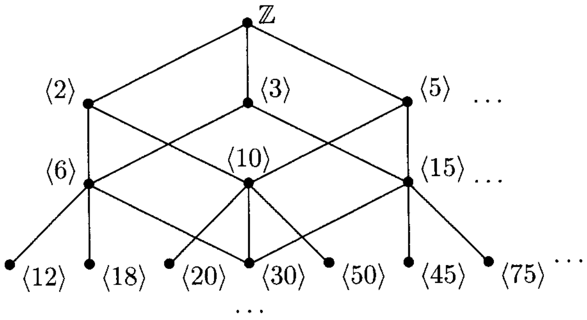
Theorem 13.3 *Let D be a domain, and $a, b \in D$.*

a. $\langle a \rangle = D$ if and only if a is a unit.

b. $\langle b \rangle \subseteq \langle a \rangle$ if and only if a divides b; this inclusion is proper if and only if $b = ax$ where x is a non-unit.

c. a is irreducible if and only if $\langle a \rangle$ is maximal among all proper principal ideals. If D is a PID, this is true if and only if $\langle a \rangle$ is a maximal ideal.

Let's now look at the whole ideal structure of $\mathbb{Z}$ in light of the above discussion. We know that the maximal ideals are given exactly by the ideals of the form $\langle p \rangle$, where p is an irreducible (or prime) integer, and containments reflect divisibility:



What is the meaning in a computational sense of the maximality of a principal ideal in a domain? If $\langle p\rangle$ is maximal and $a \notin \langle p\rangle$, then the ideal $\langle p, a\rangle$ must be the entire domain, because it is certainly a strictly larger ideal than $\langle p\rangle$. But $\langle p, a\rangle$ is the entire domain exactly if it contains (any) unit, and in particular if $1 \in \langle p, a\rangle$. That is, $1 = px + ay$, for some x, y in the domain.

Example 13.7

Consider the maximal ideal $\langle 7\rangle$ in $\mathbb{Z}$. Now, $12 \notin \langle 7\rangle$. But

$$3 \cdot 12 + (-5) \cdot 7 = 1,$$

and so $\langle 7, 12\rangle = \mathbb{Z}$. Of course, the fact that there exist x and y such that $12x + 7y = 1$ is just an example of the GCD identity for $\mathbb{Z}$. This reveals a more general principle operating in the ideal structure of $\mathbb{Z}$: $\langle a, b\rangle = \langle c\rangle$, where $c = \gcd(a, b)$. (See Exercise 13.6.)

Example 13.8

Because x is an irreducible element in $\mathbb{Q}[x]$, we know from Theorem 13.3 that $\langle x\rangle$ is a maximal ideal. To see this computationally, choose $f \notin \langle x\rangle$. Because $\langle x\rangle$ consists of the polynomials in $\mathbb{Q}[x]$ with zero constant term, f must have a non-zero constant term c, and so f can be written as $f = c + xg$, where g is some polynomial in $\mathbb{Q}[x]$. But direct computation shows that

$$1 = c^{-1}f + (-c^{-1}g)x.$$

This last expression is a linear combination of f and x, and so $1 \in \langle x, f\rangle$. Thus, $\langle x, f\rangle = \mathbb{Q}[x]$, and so $\langle x\rangle$ is a maximal ideal, as claimed.

13.4 Ideals in $\mathbb{Z}[\sqrt{-5}]$

Unfortunately, this sort of structure fails in domains that are less nice than $\mathbb{Z}$. Consider for example the irreducible 3 in $\mathbb{Z}[\sqrt{-5}]$. Because 3 is irreducible, we know by Theorem 13.3c that $\langle 3\rangle$ is maximal among proper principal ideals; but it need not be maximal among *all* proper ideals (because $\mathbb{Z}[\sqrt{-5}]$ has non-principal ideals). And indeed, we have already seen (Example 12.3) that $\langle 3, 1+\sqrt{-5}\rangle$ is a non-principal ideal with $\langle 3\rangle \subset \langle 3, 1+\sqrt{-5}\rangle$.

Now we claim that the ideal $\langle 3, 1+\sqrt{-5}\rangle$ *is* maximal in $\mathbb{Z}[\sqrt{-5}]$. How would we prove this? We need to show that if $a+b\sqrt{-5} \notin \langle 3, 1+\sqrt{-5}\rangle$, then $\langle 3, 1+\sqrt{-5}, a+b\sqrt{-5}\rangle$ is the whole domain. We show this by giving a linear combination of these three elements equal to 1. But recall that $\langle 3, 1+\sqrt{-5}\rangle = \{x + y\sqrt{-5} : 3|(x-y)\}$. Because $a + b\sqrt{-5}$ is not in this ideal, this means that $a - b$ is not divisible by 3 (and hence, relatively prime to 3). Thus, by the GCD identity for $\mathbb{Z}$, there are integers z and w with $3z + (a-b)w = 1$. But then

$$1 = z(3) + (-bw)(1+\sqrt{-5}) + (w)(a+b\sqrt{-5}),$$

the linear combination which we required.

13.5 A Comparison between $\mathbb{Z}$ and $\mathbb{Z}[\sqrt{-5}]$

Notice the distinction between $\mathbb{Z}$ and $\mathbb{Z}[\sqrt{-5}]$: In $\mathbb{Z}$ the ideal $\langle a, b\rangle$ is always principal, and a generator is $\gcd(a, b)$, which can be expressed as $ax + by$ for some x and y. In particular, if a and b have no common divisors (other than ±1), then 1 can be so expressed. That is, $\langle 1\rangle = \langle a, b\rangle$ if and only if a and b have no non-trivial common divisors.

On the other hand, in $\mathbb{Z}[\sqrt{-5}]$, 3 and $1+\sqrt{-5}$ have no common divisors (except units), but 'ought to', because

$$1 \notin \langle 3, 1+\sqrt{-5}\rangle \subset \mathbb{Z}[\sqrt{-5}].$$

This is precisely why the 19th-century German mathematician Ernst Kummer first used the term 'ideal'. He viewed an ideal like $\langle 3, 1+\sqrt{-5}\rangle$ as an 'ideal number' playing the role of a gcd for 3 and $1+\sqrt{-5}$, and thus filling in a gap in the multiplicative structure of $\mathbb{Z}[\sqrt{-5}]$. This, as we shall see, is one of the advantages of considering ideals, instead of elements.

Actually, Kummer's formulation of ideal was different than ours; our definition is due to Richard Dedekind. The crux of the Dedekind version of the definition is to identify the 'ideal number' with the set of all those numbers that ought to be its multiples. The set-theoretic nature of this definition is typical of a modern mathematical definition; Dedekind was a pioneer of this set-theoretic approach, and not only to the definition of ideal, but also in his axiomatic construction of the real numbers.

13.6 All PIDs are UFDs

We can now use what we have learned about ideals in PIDs to prove the following crucial theorem, which allows us to then conclude that *all* PIDs are UFDs:

Theorem 13.4 *In a PID, all irreducibles are prime.*

Proof: Suppose p is irreducible in the PID D. To show that p is prime, we suppose further that $p|ab$. We claim that $p|a$ or $p|b$. Suppose that p does *not* divide a. Then $a \notin \langle p \rangle$, and so $\langle p \rangle \subset \langle a, p \rangle$. Because $\langle p \rangle$ is maximal this means that $\langle a, p \rangle = D$, and so $1 \in \langle a, p \rangle$. That is, $1 = ax + py$, for some $x, y \in D$. But then $b = abx + pby$. Now $p|pby$ and $p|abx$, and so $p|b$. Thus, p is prime, as claimed. □

We have actually encountered special cases of the previous argument twice before: namely, for $\mathbb{Z}$ (in the proof of Theorem 2.7) and for $\mathbb{Q}[x]$ (in the proof of Theorem 5.2, where you did the proving). What we used in those arguments was the GCD identity. In the argument above we arrive at the conclusion $1 = ax + py$ by using the fact that $\langle p \rangle$ is a maximal ideal.

From Chapter 12 we know that any non-unit in a PID is irreducible or factorable into irreducibles (Theorem 12.1). From Theorem 13.2 we know that uniqueness of factorization is equivalent to irreducible elements being prime. Because we've just proved that this is true in a PID, we have the following theorem:

Theorem 13.5 *Any PID is a UFD.*

This theorem thus encompasses both the Fundamental Theorem of Arithmetic for $\mathbb{Z}$, and the Unique Factorization Theorem for $\mathbb{Q}[x]$. In Chapter 15 we will discover other PIDs as well, and hence other UFDs too.

The natural question to ask at this point is: Is the converse of this theorem true? That is, are all UFDs PIDs? The answer is no, and we have already encountered a domain which will serve as a counterexample. That domain is $\mathbb{Z}[x]$. In Example 12.2 we argued that $\langle 2, x \rangle$ is a non-principal ideal, and so $\mathbb{Z}[x]$ is not a PID. It thus remains to show that this domain is in fact a UFD. In order to accomplish this, we will need to inquire more carefully into the relationship between $\mathbb{Z}[x]$ and $\mathbb{Q}[x]$. Note that the relationship cannot be too simple, because the former ring is not a PID, while the latter ring is. This inquiry is the subject of Chapter 14; it requires Gauss's Lemma 5.5.

Chapter Summary

In this chapter we considered the notion of *prime* element in an arbitrary domain. We showed that prime elements are always irreducible, but the converse need not be true. If a domain has a factorization theorem and all irreducible elements are prime, then factorizations are essentially unique. We call such a domain a UFD. We then proved that all PIDs are UFDs.

Warm-up Exercises

a. Give examples of the following (or explain why no example exists). You should specify both a domain and an element of it:

   (a) A prime element that isn't irreducible.

   (b) An irreducible element that isn't prime.

   (c) A non-unit that is neither prime nor irreducible.

b. Give examples of the following (that is, specify both a domain and an ideal of it):

   (a) A principal ideal that is maximal.

(b) A principal ideal that is not maximal.

(c) A maximal ideal that is not principal.

(d) An ideal that is neither maximal nor principal.

c. Translate the following statements into statements about principal ideals:

(a) 3 divides 12 in $\mathbb{Z}$.

(b) $2 + \sqrt{3}$ is a unit in $\mathbb{Z}[\sqrt{3}]$.

(c) $2 + i$ and $-1 + 2i$ are associates in $\mathbb{Z}[i]$.

(d) $2x + 1$ is irreducible in $\mathbb{Q}[x]$.

(e) $2x + 1$ is irreducible in $\mathbb{Z}[x]$.

d. Why is the following statement silly? Every commutative ring has exactly one maximal ideal because the whole ring is an ideal, which is obviously as large as possible.

e. Why do we care whether all irreducible elements of a domain are prime?

f. Explain which implies which: UFD and PID.

g. What does the GCD identity have to do with the fact that in $\mathbb{Z}$, $\langle 9, 50\rangle = \mathbb{Z}$?

Exercises

1. Prove Theorem 13.1: In any domain, prime elements are irreducible.

2. Complete the proof of Theorem 13.2. That is, suppose we have a domain in which all non-zero non-units are irreducible, or a product of irreducibles. Furthermore, suppose all irreducible elements are prime. Prove that the factorization of any non-unit into irreducibles is unique (up to order and unit factors).

3. Exhibit in the ring $\mathbb{Z}_6$ a non-unit a for which $a^n = a$, for all positive integers n. Why does this mean that there is no unique factorization into irreducibles for this ring? Now repeat this exercise for $\mathbb{Z} \times \mathbb{Z}$.

4. Prove that x is a prime element of $\mathbb{Z}[x]$.

5. Consider $1 + i \in \mathbb{Z}[i]$.

   (a) Show that

$$\begin{aligned}\langle 1 + i\rangle &= \{a + bi : a + b \text{ is even}\}\\ &= \{\alpha \in \mathbb{Z}[i] : N(\alpha) \text{ is even}\}.\end{aligned}$$

   (b) Use part a to prove that $1 + i$ is a prime element of $\mathbb{Z}[i]$.

6. Suppose that $a, b, c \in \mathbb{Z}$ and $c = \gcd(a, b)$. Show that $\langle a, b\rangle = \langle c\rangle$. (See Example 13.7 for an illustration of this.)

7. Find two integers n with $n \leq -5$ where 2 is not prime in $\mathbb{Z}[\sqrt{n}]$. (Note by Exercise 10.11 that 2 is irreducible in these rings.)

8. Describe the elements of the ideal $\langle 3, x\rangle$ in $\mathbb{Z}[x]$. Show that $\langle 3, x\rangle$ is not a principal ideal. (See Example 12.2.)

9. Show that $\langle 3, x\rangle$ is a maximal ideal in $\mathbb{Z}[x]$.

10. Show that $\langle 9, x\rangle$ is an ideal in $\mathbb{Z}[x]$, which is neither principal nor maximal.

11. Consider the ring $\mathbb{Z}[\sqrt{2}]$.

   (a) Show that

$$I = \{a + b\sqrt{2} \in \mathbb{Z}[\sqrt{2}] : a \text{ is even}\}$$

   is an ideal.

   (b) Show that I is principal. *Hint:* Think of 'small' elements of I.

   (c) Show that I is a maximal ideal.

   (d) Show that

$$J = \{a + b\sqrt{2} \in \mathbb{Z}[\sqrt{2}] : b \text{ is even}\}$$

   is closed under subtraction but is not an ideal.

12. Using arguments similar to those used in Section 13.4 for $\mathbb{Z}[\sqrt{-5}]$, show that $\langle 2, 1 + \sqrt{-7}\rangle$ is a maximal ideal in $\mathbb{Z}[\sqrt{-7}]$ that is not principal.

13. Consider the ideal $\langle 7+\sqrt{-5}, 9\rangle$ in $\mathbb{Z}[\sqrt{-5}]$. Show that this ideal is neither maximal nor principal.

14. (a) Why is x^2+1 irreducible in $\mathbb{Q}[x]$?

    (b) By part a we now know that $\langle x^2+1\rangle$ is a maximal ideal in $\mathbb{Q}[x]$. Provide a direct argument for this, similar to the corresponding argument for $\langle x\rangle$ in Example 13.8.

15. Let n be a square-free integer and suppose that $n \equiv 1(\text{mod } 4)$. Prove that $\mathbb{Z}[\sqrt{n}]$ is not a UFD.

16. Consider the ideal $\langle x\rangle$ in $\mathbb{Q}[x,y]$; see Exercise 12.12 for more about this ring and this ideal. Show that this ideal is not maximal.

17. Consider the element $x+y$ in $\mathbb{Q}[x,y]$; see Exercise 12.12 for more about this ring. Argue that $x+y$ is irreducible, and so $\langle x+y\rangle$ is maximal among all principal ideals in $\mathbb{Q}[x,y]$. Show that $\langle x+y\rangle$ is not maximal among all ideals of $\mathbb{Q}[x,y]$.

18. In Exercise c part e, your translation should have been that

$$\langle 2x+1\rangle$$

is maximal among all principal ideals in $\mathbb{Z}[x]$. Show that this ideal is *not* maximal among all ideals, by considering the ideal

$$\langle 2x+1, 3\rangle.$$

Chapter 14

Polynomials with Integer Coefficients

Our aim in this chapter is to prove that $\mathbb{Z}[x]$, the ring of polynomials with integer coefficients, is a UFD. It probably seems plausible that every polynomial with integer coefficients can be factored uniquely into irreducibles, but the proof of the analogous statement for $\mathbb{Q}[x]$ will not work for $\mathbb{Z}[x]$.

14.1 The Proof that $\mathbb{Q}[x]$ is a UFD

Let's recall how we prove that $\mathbb{Q}[x]$ is a UFD, to see where the argument fails for $\mathbb{Z}[x]$. We proved this result for $\mathbb{Q}[x]$ originally in Theorems 5.1 and 5.4 and then provided another proof in Theorem 13.5 which depends on the fact that $\mathbb{Q}[x]$ is a PID. In either proof we showed two things: (1) non-units factor into irreducibles, and (2) such factorizations are unique. We proved in Chapter 5 that non-units factor by using induction on degree, and in Chapter 13 by using the fact that $\mathbb{Q}[x]$ is a PID. We then proved that such factorizations are unique by showing that in $\mathbb{Q}[x]$ irreducible elements are prime. In the first version of our proof, this latter depended on the GCD identity; in the proof in Chapter 13, this depended on the fact that in a PID an element a is irreducible if and only if $\langle a\rangle$ is a maximal ideal.

Because the idea of degree still makes good sense, we will find that proving the existence of factorizations is not difficult. However, there is no GCD identity for $\mathbb{Z}[x]$. Consider 2 and x in $\mathbb{Z}[x]$. A gcd in $\mathbb{Z}[x]$ for these two elements is 1, but 1 cannot be written as a linear combination of 2 and x. Furthermore, $\mathbb{Z}[x]$ is not a PID: We saw already in Example 12.2 that $\langle 2, x\rangle$ is not a principal ideal. Note that the lack of a GCD identity (for the elements 2 and x) and the fact that $\langle 2, x\rangle$ is not principal, amount to the same thing. So, we cannot use either of the proofs for $\mathbb{Q}[x]$ to show that all irreducible elements are prime in $\mathbb{Z}[x]$.

14.2 Factoring Integers out of Polynomials

We shall now concentrate on showing that every non-zero non-unit in $\mathbb{Z}[x]$ can be factored into irreducibles (recall that 1 and -1 are the only units in $\mathbb{Z}[x]$). Let's restrict our attention first to those factorizations that involve elements of $\mathbb{Z}$. For example, the factorization

$$6x^2 - 12x + 24 = 2 \cdot 3(x^2 - 2x + 4),$$

while considered trivial in $\mathbb{Q}[x]$, is of interest in $\mathbb{Z}[x]$ because 2 and 3 are *not* units in $\mathbb{Z}[x]$.

Consider now an irreducible p from $\mathbb{Z}$. We can also consider p as an element of $\mathbb{Z}[x]$ (of degree 0), and so we can ask whether p is an irreducible element of $\mathbb{Z}[x]$. But if $p = fg$, where $f, g \in \mathbb{Z}[x]$, then f and g would both have to have degree 0 (because $\deg(fg) = \deg(f) + \deg(g)$). This means that $p = fg$ can be considered a factorization in $\mathbb{Z}$, and so one of f and g must be a unit in $\mathbb{Z}$ (and hence in $\mathbb{Z}[x]$). Thus, irreducible elements in $\mathbb{Z}$ are also irreducible elements in $\mathbb{Z}[x]$. For example, 2 (considered as a degree 0 polynomial) is an irreducible element of $\mathbb{Z}[x]$.

Example 14.1

It is *not* always the case that an irreducible element in a smaller ring stays irreducible in a larger one: 2 is irreducible in $\mathbb{Z}$ but is *not* irreducible in $\mathbb{Z}[i]$, because $2 = (1+i)(1-i)$. (See Example 10.14.)

We know that an element $\mathbb{Z}$ is prime if and only if it is irreducible. On the other hand, we have not (yet) shown that $\mathbb{Z}[x]$ is a UFD, and, consequently, we cannot yet infer that irreducible elements in $\mathbb{Z}[x]$ are prime. We will show this eventually, but you should exercise care until then to preserve the (potential) distinction between irreducible and prime elements.

14.3 The Content of a Polynomial

So, a first step toward factoring elements in $\mathbb{Z}[x]$ is to factor out any non-trivial constant (that is, degree 0) elements. To describe this efficiently, we introduce the following terminology:

Given a polynomial

$$f = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0 \in \mathbb{Z}[x],$$

consider a gcd in $\mathbb{Z}$ of the elements $a_n, a_{n-1}, \cdots a_0$. We shall call such a gcd the **content** of f and denote it by cont f. Of course, the content is only well defined up to plus or minus (that is, unit multiples), but the notation cont f is convenient enough that we will live with this harmless ambiguity.

Example 14.2

The content of the polynomial $6x^2 - 12x + 24$ in $\mathbb{Z}[x]$ is 6 (or -6).

Recall from Example 10.9 that we called a polynomial in $\mathbb{Z}[x]$ **primitive** if its coefficients have no non-trivial common factor; that is, it is primitive if its content is 1. Our first important theorem asserts that the set of primitive polynomials is closed under multiplication.

Example 14.3

Both $2x+3$ and $3x^2+4x+1$ are primitive polynomials; note that their product $6x^3 + 17x^2 + 24x + 18$ is then primitive as well.

▷ **Quick Exercise.** Choose two other primitive polynomials and check that their product remains primitive. ◁

Theorem 14.1 *Suppose that f and g are primitive polynomials in $\mathbb{Z}[x]$. Then fg is primitive.*

Proof: This is just Gauss's Lemma 5.5 in disguise. Suppose that f and g are primitive polynomials, and let $d = \text{cont}\, fg$. Consider the polynomial $h = (1/d)fg$; this is an element of $\mathbb{Z}[x]$. Now $h = ((1/d)f)(g)$ is a factorization in $\mathbb{Q}[x]$; by Gauss's Lemma, this leads to a factorization in $\mathbb{Z}[x]$. That is, there are rational numbers A, B so that $h = (A(1/d)f)(Bg)$, with $A(1/d)f$ and Bg being elements of $\mathbb{Z}[x]$; note that $AB = 1$. Because g is primitive and $Bg \in \mathbb{Z}[x]$, we must have that B is an integer, because if B were not an integer, then the denominator of B would be cancelled by a non-trivial integer factor of all the coefficients of g; that is, cont g would not be 1. But because f is primitive, this means that A/d must be an integer too. Because $AB = 1$, this means that A, B, and d must each be ± 1. That is, cont $fg = \pm 1$, and so fg is primitive. □

An important consequence of this theorem is that the content function preserves multiplication:

Corollary 14.2 *Given $f, g \in \mathbb{Z}[x]$, we have that*

$$\text{cont } fg = \text{cont } f \cdot \text{cont } g.$$

Proof: Given $f, g \in \mathbb{Z}[x]$, we have that $f = (\text{cont } f)f_1$ and $g = (\text{cont } g)g_1$, where f_1, g_1 are primitive. But then

$$fg = (\text{cont } f \cdot \text{cont } g)f_1 g_1$$

and by Theorem 14.1, $f_1 g_1$ is primitive. Hence, cont fg = cont $f \cdot$ cont g. □

▷ **Quick Exercise.** Check this corollary by multiplying together two polynomials from $\mathbb{Z}[x]$ of your choice. ◁

To achieve our goal of proving that $\mathbb{Z}[x]$ is a UFD we must first prove that every element of $\mathbb{Z}[x]$ either is irreducible or can be factored into a product of irreducibles. As promised, this is easy to prove, using induction on degree.

Theorem 14.3 *Every non-zero non-unit of $\mathbb{Z}[x]$ is either irreducible or a product of irreducibles.*

Proof: Suppose that $0 \neq f \in \mathbb{Z}[x]$ is a non-unit. We proceed by induction on $\deg(f)$. If $\deg(f) = 0$, then $f \in \mathbb{Z}$. Because $\mathbb{Z}$ is a UFD, f is either irreducible in $\mathbb{Z}$, or a product of irreducibles in $\mathbb{Z}$; but irreducibles in $\mathbb{Z}$ are irreducible in $\mathbb{Z}[x]$, and so we have the required result.

Now suppose that $\deg(f) = n > 0$. If f itself is irreducible, we are done. Otherwise, we first factor out of f the element cont f; this is an element of $\mathbb{Z}$ that we can factor into irreducibles of $\mathbb{Z}$ (and hence of $\mathbb{Z}[x]$). So we may suppose that f is primitive. If f isn't irreducible, we can then factor it as $f = gh$. Because f is primitive, we must have that $\deg(g) > 0$ and $\deg(h) > 0$. So by the induction hypothesis, g and h are irreducible or the product of irreducibles, and therefore so is f. By the Principle of Mathematical Induction, this proves the theorem. □

14.4 Irreducibles in $\mathbb{Z}[x]$ are Prime

So, what remains to be done to prove that $\mathbb{Z}[x]$ is a UFD? Theorem 13.2 asserts that a domain in which factorization into irreducibles is possible has unique factorization if and only if all irreducible elements are prime (remember, of course, that prime elements are *always* irreducible). We thus must show that all irreducible elements of $\mathbb{Z}[x]$ are actually prime. We will do this by relating factorizations of polynomials in $\mathbb{Z}[x]$ to factorizations in $\mathbb{Q}[x]$, using Gauss's Lemma 5.5.

Theorem 14.4 *In $\mathbb{Z}[x]$, all irreducible elements are prime; consequently $\mathbb{Z}[x]$ is a UFD.*

Proof: Suppose that f is an irreducible polynomial in $\mathbb{Z}[x]$. Note first that because f has no non-trivial factors from $\mathbb{Z}$, this means that f is primitive. By Gauss's Lemma, f must be irreducible in $\mathbb{Q}[x]$ as well. Because $\mathbb{Q}[x]$ is a PID, f is then a prime element of $\mathbb{Q}[x]$; we must show that f is a prime element of $\mathbb{Z}[x]$.

To show this, suppose that f divides gh, where $g, h \in \mathbb{Z}[x]$. Now consider g and h as elements of $\mathbb{Q}[x]$. Because f *is* prime in $\mathbb{Q}[x]$, this means that f must divide one of them (say, g) in $\mathbb{Q}[x]$. That is, $g = fg_1$, where $g_1 \in \mathbb{Q}[x]$. But by Gauss's Lemma 5.5 there exist rational numbers A and B, so that $AB = 1$ and $Af, Bg_1 \in \mathbb{Z}[x]$. But because f is primitive and $Af \in \mathbb{Z}[x]$, A must be an integer; otherwise, the denominator of A would be cancelled by a non-trivial factor of cont f. Thus, $g = f(ABg_1)$ is a factorization in $\mathbb{Z}[x]$, and so f divides g in $\mathbb{Z}[x]$. Hence, f is prime in $\mathbb{Z}[x]$, as required. It then follows immediately that $\mathbb{Z}[x]$ is a UFD. □

This means that $\mathbb{Z}[x]$ is an example of a UFD that is not a PID, thus showing that the converse of Theorem 13.5 is false.

The method of this chapter can actually be generalized to prove that anytime D is a PID, then $D[x]$ is a UFD that is not a PID. The general proof, while quite similar, requires a bit more machinery than we presently have available to us, and so we will not pursue it here.

Chapter Summary

In this chapter, we used Gauss's Lemma to prove that in $\mathbb{Z}[x]$, irreducible elements are prime, and so $\mathbb{Z}[x]$ is a UFD.

Warm-up Exercises

a. What properties does $\mathbb{Z}[x]$ lack that prevent us from proving that it is a UFD just as we did for $\mathbb{Q}[x]$?

b. Determine the content of

$$30x^4 - 12x^2 + 42x - 54 \text{ and } 49x^3 + 70x^2 - 14.$$

What is the content of the product of these two polynomials?

c. Why would it be silly to try to talk about the content of polynomials from $\mathbb{Q}[x]$?

d. Factor the following polynomials completely into irreducibles in $\mathbb{Z}[x]$; do they have the same irreducibles as factors in $\mathbb{Q}[x]$?

(a) $6x^3 - 6$.

(b) $3x^4 - 6x$.

(c) $5x^4 - x^3 - 15x^2 - 7x + 2$.

e. Give examples of the following (or say why they don't exist):

(a) A PID that isn't a UFD.

(b) A UFD that isn't a PID.

f. Give examples of the following polynomials from $\mathbb{Z}[x]$ (or say why they don't exist):

(a) An irreducible polynomial of degree 4.

(b) An irreducible polynomial that isn't prime.

(c) An irreducible polynomial of degree 0.

(d) A unit.

(e) A polynomial that can't be non-trivially factored in $\mathbb{Z}[x]$, but can be non-trivially factored in $\mathbb{Q}[x]$.

(f) A polynomial that can't be non-trivially factored in $\mathbb{Z}[x]$, but can be non-trivially factored in $\mathbb{R}[x]$.

(g) A polynomial that can't be non-trivially factored in $\mathbb{Q}[x]$, but can be non-trivially factored in $\mathbb{Z}[x]$.

Exercises

1. Suppose that $f \in \mathbb{Z}[x]$ and the coefficient on its highest power of x is 1 (we say that such a polynomial is called **monic**).

   (a) Why is a monic polynomial necessarily primitive?

   (b) Give an example of a primitive polynomial in $\mathbb{Z}[x]$ that isn't monic.

2. Suppose that $f \in \mathbb{Z}[x]$ is monic (see Exercise 1).

   (a) Prove that all rational roots of f are integers.

   (b) Prove that all integer roots of f divide its constant term.

   (c) Give examples of primitive polynomials for which parts a and b fail.

3. Use Exercise 2 and the Root Theorem 4.3 to show that x^3+2x+7 is prime in $\mathbb{Z}[x]$.

4. Consider the ring $\mathbb{Z}[i][x]$ of polynomials with coefficients from the Gaussian integers $\mathbb{Z}[i]$. Argue that this ring is not a PID, by an argument analogous to that for $\mathbb{Z}$. (The argument in the text that $\mathbb{Z}[x]$ is a UFD can actually be generalized to this case too, but we will not do this here.)

5. Prove that in $\mathbb{Z}[x]$, $\langle 3x + 1, x + 1\rangle = \langle 2, x - 1\rangle$. Show that this ideal is not principal.

Chapter 15

Euclidean Domains

The difficulty we encountered in the last chapter in proving that $\mathbb{Z}[x]$ is a UFD might convince you of the advantage of having a Division Theorem available: The Division Theorem makes proving that $\mathbb{Z}$ and $\mathbb{Q}[x]$ are UFDs relatively easy. It seems natural then to define a more general class of domains (including both $\mathbb{Z}$ and $\mathbb{Q}[x]$) that *have* a Division Theorem. We name this class of domains in honor of Euclid, in whose *Elements* we find the first reference to that corollary of the Division Theorem, Euclid's Algorithm. This is a typical gambit of mathematicians. We have identified an important tool we'd like to study in general (in this case, a Division Theorem for domains), and so we isolate those domains having this tool by means of a definition.

15.1 Euclidean Domains

A **Euclidean domain** is a domain D that can be equipped with a function $v : D\backslash\{0\} \to \mathbb{N}$ that satisfies the following two criteria:

a. For $a, b \in D$ with $a \neq 0$, there exist $q, r \in D$ (called the **quotient** and **remainder**, respectively) such that $b = aq + r$, with $r = 0$ or $v(r) < v(a)$.

b. For all $a, b \neq 0$, $v(a)v(b) = v(ab)$.

The function v is called a **Euclidean valuation** for D.

You should think of the function v as a measure of 'size' of the elements. Thus, the first condition says that we can always divide an element of D by a non-zero element; either the division is exact, or we have a remainder 'smaller' than the divisor. To be consistent with our earlier terminology, we will call this the **Division Theorem** for Euclidean domains (even though we have built it right into the definition). The second condition allows us to relate divisibility in D to divisibility in $\mathbb{Z}$ (which we presumably know more about). Note that because the function v takes on its values in $\mathbb{N}$, we have available to us all we know about this set, including such tools as the Well-ordering Principle and Mathematical Induction.

Before making any abstract inferences about Euclidean domains, we shall first list a number of examples, showing that this concept is in fact a common generalization of a number of the domains we have already considered.

Example 15.1

$\mathbb{Z}$ is a Euclidean domain. Here, the valuation (or 'size' function) is obvious; just let $v(n) = |n|$. The first condition is then the Division Theorem of $\mathbb{Z}$ (Theorem 2.1), and the second condition holds because $|nm| = |n||m|$.

Example 15.2

$\mathbb{Q}[x]$ is a Euclidean domain. How did we measure 'size' of elements for polynomials? We just used their degree; thus, $v(f) = \deg(f)$ seems a natural definition; condition (1) is now just the Division Theorem for $\mathbb{Q}[x]$ (Theorem 4.2). However, the second condition fails, because after all, $\deg(fg) = \deg(f) + \deg(g)$ (rather than $\deg(f)\deg(g)$). We can escape this problem by means of a trick: Just let $v(f) = 2^{\deg(f)}$. Because exponentiation turns addition into multiplication, condition (2) is now satisfied. But what about condition (1)? This still works because $\deg(f) < \deg(g)$ exactly when $v(f) < v(g)$.

Example 15.3

Any field is a Euclidean domain. In a field we can divide any element by any non-zero element, with no remainder. Thus, all non-zero elements should be 'small'. A function that works easily is just $v(a) = 1$, for all $a \neq 0$. This seems almost too easy, but, after all, we are only really interested in a Division Theorem where we have at least some elements that do *not* divide one another exactly.

15.2 The Gaussian Integers

We will now show that the ring of Gaussian integers $\mathbb{Z}[i]$ is a Euclidean domain. In Chapter 10 we had already measured the 'size' of Gaussian integers, by means of the norm function $N(a + bi) = a^2 + b^2$. We wish to show that the norm function serves as a Euclidean valuation. Note that the second condition holds because $N(\alpha\beta) = N(\alpha)N(\beta)$ (Theorem 10.1). Unfortunately, the Division Theorem is not really obvious. What should the quotient and remainder be if we divide one Gaussian integer by another?

Example 15.4

Let's look at a particular example: Consider $11 + 6i$ divided by $2 + 3i$. Now we can certainly perform this division in the field $\mathbb{C}$. We obtain the following:

$$\frac{11+6i}{2+3i} = \frac{(11+6i)(2-3i)}{(2+3i)(2-3i)} = \frac{40-21i}{13} = \frac{40}{13} - \frac{21}{13}i.$$

What is the Gaussian integer closest to this quotient? Because $\frac{40}{13} = 3\frac{1}{13}$ and $\frac{21}{13} = 1\frac{8}{13}$, the answer to this question seems to be $3 - 2i$. If this is to be our quotient q, then the remainder r must be given by

$$(11 + 6i) - (2 + 3i)(3 - 2i) = -1 + i.$$

Note that $v(-1 + i) = 2 < 13 = v(2 + 3i)$.

This example makes plausible the assertion that the norm function is in fact a Euclidean valuation for the Gaussian integers. Let us now prove this carefully:

Theorem 15.1 *$\mathbb{Z}[i]$ is a Euclidean domain.*

Proof: As observed above, we clearly need check only that the Division Theorem (that is, condition (1) above) really works. So suppose that $\beta = m + ni$ and $\alpha = r + si$ are Gaussian integers, and we wish to verify the Division Theorem for them, where $\alpha \neq 0$ will serve as the divisor. Using the example above as our model, consider the following computation in $\mathbb{C}$:

$$\frac{\beta}{\alpha} = \frac{m+ni}{r+si} = \frac{mr+ns}{r^2+s^2} + \frac{nr-ms}{r^2+s^2}i.$$

The real and imaginary parts of this complex number certainly need not be integers, but we shall now choose integers q_1 and q_2 as close as possible to them. Any real number is within $\frac{1}{2}$ of an integer, and so this amounts to saying that

$$\left|\frac{mr+ns}{r^2+s^2} - q_1\right| \leq \frac{1}{2}$$

and

$$\left|\frac{nr-ms}{r^2+s^2} - q_2\right| \leq \frac{1}{2}.$$

We will use $\gamma = q_1 + q_2 i$ as the proposed quotient in the Division Theorem. The remainder ρ will then necessarily be given by

$$\rho = \beta - \alpha\gamma = (m - rq_1 + sq_2) + (n - rq_2 - sq_1)i.$$

It remains to check that the valuation of ρ is less than the valuation of α (or else $\rho = 0$). To verify this we use the fact that the function N preserves multiplication even in $\mathbb{C}$ (Theorem 8.3). We can thus do the following computation:

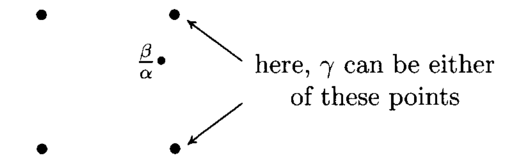
$$\begin{aligned}
v(\rho) = N(\rho) = N(\beta - \alpha\gamma) &= N\left(\alpha \cdot \left(\frac{\beta}{\alpha} - \gamma\right)\right) \\
&= N(\alpha)N\left(\frac{\beta}{\alpha} - \gamma\right) \\
&= N(\alpha)\left(\left(\frac{mr+ns}{r^2+s^2} - q_1\right)^2 + \left(\frac{nr-ms}{r^2+s^2} - q_2\right)^2\right) \\
&\leq N(\alpha)\left(\frac{1}{4} + \frac{1}{4}\right) < N(\alpha) = v(\alpha).
\end{aligned}$$

This shows that the remainder ρ is indeed suitably 'small'. □

You will use similar means to prove that $\mathbb{Z}[\sqrt{2}]$ is a Euclidean domain in Exercise 15.1. Not all such domains are; it is a deep and not completely solved problem of number theory to distinguish which are Euclidean domains, and which aren't.

The proof of Theorem 15.1 can be interpreted geometrically. The crucial step is the choice of the quotient γ; we do this by finding a point in the complex plane with integer coordinates, whose distance to β/α (the quotient in $\mathbb{C}$) is less than 1. If the complex number β/α happened to fall between points with consecutive integer coordinates,

as illustrated in the diagram below, we would actually have two possible choices for the quotient:



Thus, the quotient and remainder *need not be unique* in a Euclidean domain; the careful reader might have noted earlier the absence of the word 'unique' in the original definition. We will illustrate this point algebraically, in the following example, where we actually have four choices for the quotient available:

Example 15.5

Let's divide $3+5i$ by 2. Because the quotient in $\mathbb{C}$ is $3/2+(5/2)i$, which element of $\mathbb{Z}[i]$ should we choose as the quotient? In this case, each of the four nearby points with integer coordinates are at distance less than 1 from $3/2 + (5/2)i$, and so we have four different quotient/remainder pairs:

$$\begin{aligned}
3 + 5i &= (1 + 2i)(2) + (1 + i) \\
&= (1 + 3i)(2) + (1 - i) \\
&= (2 + 2i)(2) + (-1 + i) \\
&= (2 + 3i)(2) + (-1 - i).
\end{aligned}$$

15.3 Euclidean Domains are PIDs

We are now ready to make some abstract observations about Euclidean domains. We first prove that the units of a Euclidean domain are identifiable by means of its valuation:

Theorem 15.2 *Let D be a Euclidean domain with valuation v. Then u is a unit of D if and only if $v(u) = 1$.*

Proof: Notice first that because $1^2 = 1$, $v(1)^2 = v(1)$; the only positive integer with this property is 1, and so $v(1) = 1$. But then if u is a unit, $v(u)v(u^{-1}) = v(uu^{-1}) = v(1) = 1$, and so $v(u) = 1$.

Conversely, if $v(u) = 1$, apply the Division Theorem to 1 and u to obtain $1 = qu + r$. But $v(r) < v(u) = 1$, which is impossible, and so $r = 0$. That is, u is a unit with inverse q. □

We next show that all Euclidean domains are PIDs, using very much the same sort of proof as we used for $\mathbb{Z}$ and $\mathbb{Q}[x]$.

Theorem 15.3 *Each Euclidean domain is a PID.*

Proof: Suppose that I is a proper ideal of the Euclidean domain D. What element of I might serve as its generator? We answer this question (as we did for $\mathbb{Z}$) by using the Well-ordering Principle. Choose an element d of I with smallest valuation; there may be many choices for d. Note that this valuation is necessarily larger than 1 because I contains no units. We claim that $\langle d \rangle = I$. Because I is an ideal, it is clear that $\langle d \rangle \subseteq I$. Suppose now that $b \in I$. We claim b is a multiple of d. To check this, we should clearly use the Division Theorem to obtain a quotient q and remainder r: $b = qd + r$. Because $r = b - qd$, this means that $r \in I$. But $v(r) < v(d)$, which is impossible unless $r = 0$. Thus, $b = qd \in \langle d \rangle$. □

Because every PID is a UFD, we have:

Corollary 15.4 *Each Euclidean domain is a UFD.*

Example 15.6

We thus know that the ring $\mathbb{Z}[i]$ of Gaussian integers is a PID. The crucial idea in the proof above is that a generator for an ideal is an element with smallest valuation. As an example of this procedure in $\mathbb{Z}[i]$, consider the set

$$I = \{a + bi \in \mathbb{Z}[i] : 5|(2a - b)\}.$$

We claim that I is an ideal. It is quite easy to show that I is closed under subtraction.

▷ **Quick Exercise.** Check this. ◁

Suppose now that $a + bi \in I$ and $c + di \in \mathbb{Z}[i]$. Then

$$(a + bi)(c + di) = (ac - bd) + i(bc + ad)$$

and so we require that $2ac - 2bd - bc - ad$ is an integer divisible by 5. But

$$\begin{aligned} &2ac - 2bd - bc - ad = \\ &c(2a - b) - d(2b + a) = c(2a - b) - d(5a - 2(2a - b)), \end{aligned}$$

and this is divisible by 5 because $2a - b$ is.

The theorem asserts that I is a principal ideal, and the proof tells that we can find a generator by picking an element of I with smallest valuation. To find out what this is, suppose that $a + bi \in I$; then $2a - b = 5k$, for $k \in \mathbb{Z}$. So

$$v(a + bi) = a^2 + b^2 = a^2 + (2a - 5k)^2 = 5\left(a^2 - 4ak + 5k^2\right),$$

and so this smallest valuation is divisible by 5. But note that $2 - i$ has valuation 5 and $2 - i \in I$; consequently, we must have that $\langle 2 - i \rangle = I$. Let's do some arithmetic in $\mathbb{C}$, to see explicitly that every element of I is a multiple of $2 - i$. For that purpose, let $a + bi \in I$; then

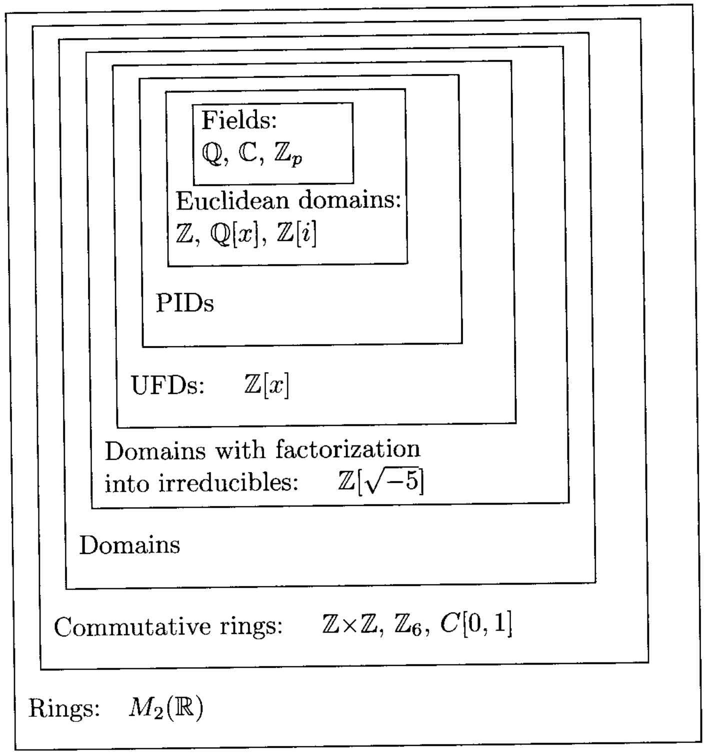
$$\begin{aligned} \frac{a + bi}{2 - i} = \frac{(a + bi)(2 + i)}{(2 - i)(2 + i)} &= \frac{1}{5}((2a - b) + (a + 2b)i) \\ &= \frac{1}{5}((2a - b) + (3(2a - b) - 5(a - b))i) \end{aligned}$$

and this latter element is in $\mathbb{Z}[i]$, because the real and imaginary parts of the numerator are divisible by 5. Thus, any element of I is a multiple of $2 - i$, as claimed. Note that in practice it might be quite difficult to find the generator for an ideal by this method.

15.4 Some PIDs are not Euclidean

A natural question to ask at this point is: Are there PIDs that are not Euclidean? The answer is yes, but it is difficult to prove this. It is usually quite easy to show that a domain is not a PID: Just exhibit a particular ideal and prove it is not principal. However, to show that a domain D is not Euclidean, one must prove that *all* functions $v : D \backslash \{0\} \to \mathbb{N}$ fail to satisfy at least one of the two defining conditions. The difference between these two tasks is that the definition of PID is a *universal* statement—*all* ideals are principal—while the definition of a Euclidean domain is an *existential* statement—*there exists* a function satisfying particular properties. And the negation of a universal statement is existential, while the negation of an existential statement is universal.

The picture below provides a nice summary of the relationships between the various sorts of commutative rings we have studied in the last several chapters. All inclusions shown are proper, whether or not we have been able to provide examples in this book.


Fields:
$\mathbb{Q}$, $\mathbb{C}$, $\mathbb{Z}_p$
Euclidean domains:
$\mathbb{Z}$, $\mathbb{Q}[x]$, $\mathbb{Z}[i]$
PIDs
UFDs: $\mathbb{Z}[x]$
Domains with factorization into irreducibles: $\mathbb{Z}[\sqrt{-5}]$
Domains
Commutative rings: $\mathbb{Z}\times\mathbb{Z}$, $\mathbb{Z}_6$, $C[0,1]$
Rings: $M_2(\mathbb{R})$

Chapter Summary

In this chapter we defined the concept of *Euclidean domain* and proved that all Euclidean domains are PIDs. We showed that $\mathbb{Z}[i]$ is a Euclidean domain, in addition to $\mathbb{Z}$ and $\mathbb{Q}[x]$.

Warm-up Exercises

a. Compute a quotient and remainder for the following pairs of elements, in the given Euclidean domain.

   (a) 116 divided by 7 in $\mathbb{Z}$.

   (b) $x^4 - 2x^3 + 5x - 3$ divided by $2x^2 - 1$ in $\mathbb{Q}[x]$.

   (c) $13 + 5i$ divided by $3 + 2i$ in $\mathbb{Z}[i]$.

   (d) 7 divided by π in $\mathbb{R}$.

b. Given an ideal I in a Euclidean domain, we know that I is principal. Describe conceptually how to determine the generator of I. (This description is very simple; it might not be so simple to carry out in practice.)

Exercises

1. Using a similar argument to that in the text for $\mathbb{Z}[i]$, prove that $\mathbb{Z}[\sqrt{2}]$ is a Euclidean domain, using the function N as the valuation.

2. Compute a quotient and remainder for $17 + 32\sqrt{2}$ divided by $3 - 4\sqrt{2}$ in the Euclidean domain $\mathbb{Z}[\sqrt{2}]$. Check that your remainder has a smaller valuation than $3 - 4\sqrt{2}$.

3. Show by example that quotients and remainders are not unique in the Euclidean domain $\mathbb{Z}[\sqrt{2}]$.

4. Make a definition for *greatest common divisor* of two elements in a Euclidean domain.

5. Suppose that $a, b \in D$, and D is a Euclidean domain. Consider the ideal $\langle a, b\rangle = \{ax + by : x, y \in D\}$; see Exercise 12.1. Let d be an element of $\langle a, b\rangle$ with least valuation.

   (a) Why does the element d exist?

   (b) Prove that d is a gcd of a, b. (See Exercise 4.)

   (c) Why does this mean that Euclidean domains possess a GCD identity?

(d) Rephrase the GCD identity for Euclidean domains in terms of ideals of the form $\langle a, b\rangle$.

6. (a) Explain why Euclid's Algorithm for finding a gcd makes sense in a Euclidean domain.

   (b) Apply Euclid's Algorithm to $5 + 133i$ and $17 + 34i$ in $\mathbb{Z}[i]$.

   (c) Backtrack through Euclid's Algorithm to show explicitly that $\langle 5 + 133i, 17 + 34i\rangle$ is a principal ideal.

7. Suppose that a and b are integers. Prove that their gcd in $\mathbb{Z}[i]$ is the same as in $\mathbb{Z}$, up to unit multiples in $\mathbb{Z}[i]$.

8. Prove that if D is a Euclidean domain and $a, b \in D$ are associates, then $v(a) = v(b)$. (This is easy!) Show by example that the converse is false in $\mathbb{Z}[\sqrt{2}]$. (You should compare this exercise to Exercise 10.14, where two non-associates of the same norm are exhibited in $\mathbb{Z}[\sqrt{-5}]$; of course, the latter ring is not a Euclidean domain.)

9. Provide a geometric interpretation of the proof that $\mathbb{Z}[\sqrt{2}]$ is a Euclidean domain, analogous to the interpretation we discussed for $\mathbb{Z}[i]$.

Section III in a Nutshell

This section examines conditions that $\mathbb{Z}$ and $\mathbb{Q}[x]$ share, which provide them with a unique factorization theorem into irreducibles, like the Fundamental Theorem of Arithmetic.

To accomplish this goal, the section first considers analogues for the concepts of irreducible, prime and associate. The ultimately insignificant distinction between two elements in a ring that are associates of one another leads to the definition of a *principal ideal* in a commutative ring R: $\langle a\rangle = \{b \in R : b = ac,\ c \in R\}$. Statements about elements can be efficiently and elegantly translated into statements about principal ideals, especially for integral domains D (Theorem 13.3): $\langle a\rangle = D$ if and only if a is a unit; $\langle a\rangle \subseteq \langle b\rangle$ if and only a divides b; a is irreducible if and only if $\langle a\rangle$ is maximal among all principal ideals.

The idea of principal ideal can be generalized to that of *ideal*, which is a subring I of a commutative ring R satisfying the *multiplicative absorption property*: if $a \in I$ and $r \in R$, then $ar \in I$.

For $\mathbb{Z}$ and $\mathbb{Q}[x]$, all ideals are principal (Theorem 11.6). An integral domain where this holds is called a *principal ideal domain* (or PID). For a PID, we have factorization into irreducibles (Theorem 12.1). To prove uniqueness of such factorization requires precisely that the concepts of irreducibility and primeness coincide (Theorems 13.4 and 13.5), as they do for the integers. An integral domain that has unique factorization into irreducibles is called a *unique factorization domains* (or UFD).

Integral domains of the form $\mathbb{Z}[\sqrt{n}] = \{a + b\sqrt{n} : a, b \in \mathbb{Z}\}$ are called *quadratic extensions* of the integers. These are important examples; some such rings are PIDs, but not all. The ring $\mathbb{Z}[\sqrt{-5}]$ is not a PID and is not even a UFD.

The ring $\mathbb{Z}[x]$ of polynomials with integer coefficients is not a PID, but it is a UFD; this follows essentially from Gauss's Lemma.

Some quadratic extensions (such as the *Gaussian integers* $\mathbb{Z}[\sqrt{-1}]$) share even more properties in common with $\mathbb{Z}$ and $\mathbb{Q}[x]$ and are called Euclidean domains: elements in such domains have a notion of 'size', which equips them with a Division algorithm. This makes it easy to prove (Theorem 15.3) that they are PIDs, and hence UFDs.

IV

Ring Homomorphisms and Ideals

Chapter 16

Ring Homomorphisms

Up to now we have examined the relationship between rings by looking at properties they might have in common. Some pairs of rings can actually be placed into a rather closer relationship by means of a function between them. The most important example of this idea is the relationship between $\mathbb{Z}$ and $\mathbb{Z}_n$, as given by the residue function $\varphi : \mathbb{Z} \to \mathbb{Z}_n$ defined by $\varphi(m) = [m]_n$. Another example is the evaluation function $\psi : \mathbb{Q}[x] \to \mathbb{Q}$ defined by $\psi(f) = f(a)$ (where a is some fixed rational number). Exploring the general context of these examples will then give us a new tool with which to study rings.

Because we deal extensively with functions in this chapter, we should remind you of some terminology regarding them. If R and S are sets and $\varphi : R \to S$ is a function, we call R the **domain** of the function and S its **range**. Please note that this is a dramatically different use of the word domain than we have already encountered (where domain is really a short version of integral domain), but standard usage forces this ambiguity upon us; we hope you will be able to tell from context which meaning is intended.

16.1 Homomorphisms

Now consider the functions φ and ψ defined above. They certainly have rings as their ranges and domains; however, they also carry significant parts of the structure of the domain ring over to the range ring. We make this precise in the following definition: Let R and S be rings and $\varphi : R \to S$ a function such that

$$\varphi(a + b) = \varphi(a) + \varphi(b)$$

and

$$\varphi(ab) = \varphi(a)\varphi(b),$$

for all $a, b \in R$. We call this a **ring homomorphism**. Speaking more colloquially, we say that φ *preserves addition* and φ *preserves multiplication.*

Later in the book we will consider homomorphisms between other algebraic structures than rings and will then have to be careful to refer to *ring* homomorphisms, as opposed to others we might discuss; but for now the term homomorphism will suffice. The word homomorphism comes from Greek roots meaning 'same shape'. We shall see as we explore this concept how appropriate this term is.

It is time now to look at some examples of ring homomorphisms. We will begin with specific examples of the functions with which we started the chapter.

Example 16.1

Consider the residue function $\varphi : \mathbb{Z} \to \mathbb{Z}_4$ defined by $\varphi(m) = [m]_4$. Thus, for example, $\varphi(7) = [3]$. Is this function a homomorphism? That is, does it preserve addition and multiplication? Let's check addition:

$$\varphi(a + b) = [a + b] = [a] + [b] = \varphi(a) + \varphi(b).$$

The crucial step here is the second equal sign; it holds because of the way we defined addition in $\mathbb{Z}_4$, back in Chapter 3. The proof for multiplication works just the same. We can view the homomorphism φ as a somewhat more abstract and precise way of conveying the fact that we already know: The operations on $\mathbb{Z}_4$ are related to those on $\mathbb{Z}$. Notice of course that there's nothing special about 4: The residue function is a homomorphism, for any modulus.

Example 16.2

Consider the evaluation function $\varphi : \mathbb{Q}[x] \to \mathbb{Q}$ defined by $\varphi(f) = f(2)$: evaluation of polynomials at 2. For example, if $f = x^2 - 3x + 1$, then $\varphi(f) = 4 - 6 + 1 = -1$. This is also a homomorphism:

$$\varphi(f + g) = (f + g)(2) = f(2) + g(2) = \varphi(f) + \varphi(g).$$

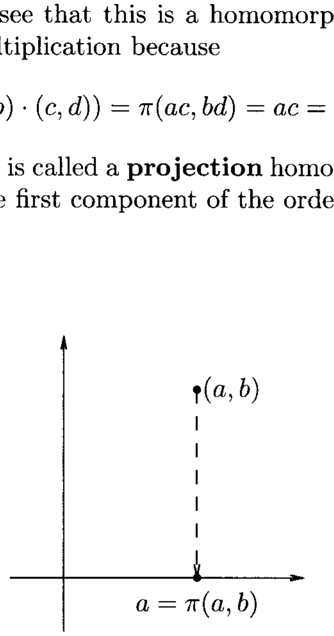
Once again, the crucial step is the second equal sign. It holds because of the way we defined the addition of polynomials in Chapter 4. Once again, the proof for multiplication is just the same. Notice of course that there's nothing special about 2: The evaluation function is a homomorphism, for any element of $\mathbb{Q}$.

Example 16.3

Consider the function $\pi : \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$ defined by $\pi(a, b) = a$. It is easy to see that this is a homomorphism. For example, it preserves multiplication because

$$\pi((a, b) \cdot (c, d)) = \pi(ac, bd) = ac = \pi(a, b) \cdot \pi(c, d).$$

This function is called a **projection** homomorphism (we are projecting on the first component of the ordered pair). See the diagram below.



Example 16.4

Consider the function $\varphi : \mathbb{C} \to \mathbb{C}$ defined by $\varphi(\alpha) = \bar{\alpha}$, which takes a complex number to its complex conjugate. For example, $\varphi(3 - i) = 3 + i$. This preserves addition:

$$\begin{aligned} \varphi((a + bi) + (c + di)) &= \varphi((a + c) + (b + d)i) = \\ (a + c) - (b + d)i &= (a - bi) + (c - di) = \\ &\quad \varphi(a + bi) + \varphi(c + di). \end{aligned}$$

It also preserves multiplication:

$$\begin{aligned} \varphi((a + bi)(c + di)) &= \varphi((ac - bd) + (ad + bc)i) = \\ (ac - bd) - (ad + bc)i &= (a - bi)(c - di) = \\ &\quad \varphi(a + bi)\varphi(c + di). \end{aligned}$$

Of course, we can define many functions whose ranges and domains are rings that are not homomorphisms; here is one example:

Example 16.5

Consider the function $\rho : \mathbb{Z} \to \mathbb{Z}$ defined by $\rho(n) = 3n$; this function merely multiplies by 3. Note that $\rho(nm) = 3nm$, while $\rho(n)\rho(m) = 9nm$, and so this function certainly does not preserve multiplication; it is consequently not a ring homomorphism.

▷ **Quick Exercise.** Does this function preserve addition? ◁

16.2 One-to-one and Onto Functions

We now must remind you of two very important properties descriptive of functions (whether homomorphisms or not); namely, *onto* and *one-to-one*. In order to review the definitions of these properties, let's suppose that X and Y are sets, and $\varphi : X \to Y$ is a function.

It may well be the case that Y contains elements that do not occur as $\varphi(x)$ for any $x \in X$. Writing $\{\varphi(x) : x \in X\}$ as $\varphi(X)$, we are saying that it may be the case that $\varphi(X) \subset Y$. If $\varphi(X) = Y$, we say that φ is an **onto** function. (An alternate term in common use is to call φ a **surjective** function in this case.)

For each element y of $\varphi(X)$, there exists an element x of X such that $\varphi(x) = y$ (this is just the definition of $\varphi(X)$). If this element is unique for each such $y \in \varphi(X)$, we say that φ is a **one-to-one** function. Another way of saying this is to assert that φ never takes two distinct elements of X to the same element of Y. (An alternate term in common use is to call φ an **injective** function in this case.)

For numerous examples of these concepts in action, see the examples of homomorphisms below.

▷ **Quick Exercise.** Which of the functions in Examples 16.1 to 16.5 are one-to-one? Which are onto? ◁

16.3 Properties Preserved by Homomorphisms

Although we have in our definition required only that a homomorphism preserve the ring operations addition and multiplication, a homomorphism actually preserves much more of the ring structure. We catalog such properties of a homomorphism in the following theorem:

Theorem 16.1 *Let $\varphi : R \to S$ be a homomorphism between the rings R and S. Let 0_R and 0_S be the additive identities of R and S, respectively.*

a. $\varphi(0_R) = 0_S$.

b. $\varphi(-a) = -\varphi(a)$, *and so* $\varphi(a - b) = \varphi(a) - \varphi(b)$.

c. If R has unity 1_R and φ is onto S, and S is not the zero ring, then $\varphi(1_R)$ is unity for S.

d. If a is a unit of R and φ is onto S, and S is not the zero ring, then $\varphi(a)$ is a unit of S. In this case, $\varphi(a)^{-1} = \varphi(a^{-1})$.

In other words, a homomorphism always preserves the zero of the ring and the additive inverses. Furthermore, a homomorphism preserves unity and units *if the homomorphism is onto.* This theorem thus asserts that the assumption that a function preserve addition and multiplication is enough to conclude that it preserve other additive and multiplicative structures. Note that a ring always has an additive identity and additive inverses, and so (a) and (b) are phrased to reflect that fact. On the other hand, a ring need not have unity or units, and so (c) and (d) are phrased rather differently.

Proof: (a): Now $\varphi(0_R) + \varphi(0_R) = \varphi(0_R + 0_R) = \varphi(0_R)$. By subtracting $\varphi(0_R)$ from both sides, we obtain $\varphi(0_R) = 0_S$.

(b): Now $\varphi(-a) + \varphi(a) = \varphi(-a + a) = \varphi(0_R) = 0_S$; thus by the definition and uniqueness of $-\varphi(a)$, we have that $\varphi(-a) = -\varphi(a)$. But then

$$\varphi(a - b) = \varphi(a + (-b)) = \varphi(a) + \varphi(-b) = \varphi(a) - \varphi(b),$$

as claimed.

(c): If S is the zero ring, we have previously decreed that we won't call its unique element 0 unity; we've excluded this case in our hypothesis.

Let s be any element of S; then there exists $r \in R$ with $\varphi(r) = s$, because φ is onto. Then

$$s\varphi(1_R) = \varphi(r)\varphi(1_R) = \varphi(r \cdot 1_R) = \varphi(r) = s,$$

and so $\varphi(1_R)$ must be the unity of S.

(d): Suppose that $aa^{-1} = 1_R$. Then

$$\varphi(a)\varphi(a^{-1}) = \varphi(aa^{-1}) = \varphi(1_R) = 1_S,$$

and so $\varphi(a)$ is a unit of S with inverse $\varphi(a^{-1})$. □

16.4 More Examples

We now look at some further examples of homomorphisms, which will illustrate both the theorem and the notions of one-to-one and onto. In each case you should check that the example given does preserve multiplication and addition. The first three examples seem rather trivial but are important.

Example 16.6

Let R be a ring with unity 1, and S another ring. Define $\zeta : R \to S$ by $\zeta(r) = 0$, for all $r \in R$ (this is called the **zero homomorphism**). If S has unity, note that $\zeta(1) = 0 \neq 1$, and in this case the function is not onto, and so (c) of Theorem 16.1 does not apply. Notice that this homomorphism is (notoriously) not one-to-one.

Example 16.7

Let R be a ring and define $\iota : R \to R$ by $\iota(r) = r$ (this is called the **identity homomorphism**). It of course preserves *all* structure of the ring R; it is both one-to-one and onto.

Example 16.8

Let R be a subring of the ring S. Consider the function $\iota : R \to S$ defined by $\iota(r) = r$ (this is called the **inclusion homomorphism**). It is always one-to-one, but not onto if $R \subset S$.

Example 16.9

Reconsider the residue map $\varphi : \mathbb{Z} \to \mathbb{Z}_4$. Notice that although 2 is not a zero divisor in $\mathbb{Z}$, $\varphi(2)$ *is* a zero divisor in $\mathbb{Z}_4$ because $[2][2] = [4] = [0]$. Another thing to notice about this function is that although the domain is infinite, the range is finite; although onto, it is not one-to-one.

Example 16.10

Consider the rings $\mathbb{Q}$ and $\mathbb{Q} \times \mathbb{Q}$ (recall from Example 6.10 how the direct product $\mathbb{Q} \times \mathbb{Q}$ is made into a ring). Define the function $\varphi : \mathbb{Q} \to \mathbb{Q} \times \mathbb{Q}$ by setting $\varphi(r) = (r, 0)$. This homomorphism is one-to-one, but not onto. Now 3 is a unit in $\mathbb{Q}$ because $3 \cdot \frac{1}{3} = 1$. But the unity of $\mathbb{Q} \times \mathbb{Q}$ is $(1, 1)$, and so $\varphi(3) = (3, 0)$ cannot be a unit in $\mathbb{Q} \times \mathbb{Q}$, because there is no (z, w) such that $(3, 0)(z, w) = (1, 1)$. In fact, $\varphi(3)$ (and $\varphi(1)$) are zero divisors.

Example 16.11

Reconsider the evaluation map $\varphi : \mathbb{Q}[x] \to \mathbb{Q}$ given by $\varphi(f) = f(2)$, which we considered in Example 16.2. This homomorphism is onto, but not one-to-one. Notice that the polynomial x in $\mathbb{Q}[x]$ is not a unit, but $\varphi(x) = 2$ is a unit in $\mathbb{Q}$.

Example 16.12

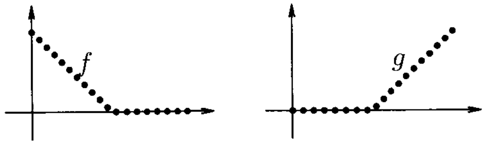
Let $C[0, 1]$ be the ring of continuous real-valued functions with domains $[0, 1]$ (we discussed this ring in Example 6.14). Consider the evaluation map $\psi : C[0, 1] \to \mathbb{R}$ defined by $\psi(f) = f(\frac{1}{4})$. Define the elements f, g of $C[0, 1]$ like this:

$$f(x) = \begin{cases} \frac{1}{2} - x, & \text{if } 0 \leq x \leq \frac{1}{2} \\ 0, & \text{if } \frac{1}{2} \leq x \leq 1 \end{cases}$$

and

$$g(x) = \begin{cases} 0, & \text{if } 0 \leq x \leq \frac{1}{2} \\ x - \frac{1}{2}, & \text{if } \frac{1}{2} \leq x \leq 1. \end{cases}$$

(The graphs of f and g are shown below.) Then $fg = 0$, and so f is a zero divisor. However, $\psi(f) = \frac{1}{4}$, which is a unit of $\mathbb{R}$.



▷ **Quick Exercise.** Verify the claims made in the above examples. ◁

16.5 Making a Homomorphism Onto

Sometimes we wish to modify a homomorphism so that it becomes onto. Given homomorphism $\varphi : R \to S$, if we restrict the range to the set $\varphi(R)$, the new function (which we still call φ) $\varphi : R \to \varphi(R)$ is obviously onto. But is it still a homomorphism? Because we have not affected preservation of addition and multiplication by restricting the range, this new function must still be a homomorphism, *if* $\varphi(R)$ *is a ring.* But this is always the case, as we prove next:

Theorem 16.2 *Let* $\varphi : R \to S$ *be a homomorphism between rings* R *and* S*. Then* $\varphi(R)$ *is a subring of* S*.*

Proof: We need only check that the set $\varphi(R)$ is closed under subtraction and multiplication; then we'll know by the Subring Theorem 7.1 that it is a subring of S. For this purpose, choose $x, y \in \varphi(R)$; there exist $a, b \in R$ such that $\varphi(a) = x$ and $\varphi(b) = y$. Then

$$x - y = \varphi(a) - \varphi(b) = \varphi(a - b) \in \varphi(R),$$

and similarly, $xy = \varphi(ab) \in \varphi(R)$. □

Historical Remarks

The functional point of view is a very powerful one in all branches of mathematics. Analysts and topologists generally restrict themselves to *continuous* functions, because these are the functions that preserve analytic or topological structure. Similarly, in algebra we generally restrict ourselves to *homomorphisms*, because they are the functions that preserve algebraic structure.

The great 19th-century German mathematician Felix Klein was the first to use the word homomorphism in the context of a function preserving algebraic structure. He was actually talking about groups rather than rings; a group is a set endowed with algebraic structure that we will meet in Chapter 24.

Chapter Summary

In this chapter we introduced the idea of *ring homomorphism*, a function between two rings which preserves algebraic structure. We looked at numerous examples of such functions.

Warm-up Exercises

a. Consider the functions f, g, h, and k with range and domain $\mathbb{R}$, defined by

$$f(x) = x^2, \quad g(x) = x^3, \quad h(x) = e^x \quad k(x) = x^3 - x.$$

Which of these functions is one-to-one? Which of these functions is onto? (*Note*: We are *not* claiming that these functions are ring homomorphisms.)

b. Suppose that $f : X \to Y$ is a one-to-one function, and X is a finite set with n elements. What can you say about the number of elements in Y?

c. Suppose that $f : X \to Y$ is an onto function, and X is a finite set with n elements. What can you say about the number of elements in Y?

d. Give an example of a ring homomorphism $\varphi : R \to S$ satisfying each of the following (or explain why they cannot exist):

(a) φ is onto but isn't one-to-one.
(b) φ is one-to-one but isn't onto.
(c) φ is both one-to-one and onto.

(d) φ is neither one-to-one nor onto.

(e) R has unity 1, but $\varphi(1)$ is not unity for S.

(f) R has additive identity 0, but $\varphi(0)$ is not the additive identity for S.

(g) a is a zero divisor for R, but $\varphi(a)$ is not a zero divisor for S.

(h) a is a unit for R, but $\varphi(a)$ is not a unit for S.

(i) a is not a zero divisor for R, but $\varphi(a)$ is a zero divisor for S.

(j) a is not a unit for R, but $\varphi(a)$ is a unit for S.

e. We often write $\mathbb{Z}_4$ as $\{0, 1, 2, 3\}$. Using this notation, define the function $\varphi : \mathbb{Z}_4 \to \mathbb{Z}$ by $\varphi(n) = n$. Is this a ring homomorphism?

f. The function $\alpha : \mathbb{Z} \to \mathbb{Z}$ defined by $\alpha(n) = |n|$ is not a ring homomorphism. Why not?

g. Suppose that R and S are rings, and $\varphi : R \to S$ is an onto ring homomorphism.

(a) If R is a domain, is S a domain?

(b) If R is commutative, is S commutative?

Exercises

In Exercises 1–24, functions are defined whose domains and ranges are rings. In each case determine whether the function is a ring homomorphism, and whether it is one-to-one or onto. Justify your answers:

1. Define $\varphi : \mathbb{Z} \to m\mathbb{Z}$ by $\varphi(n) = mn$.
2. Define $\varphi : \mathbb{Q}[x] \to \mathbb{Q}[x]$ by $\varphi(f) = f'$. (f' is the formal derivative of f, discussed in Exercise 4.7.)
3. Define $\varphi : M_2(\mathbb{Z}) \to \mathbb{Z}$ by mapping a matrix to its determinant. (The determinant function is discussed in Exercise 8.2.)
4. Let R and S be rings and define $\pi_1 : R \times S \to R$ by $\pi_1(r, s) = r$, the **projection homomorphsim** from $R \times S$ to R. Similarly, we can define $\pi_2 : R \times S \to S$. (This generalizes Example 16.3.)
5. Let R be a ring, and define $\varphi : R \to R \times R$ by $\varphi(r) = (r, r)$.

6. Let R be a ring with unity and define $\varphi(r) = -r$. *Hint*: Consider Exercises 6.3 and 6.4.

   Case 1: There is at least one element $a \in R$ with $a + a \neq 0$.

   Case 2: For all $r \in R$, $r + r = 0$.

7. Recall from Example 7.9 that $D_2(\mathbb{Z})$ is the ring of 2-by-2 matrices with entries from $\mathbb{Z}$, with all entries off the main diagonal being zero. Define
$$\varphi : D_2(\mathbb{Z}) \to \mathbb{Z} \times \mathbb{Z}$$
by $\varphi\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} = (a, b)$.

8. Define $\varphi : \mathbb{Z}[\sqrt{2}] \to \mathbb{Z}$ by $\varphi(a + b\sqrt{2}) = a$. (See Exercise 7.1 for $\mathbb{Z}[\sqrt{2}]$.)

9. Define $\varphi : \mathbb{Z}[\sqrt{2}] \to \mathbb{Z}[\sqrt{2}]$ by $\varphi(a + b\sqrt{2}) = a - b\sqrt{2}$.

10. Let X be an arbitrary set; recall the power set ring $P(X)$ of subsets of X (described in Exercise 6.20). Choose some fixed $x \in X$. Then define $\varphi : P(X) \to \mathbb{Z}_2$ by setting
$$\varphi(a) = \begin{cases} [1], & x \in a \\ [0], & x \notin a. \end{cases}$$

11. Define $\varphi : \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$ by $\varphi(a, b) = a + b$.

12. Define $\varphi : \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$ by $\varphi(a, b) = ab$.

13. Recall the ring S of all real-valued sequences. (See Exercise 6.19.) Define $\varphi : S \to S$ by
$$\varphi(\{s_1, s_2, \cdots\}) = \{s_2, s_3, \cdots\}.$$
(That is, φ obtains the new sequence merely by dropping the first term of the sequence.)

14. Recall the ring $\mathbb{Z}[\alpha]$, where $\alpha = \sqrt[3]{5}$, as described in Exercise 7.3. Define $\varphi : \mathbb{Z}[\alpha] \to \mathbb{Z}_4$ by $\varphi(a + b\alpha + c\alpha^2) = [a + b + c]_4$.

15. Define $\varphi : \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}_6$ by $\varphi(a, b) = [a - b]_6$.

16. Define $\varphi : \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}_6$ by $\varphi(a, b) = [a + b]_6$.

17. Define $\varphi : \mathbb{Z}_2 \times \mathbb{Z}_3 \to \mathbb{Z}_6$ by $\varphi([a]_2, [b]_3) = [3a + 2b]_6$.

18. Define $\varphi : \mathbb{Z}_2 \times \mathbb{Z}_3 \to \mathbb{Z}_6$ by $\varphi([a]_2, [b]_3) = [3a + 4b]_6$.

19. Define $\varphi : \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}_6$ by $\varphi(a, b) = [3a + 4b]_6$.

20. Define $\varphi : \mathbb{Z}_6 \to \mathbb{Z}_4$ by $\varphi([a]_6) = [a]_4$.

21. Define $\varphi : \mathbb{C} \to \mathbb{R}$ by $\varphi(a + bi) = a$.

22. Define $\varphi : C[0,1] \to \mathbb{R} \times \mathbb{R}$ by $\varphi(f) = (f(0), f(1))$.

23. Define $\varphi : \mathbb{Q}[x] \to \mathbb{Q}$ by letting $\varphi(f)$ be the x-coefficient in f.

24. Define $\varphi : \mathbb{Q}[x] \to M_2(\mathbb{R})$ by

$$\varphi(f) = \begin{pmatrix} f(0) & f'(0) \\ 0 & f(0) \end{pmatrix}.$$

25. Let R be a commutative ring, and suppose that $\varphi : R \to R$ is a ring homomorphism. Consider

$$S = \{r \in R : \varphi(r) = r\}.$$

   (a) Show that S is a subring of R.

   (b) Show that S might not be an ideal.

26. Suppose that R and S are arbitrary rings, and $\varphi : R \to S$ is an onto ring homomorphism. Prove that $\varphi(Z(R)) \subseteq Z(S)$. (Recall from Exercise 7.12 that $Z(R)$ is called the center of the ring R.)

27. Suppose that R and S are commutative rings, and $\varphi : R \to S$ is a ring homomorphism. Let $N(R)$ and $N(S)$ be the nilradicals of R and S, respectively. (See Exercise 7.15.) Prove that $\varphi(N(R)) \subseteq N(S)$.

28. Let R and S be rings, with $\varphi : R \to S$ a ring homomorphism. Suppose that T is a subring of S. Let

$$\varphi^{-1}(T) = \{r \in R : \varphi(r) \in T\}.$$

Prove that $\varphi^{-1}(T)$ is a subring of R. When is $\varphi^{-1}(T)$ the whole ring R?

29. Suppose that R is a (not necessarily commutative) ring with unity, and a is a unit in R. Define the function $\varphi : R \to R$ by $\varphi(r) = ara^{-1}$. Prove that φ is a one-to-one, onto ring homomorphism from R to R.

30. Consider the ring S of real-valued sequences. (See Exercise 6.19.) Let

$$C = \{(x_1, x_2, x_3, \ldots) : \lim_{n\to\infty} x_n \text{ exists}\}.$$

This is the set of **convergent** sequences.

   (a) Prove that C is a subring of S.

   (b) Define $\varphi : C \to \mathbb{R}$ by setting

$$\varphi((x_1, x_2, x_3, \ldots)) = \lim_{n\to\infty} x_n.$$

   Prove that this is a ring homomorphism.

31. In this exercise we extend Theorem 16.2 to ideals.

   (a) Suppose that R and S are commutative rings and $\varphi : R \to S$ is an onto ring homomorphism. Let I be an ideal of R. Prove that $\varphi(I)$ is an ideal of S.

   (b) Show by example that part a is false, if we do not require that φ is onto.

Chapter 17

The Kernel

Let's consider the residue homomorphism $\varphi : \mathbb{Z} \to \mathbb{Z}_4$. There are exactly four residue classes modulo 4; namely,

$$\begin{gathered}\{\cdots, -4, 0, 4, 8, \cdots\}, \\ \{\cdots, -3, 1, 5, 9, \cdots\}, \\ \{\cdots, -2, 2, 6, 10, \cdots\}, \text{ and} \\ \{\cdots, -1, 3, 7, 11, \cdots\}.\end{gathered}$$

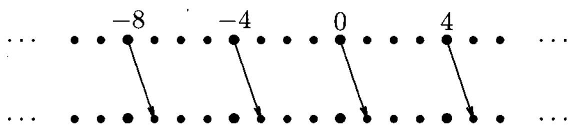
How do they relate to the function φ? The answer to this question is reasonably obvious: They consist of the four **pre-images** of the elements of $\mathbb{Z}_4$. Namely, the residue classes consist of the four sets

$$\begin{gathered}\varphi^{-1}([0]) = \{n \in \mathbb{Z} : \varphi(n) = [0]\}, \\ \varphi^{-1}([1]),\ \varphi^{-1}([2]), \text{ and } \varphi^{-1}([3]).\end{gathered}$$

Furthermore, these sets are rather nicely related: $\varphi^{-1}([0]) = 4\mathbb{Z}$, the multiples of 4, and the other three can be obtained from $4\mathbb{Z}$ by adding a fixed element to every element of $4\mathbb{Z}$;

$$\begin{aligned}\varphi^{-1}([1]) &= \{a + 1 : a \in 4\mathbb{Z}\}, \\ \varphi^{-1}([2]) &= \{a + 2 : a \in 4\mathbb{Z}\}, \text{ and} \\ \varphi^{-1}([3]) &= \{a + 3 : a \in 4\mathbb{Z}\}.\end{aligned}$$

We say that we have obtained $\varphi^{-1}([1])$, $\varphi^{-1}([2])$, and $\varphi^{-1}([3])$ by **additively translating** $4\mathbb{Z}$ by 1, 2, and 3, respectively.



translation of $4\mathbb{Z}$ by 1

17.1 Ideals

In Chapter 11 we introduced some notation to describe sets like $4\mathbb{Z}$, which we now review. For any commutative ring R and $a \in R$, let

$$\langle a \rangle = \{ar : r \in R\},$$

the multiples of a in R. Then we can write $4\mathbb{Z}$ as $\langle 4 \rangle$. And furthermore, we will denote the pre-image

$$\varphi^{-1}([1]) = \{a + 1 : a \in \langle 4 \rangle\}$$

as $\langle 4 \rangle + 1$. Using this notation, the four residue classes modulo 4 can then be written as

$$\langle 4 \rangle, \ \langle 4 \rangle + 1, \ \langle 4 \rangle + 2, \ \langle 4 \rangle + 3.$$

Note that the set $\varphi^{-1}([0]) = \langle 4 \rangle$ has some properties that the other pre-images do not have. Most obviously, $\langle 4 \rangle$ is a subring of $\mathbb{Z}$ (as we saw in Chapter 7). But in addition to this, $\langle 4 \rangle$ satisfies a stronger multiplicative closure property: namely, if a is an element of $\langle 4 \rangle$, then so is ar for any r in the ring $\mathbb{Z}$. In Chapter 11 we said that a subset I of a commutative ring R satisfies the **multiplicative absorption property** if whenever $a \in I$ and $r \in R$, then $ar \in I$. We called subrings with this stronger multiplicative closure **ideals**.

17.2 The Kernel

We'd like to generalize this situation as far as possible to arbitrary rings. This leads us to the following definition: Let $\varphi : R \to S$ be a homomorphism between the rings R and S. The **kernel** of φ is

$$\varphi^{-1}(0) = \{r \in R : \varphi(r) = 0\};$$

we will denote this set by $\ker(\varphi)$. In other words, the kernel is the pre-image of 0.

Thus, the kernel of the residue homomorphism above is exactly $\langle 4 \rangle$. Notice that the kernel always contains the additive identity of R (because homomorphisms take zero to zero). But as the example above shows, the kernel can include a great many other elements as well.

Let's go back to some of the examples from the previous chapter and compute their kernels.

Example 17.1

(From Example 16.2) This is the homomorphism that evaluates each rational polynomial at 2, and so its kernel is the set of all polynomials f such that $f(2) = 0$. By the Root Theorem, this is the set of all polynomials of the form $(x - 2)g$, where g is some arbitrary element of $\mathbb{Q}[x]$. Note that this set is precisely $\langle x - 2 \rangle$.

Example 17.2

(From Example 16.3.) This is the projection homomorphism onto the first component, from $\mathbb{Z} \times \mathbb{Z}$ to $\mathbb{Z}$. When is $\pi(a, b) = 0$? The answer is precisely when $a = 0$; thus,

$$\ker(\pi) = \{(0, b) : b \in \mathbb{Z}\}.$$

Note that this is precisely $\langle (0, 1) \rangle$.

Example 17.3

(From Example 16.4.) This is the homomorphism φ taking each complex number to its conjugate. Thus, $a + bi = \alpha \in \ker(\varphi)$ exactly if $a - bi = \bar{\alpha} = 0$. But this means that both a and b are zero, and so $\alpha = 0$. Thus,

$$\ker(\varphi) = \{0\} = \langle 0 \rangle.$$

Example 17.4

(From Example 16.6.) This is the zero homomorphism $\zeta : R \to S$, and its kernel is quite evidently the entire ring R.

Example 17.5

(From Example 16.12.) This is the homomorphism $\psi : C[0,1] \to \mathbb{R}$ that sends a continuous function to its value at 1/4. Here,

$$\ker(\psi) = \{f \in C[0,1] : f(1/4) = 0\}.$$

▷ **Quick Exercise.** Compute the kernels of the other examples of ring homomorphisms from the previous chapter. ◁

17.3 The Kernel is an Ideal

Note that every kernel from the above examples is an ideal: a subring that has the multiplicative absorption property. In the case of Example 17.4, this is evident because the kernel is the entire ring. In the case of Examples 17.1, 17.2, and 17.3, the kernel is of the form

$$\langle a \rangle = \{ar : r \in R\}.$$

We proved in Chapter 11 that this set is an ideal, for any commutative ring R. We called such ideals **principal ideals**.

▷ **Quick Exercise.** Reconstruct this proof yourself (or at least read the proof in Chapter 11). ◁

For Example 17.5, let's check explicitly that

$$\ker(\psi) = \{f \in C[0,1] : f(1/4) = 0\}$$

is an ideal. We first see that it is a non-empty set, by noting that the zero function is in $\ker(\psi)$. It is easy to check that $\ker(\psi)$ is closed under subtraction.

▷ **Quick Exercise.** Show that the kernel from Example 17.5 is closed under subtraction. ◁

This kernel also has the absorption property: If f is in $\ker(\psi)$ and g is any function in $C[0,1]$, then

$$(fg)(1/4) = f(1/4) \cdot g(1/4) = 0 \cdot g(1/4) = 0.$$

In other words, fg is also in $\ker(\psi)$.

With the examples above before us, you may be ready to believe the following theorem:

Theorem 17.1 *Let $\varphi : R \to S$ be a homomorphism between the commutative rings R and S. Then* $\ker(\varphi)$ *is an ideal of R.*

Proof: To show that $\ker(\varphi)$ is an ideal, we must first check that it is a non-empty set. But evidently, the zero element of R belongs to the kernel. We next check that the kernel is closed under subtraction. So suppose a and b are elements of the kernel. We need to check that $a - b \in \ker(\varphi)$. But $\varphi(a-b) = \varphi(a) - \varphi(b) = 0 - 0 = 0$. That is, $a - b \in \ker(\varphi)$, as required. Similarly, if $r \in R$, then $\varphi(ar) = \varphi(a)\varphi(r) = 0\varphi(r) = 0$, and so $ar \in \ker(\varphi)$. □

17.4 All Pre-images Can Be Obtained from the Kernel

The residue homomorphism example suggests that we can capture *all* pre-images by additively translating the kernel. Let's look at the pre-images of Example 17.1, the evaluation map

$$\varphi : \mathbb{Q}[x] \to \mathbb{Q}$$

defined by $\varphi(f) = f(2)$. This is more complicated than the residue homomorphism example because the range $\mathbb{Q}$ is infinite, and so there are infinitely many pre-images. Recall that the kernel of φ is $\langle x-2 \rangle$. We wish to show that it is also the case in this example that each pre-image can be obtained by additively translating the kernel. That is, we claim that each pre-image can be written as

$$\langle x-2 \rangle + g = \{f \in \mathbb{Q}[x] : f = h + g \text{ where } h \in \langle x-2 \rangle\},$$

for some choice of g. To show this, choose $a \in \mathbb{Q}$ and consider g in $\varphi^{-1}(a)$. This means that $\varphi(g) = g(2) = a$. We claim that every f in $\varphi^{-1}(a)$ can be written in the form $f = h + g$ where $h \in \langle x-2 \rangle$. This would show that $\varphi^{-1}(a) = \langle x-2 \rangle + g$. Now because $f(2) = a$, and $g(2) = a$, we have that $(f-g)(2) = 0$. That is, $f - g \in \ker(\varphi)$. But $\ker(\varphi) = \langle x-2 \rangle$, so $f - g = h$, where h is some multiple of $x - 2$. But then $f = h + g$, as we wish.

Note that the g we picked to represent $\varphi^{-1}(a)$ was chosen arbitrarily from all the elements of $\varphi^{-1}(a)$. Another element in $\varphi^{-1}(a)$ would have served just as well. The form of the translation would be different but would give the same set. In other words, if g_1 and g_2 are both in $\varphi^{-1}(a)$, then

$$\varphi^{-1}(a) = \langle x-2\rangle + g_1 = \langle x-2\rangle + g_2.$$

This is analogous to our freedom of choice in representing residue classes in $\mathbb{Z}_m$: for instance, $[4]_6 = [10]_6 = [16]_6$; or, to express this using the notation of ideals,

$$\langle 6\rangle + 4 = \langle 6\rangle + 10 = \langle 6\rangle + 16.$$

It turns out that we can always capture all the pre-images of a ring homomorphism by additively translating the kernel. To state this formally, we need a definition. Let I be an ideal of the ring R. Given $r \in R$, the **coset** of I determined by r consists of the set $\{a+r : a \in I\}$, which we write as $I + r$. Thus, $\langle 4\rangle + 3$ is a coset of the ideal $\langle 4\rangle$ in $\mathbb{Z}$.

▷ **Quick Exercise.** Consider the function φ in Example 17.1; describe the coset $\ker(\varphi) + 5$. What values do the polynomials in this set take on at 2? ◁

Theorem 17.2 *Let $\varphi : R \to S$ be a homomorphism between the rings R and S, and $s \in \varphi(R)$. Then $\varphi^{-1}(s)$ equals the coset $\ker(\varphi)+r$, where r is any given element of $\varphi^{-1}(s)$.*

Proof: Let $s \in \varphi(R)$, and choose any $r \in \varphi^{-1}(s)$ (which is nonempty by assumption). We must show that the sets $\ker(\varphi) + r$ and $\varphi^{-1}(s)$ are equal. Choose an arbitrary element $a + r \in \ker(\varphi) + r$, where $a \in \ker(\varphi)$. Then $\varphi(a+r) = \varphi(a) + \varphi(r) = 0 + \varphi(r) = s$. Thus, $a + r \in \varphi^{-1}(s)$, as claimed.

Conversely, choose $t \in \varphi^{-1}(s)$. Then consider $t - r$;

$$\varphi(t-r) = \varphi(t) - \varphi(r) = s - s = 0,$$

and so $t-r \in \ker(\varphi)$. But then $t = (t-r)+r \in \ker(\varphi)+r$, as required. □

Now because pre-images of distinct elements are clearly disjoint from one another, we have that the set of cosets of the kernel decomposes the domain ring into a set of pairwise disjoint subsets. That is, the set of cosets of the kernel partitions the ring. The unique one of these sets containing 0 is an ideal (because an ideal *has* to contain 0, it is clear that only one of the cosets can be an ideal). Notice that we can (and will) think of the ideal itself as a coset, namely, as $\ker(\varphi) + 0$.

Example 17.6

Let's explicitly compute this decomposition for the function $\varphi : \mathbb{Z}_{12} \to \mathbb{Z}_4$ defined by $\varphi([a]_{12}) = [a]_4$.

▷ **Quick Exercise.** Check that this function is a ring homomorphism. ◁

The kernel of φ is

$$\varphi^{-1}(0) = \{0,4,8\} = \langle 4\rangle = \langle 4\rangle + 0 = \langle 4\rangle + 4 = \langle 4\rangle + 8,$$

(where we have omitted the square brackets for simplicity's sake). The other distinct cosets of $\langle 4\rangle$ are

$$\varphi^{-1}(1) = \{1,5,9\} = \langle 4\rangle + 1 = \langle 4\rangle + 5 = \langle 4\rangle + 9,$$
$$\varphi^{-1}(2) = \{2,6,10\} = \langle 4\rangle + 2 = \langle 4\rangle + 6 = \langle 4\rangle + 10, \text{ and}$$
$$\varphi^{-1}(3) = \{3,7,11\} = \langle 4\rangle + 3 = \langle 4\rangle + 7 = \langle 4\rangle + 11.$$

Notice that in Example 17.6, each of the cosets has exactly the same number of elements. This is true in general because there is a one-to-one correspondence between any pair of cosets of an ideal: Given two cosets $I + a$ and $I + b$, the function $\alpha : I + a \to I + b$ defined by $\alpha(x) = x - a + b$ is both one-to-one and onto.

▷ **Quick Exercise.** Show that α is both one-to-one and onto. ◁

It is important to make clear that the function α is *not* a homomorphism; it does *not* preserve the operations, and the domain and range (except in the special case of the coset I) are not even rings. Two sets have the same number of elements exactly if there is a one-to-one correspondence between them; this is the situation in our example above. (Although we will not inquire into this here, this one-to-one correspondence is even useful if the ideal (and hence its cosets) is an infinite set, because it turns out that not all infinite sets can be put into one-to-one correspondence with one another; some are 'bigger' than others.)

We record our result formally:

Theorem 17.3 *Let I be an ideal of the commutative ring R, and let $I + a$ and $I + b$ be any two cosets of I. Then there is a one-to-one correspondence between the elements of $I + a$ and $I + b$. In particular, if these sets are finite, they have the same number of elements.*

17.5 When is the Kernel Trivial?

An important special case to consider of this one-to-one correspondence is when the kernel consists of only a single element; namely, when the ideal is $\{0\}$ (because an ideal is a subring this is the only possible one element ideal). Thus, *every* coset consists of a single element. Because the cosets consist of the pre-images of elements from the range ring, this means that a homomorphism with kernel $\{0\}$ is necessarily one-to-one. To rephrase: If a homomorphism takes *only* zero to zero, then there is only a single element that it takes to *any* of its values. We state this formally as a corollary:

Corollary 17.4 *A ring homomorphism is one-to-one if and only if its kernel is* $\{0\}$.

17.6 A Summary and Example

We have thus concluded that each homomorphism gives rise to an ideal, and this ideal in essence determines the pre-images of elements from the range of the function. We thus know (from just knowing the kernel) which elements in the ring are sent by the homomorphism to the same elements.

For example, consider the ring $\mathbb{Z} \times \mathbb{Z}$ and the ideal $I = \{(x, 0) : x \in \mathbb{Z}\}$.

▷ **Quick Exercise.** Check that this is an ideal. ◁

Suppose we are told that this is the kernel of some homomorphism. Then we know that all elements in the coset

$$I + (3,4) = I + (-5,4) = \{(x,4) : x \in \mathbb{Z}\}$$

must be sent by the homomorphism to the same element in the range ring (whatever that might be). Can you in fact think of a homomorphism with domain $\mathbb{Z} \times \mathbb{Z}$ of which I is the kernel? (If not, rest assured that in Example 19.8 we will return to this example.)

This all suggests that knowing the kernel of a homomorphism in essence gives us the homomorphism. This is precisely the content of the next chapter, where we will prove that *every* ideal of a ring is the kernel of some homomorphism. That is, we will prove the converse of the Theorem 17.1, which asserts that the kernel of a homomorphism is always an ideal.

Chapter Summary

In this chapter we defined the notion of the *kernel* of a ring homomorphism and proved that the kernel is always an *ideal*. Furthermore, the pre-images of a ring homomorphism are exactly *cosets* of the kernel.

Warm-up Exercises

a. What is the kernel of

$$\varphi : \mathbb{Z}_6 \to \mathbb{Z}_2$$

given by $\varphi([a]_6) = [a]_2$?

b. Write down the pre-images from the example in part a, and check that they are cosets.

c. Consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$. How many real numbers are there for which $f(x) = 0$? How about for $f(x) = 4$? Why does this tell us that f is *not* a homomorphism?

d. Give examples of ring homomorphisms φ satisfying the following (or say why you can't):

   (a) φ is onto, and its kernel is $\{0\}$.

   (b) φ is onto, and its kernel is not $\{0\}$.

   (c) φ is one-to-one, and its kernel is $\{0\}$.

   (d) φ is one-to-one, and its kernel is not $\{0\}$.

e. If you've read Chapter 11, give an example of a subring that is not the kernel of any ring homomorphism (if you can).

f. If you've read Chapter 11, give an example of a kernel of a ring homomorphism that is not a subring (if you can).

Exercises

1. Consider the homomorphism given in Exercise 16.5. What is its kernel? What does this mean about the homomorphism?

2. Consider the homomorphism given by Exercise 16.7. What is its kernel?

3. Consider the homomorphism given by Exercise 16.10. What is its kernel? Can you describe this kernel as a principal ideal $\langle b \rangle$, for some subset $b \in P(X)$?

4. What is the kernel of the homomorphism given by Exercise 16.13?

5. Consider the homomorphism φ given by Exercise 16.14. Prove that $\ker(\varphi) = \langle 3 + \alpha, 4 \rangle$.

6. What is the kernel of the homomorphism given by Exercise 16.17?

7. What is the kernel of the homomorphism given by Exercise 16.22?

8. Consider again the ring homomorphism of Example 17.1, namely, the function $\varphi : \mathbb{Q}[x] \to \mathbb{Q}$ defined by $\varphi(f) = f(2)$. Consider the translates of $\langle x - 2 \rangle$ by each of the constant polynomials (one for each rational). Show that each of these gives rise to a different pre-image of φ.

9. (Continuation of Exercise 8.) Now show that all the pre-images of φ can be obtained in this manner.

10. If F is a field and φ is a ring homomorphism, is $\varphi(F)$ also a field? If yes, prove it. If no, give a counterexample.

11. Consider the projection homomorphism $\pi_1 : R \times S \to R$ given in Exercise 16.4. What is the kernel of π_1? When do two elements of $R \times S$ get mapped to the same element of R? The set of pre-images of π_1 is naturally in one-to-one correspondence with what ring?

12. If I is an ideal of a commutative ring R, then show that two cosets of I (say, $I + a$ and $I + b$) are either equal or disjoint. (That is, the set of all translates of I *partition* R.) If you've previously encountered the idea of *equivalence relation*, rephrase this result in terms of that concept.

13. Consider the homomorphism given by Exercise 16.24. What is the kernel of this homomorphism? Can you describe this kernel as a principal ideal?

14. Let R be a commutative ring with unity, and suppose that e is an idempotent element. (That is, $e^2 = e$.) See Exercise 7.25 for more about idempotents. Define $\varphi : R \to R$ by setting $\varphi(r) = er$. Prove that φ is a ring homomorphism. Describe its kernel.

15. This is a converse for Exercise 14. Suppose that R is a commutative ring with unity, $a \in R$, and $\varphi(r) = ar$ defines a ring homomorphism. Prove that a is idempotent.

16. Let R be a finite commutative ring, with n elements. Suppose that I is an ideal of R with m elements. Prove that m divides n. (This result will be put in a more general context in Theorem 31.2, which is called Lagrange's Theorem.)

17. Suppose that F is a finite field with characteristic 2. (See Exercise 8.11 for a definition of characteristic.)

    (a) Prove that $\varphi : F \to F$, defined by $\varphi(r) = r^2$ is a ring isomorphism.

    (b) One example of a field with characteristic 2 is $\mathbb{Z}_2$. Describe the isomorphism φ explicitly in this case.

    (c) Another example of a field with characteristic 2 is the field described in Exercise 8.12, which consists of the elements

    $$\{0, 1, \alpha, 1 + \alpha\}.$$

    Describe the isomorphism φ explicitly in this case.

18. Generalize Exercise 17. That is, suppose that F is a finite field with characteristic p.

    (a) Prove that $\varphi : F \to F$ defined by $\varphi(r) = r^p$ is a ring isomorphism. This function is called the **Frobenius isomorphism**.

(b) Suppose now that the finite field F is the field $\mathbb{Z}_p$. Explain why Fermat's Little Theorem 8.7 implies that in this case the Frobenius isomorphism is actually the identity map. (Note that in Chapter 46 we will encounter many finite fields that are not of the form $\mathbb{Z}_p$.)

Chapter 18

Rings of Cosets

In practice when we think of the ring $\mathbb{Z}_4$ we think of the four elements $[0], [1], [2], [3]$ (or even $0, 1, 2, 3$) together with the appropriate operations. But technically, when we first considered $\mathbb{Z}_4$ and its operations in Chapter 3, we defined those elements as infinite sets of integers; that is,

$$[a] = \{a + 4m : m \in \mathbb{Z}\} = \{n \in \mathbb{Z} : 4|(a - n)\}.$$

We then defined addition and multiplication by setting

$$[a] + [b] = [a + b] \quad \text{and} \quad [a][b] = [ab].$$

While these definitions look innocuous enough, what we first had to do was to check that they make sense (or are *well defined*): if $[c] = [a]$ and $[b] = [d]$, do $[c + d] = [a + b]$ and $[cd] = [ab]$? We now wish to follow the same line of thought for the cosets of an arbitrary ideal of a commutative ring.

18.1 The Ring of Cosets

We first need some notation. Let R be a commutative ring with ideal I. Then we denote the set $\{I + r : r \in R\}$ of all cosets of I in R by R/I; we read this as 'R modulo I'. Note that by defining R/I in this way, it looks as if R/I has as many elements as R does. But this is certainly not the case, for different elements of R may well give rise to the same coset.

For example, we saw in Section 17.4 that we can write

$$\mathbb{Z}/\langle 4 \rangle = \{\langle 4 \rangle + n : n \in \mathbb{Z}\} = \{\langle 4 \rangle + 0, \langle 4 \rangle + 1, \langle 4 \rangle + 2, \langle 4 \rangle + 3\}.$$

Another example we discussed in the previous chapter is this:

$$\mathbb{Q}[x]/\langle x - 2 \rangle = \{\langle x - 2 \rangle + f : f \in \mathbb{Q}[x]\} = \{\langle x - 2 \rangle + q : q \in \mathbb{Q}\}.$$

What we are going to do now is define an addition and multiplication on R/I in such a way as to make it a commutative ring. We do this so that $\mathbb{Z}/\langle 4\rangle$ becomes a ring essentially the same as $\mathbb{Z}_4$, and $\mathbb{Q}[x]/\langle x-2\rangle$ becomes a ring essentially the same as $\mathbb{Q}$.

Given a commutative ring R with ideal I, we define

$$(I+a)+(I+r)=I+(a+r) \quad \text{and} \quad (I+a)(I+r)=I+ar.$$

We must now check that these definitions make sense. As when checking that addition and multiplication are well defined on $\mathbb{Z}_m$, we must show these definitions for addition and multiplication on cosets of I are well defined by showing that different representations of the cosets yield the same sum (and product). That is, suppose that $I+a=I+c$ and $I+r=I+s$: Is

$$(I+a)+(I+r)=(I+b)+(I+s)?$$

Is

$$(I+a)(I+r)=(I+b)(I+s)?$$

What does such an assumption as $I+a=I+b$ mean? In other words: When do two elements determine the same coset of I? This is important enough to characterize in the following theorem. But before stating the theorem, think again about the example $\mathbb{Z}/\langle 4\rangle$. When do two integers a and b determine the same coset (or in this case, residue class modulo 4)? The answer is exactly if their difference $a-b$ belongs to the ideal $\langle 4\rangle$. This is the answer in general; we shall prove this now, along with some other important observations about cosets, in the Coset Theorem.

Theorem 18.1 The Coset Theorem *Let I be an ideal of the commutative ring R with $a,b\in R$.*

a. If $I+a\subseteq I+b$, then $I+a=I+b$.

b. If $I+a\cap I+b\neq\emptyset$, then $I+a=I+b$.

c. $I+a=I+b$ if and only if $a-b\in I$.

d. There exists a one-to-one and onto function between any two cosets $I+a$ and $I+b$. Thus, if I has finitely many elements, every coset has that same number of elements.

Notice that if you have done Exercise 17.12, you have already checked that parts (a) and (b) are true.

Proof: (a) Suppose that I is an ideal of the commutative ring R, and a and b are elements of the ring for which $I+a\subseteq I+b$. Then

$$a=0+a\in I+a\subseteq I+b,$$

and so there exists $x\in I$ such that $a=x+b$. But then $b=-x+a\in I+a$. Now, if $k\in I$, $k+b=(k-x)+a\in I+a$, and so $I+b\subseteq I+a$. That is, $I+a=I+b$.

(b): Suppose that $I+a\cap I+b\neq\emptyset$. Choose c in this intersection. Then $c\in I+a$, and so $I+c\subseteq I+a$. But then by part (a), $I+c=I+a$. But similarly, $I+c=I+b$, and so $I+a=I+b$.

(c): If $I+a=I+b$, then $a=0+a\in I+a=I+b$, and so there exists $k\in I$ such that $a=k+b$. But then $a-b=k\in I$, as required. Conversely, suppose if $a-b\in I$, then $a=(a-b)+b\in I+b$. But then $a\in I+a\cap I+b$, and so by part (b), $I+a=I+b$.

(d): In the Quick Exercise following Example 17.6, you argued that the function $\alpha: I+a\to I+b$ defined by $\alpha(x)=x-a+b$ is one-to-one and onto. □

We now use the Coset Theorem 18.1, as promised, to show that the addition and multiplication we have defined above on R/I are well defined:

Proof that Operations are Well Defined: We can now return to the task of checking that the above definitions of addition and multiplication for R/I make sense. Suppose that $I+a=I+b$ and $I+r=I+s$; we claim first that $I+(a+r)=I+(b+s)$. But this amounts to claiming that $(a+r)-(b+s)\in I$, and because $(a+r)-(b+s)=(a-b)+(r-s)$, this is clear. We claim next that $I+ar=I+bs$, or in other words, that $ar-bs\in I$. But

$$ar-bs=ar-br+br-bs=(a-b)r+b(r-s).$$

Because I has the multiplicative absorption property, $(a-b)r\in I$ and $b(r-s)\in I$, and so therefore $ar-bs\in I$. □

Notice that in order to show that multiplication makes sense for cosets, we needed the full strength of the definition of ideal. You will see by example in Exercise 18.8, that multiplication of cosets does *not* make sense if I is merely a subring.

Now that we have the appropriate operations defined on R/I, the rest of the following theorem is easy:

Theorem 18.2 *Let I be an ideal of the commutative ring R. Then the set R/I of cosets of I in R, under the operations defined above, is a commutative ring.*

Proof: We know from above that the addition and multiplication defined on R/I are in fact binary operations. They are associative and commutative because the corresponding operations on R are:

$$\begin{aligned}(I+a)((I+b)(I+c)) &= (I+a)(I+bc) = I+a(bc)\\ &= I+(ab)c = ((I+a)(I+b))(I+c),\end{aligned}$$

and similarly for addition. In a similar fashion the distributive law for R/I carries over from the distributive law for R. You should check that the additive identity for R/I is $I+0$, and the additive inverse of $I+a$ is $I+(-a)$.

▷ **Quick Exercise.** Perform these verifications. ◁

□

The ring R/I is called the **ring of cosets**, or the **quotient ring of R modulo I**.

Example 18.1

Let's look at an example of this construction in the commutative ring $\mathbb{Z}_{12}$. Consider the ideal

$$\langle 4\rangle = \{0,4,8\}.$$

This ideal has 4 distinct cosets:

$$\mathbb{Z}_{12}/\langle 4\rangle = \{\langle 4\rangle, \langle 4\rangle+1, \langle 4\rangle+2, \langle 4\rangle+3\}.$$

▷ **Quick Exercise.** Write down the multiplication and addition tables for the ring $\mathbb{Z}_{12}/\langle 4\rangle$, and thus check quite explicitly that this is a ring. ◁

18.2 The Natural Homomorphism

Suppose that R is a commutative ring with ideal I. Consider the function $\nu : R \to R/I$ defined by

$$\nu(a) = I + a.$$

Because of the definition of the operations on R/I, it is obvious that this function preserves them both. Thus, ν is a homomorphism from R onto R/I. We call it the **natural homomorphism from R onto R/I.**

Our experience in the previous chapter suggests the following question: What is the kernel of the natural homomorphism? If $\nu(a) = I+0$ (the additive identity of R/I), then $I+a = I+0$. But by the Coset Theorem 18.1, this is true exactly if $a = a-0 \in I$. Thus, the kernel of the natural homomorphism from R to R/I is exactly I. This is very easy but is important enough to record as a theorem:

Theorem 18.3 *Let R be a commutative ring with ideal I, and $\nu : R \to R/I$ the natural homomorphism. Then* $\ker(\nu) = I$.

In the last chapter we saw that every kernel of a homomorphism is an ideal. In this chapter we have seen that every ideal is the kernel of some homomorphism, namely, the natural homomorphism from R to R/I. We have thus obtained a completely different way of thinking about ideals: *Ideals are kernels of homomorphisms.* This is the sort of surprise that mathematicians particularly enjoy: Two apparently quite different ideas (in this case, homomorphisms and ideals) turn out to be inextricably linked! But to really say that the ideas of homomorphisms and ideals amount to the same thing, we need a bit more. To see this, let's look again at an example.

Consider again the evaluation homomorphism $\psi : \mathbb{Q}[x] \to \mathbb{Q}$ where $\psi(f) = f(2)$. We know that $\psi(\mathbb{Q}[x]) = \mathbb{Q}$. Now ψ has kernel $\langle x-2\rangle$, which is an ideal. In this chapter we have constructed a ring and a homomorphism of which the ideal $\langle x-2\rangle$ is the kernel; namely, we have the ring of cosets $\mathbb{Q}[x]/\langle x-2\rangle$ and the natural homomorphism $\nu : \mathbb{Q}[x] \to \mathbb{Q}[x]/\langle x-2\rangle$. The two functions ψ and ν are certainly different; the range of one is the set of rational numbers, while the range of the other is a certain set of subsets of $\mathbb{Q}[x]$. However, in structure these two ranges (and the functions ψ and ν which connect them to the domain $\mathbb{Q}[x]$) *seem essentially the same.* The ring of cosets $\mathbb{Q}[x]/\langle x-2\rangle$ appears to be just our old friend the rationals, in disguised garb. Our goal in the next chapter is to show that this is no accident.

Chapter Summary

In this chapter we saw how to make a ring out of the set of cosets of an ideal in a commutative ring, generalizing the construction of $\mathbb{Z}_n$ in Chapter 3.

Warm-up Exercises

a. How many elements does the ring

$$\mathbb{Z}_{16}/\langle 4\rangle$$

have? Which one is its additive identity? Does this ring have unity? Does it have any zero divisors?

b. Compute

$$(\langle 4\rangle + 3)(\langle 4\rangle + 2)$$

in $\mathbb{Z}_{12}/\langle 4\rangle$ twice, using different representatives from these two cosets.

c. What is the additive inverse of

$$\langle x-2\rangle + (x^2-2)$$

in $\mathbb{Q}[x]/\langle x-2\rangle$?

d. What is the multiplicative inverse of

$$\langle x-2\rangle + (x^2-2)$$

in $\mathbb{Q}[x]/\langle x-2\rangle$?

e. Let R be a commutative ring and I one of its ideals. What is the nature of the elements belonging to the set R/I?

f. What is the kernel of the natural homomorphism $\nu : R \to R/I$?

g. Is the natural homomorphism always onto? Why or why not?

h. Is the natural homomorphism always one-to-one? Why or why not?

i. Is every ideal of a commutative ring a kernel of some homomorphism? Why or why not?

Exercises

1. Consider the ideal $\langle x^2\rangle$ in $\mathbb{Q}[x]$.

   (a) Prove that for each $f \in \mathbb{Q}[x]$, there exists $a + bx \in \mathbb{Q}[x]$, so that
   $$\langle x^2\rangle + f = \langle x^2\rangle + (a+bx).$$

   (b) Show that $\mathbb{Q}[x]/\langle x^2\rangle$ is not a domain.

2. Consider the ring $\mathbb{Z}/I$, for some ideal I. For what ideals I does this ring have infinitely many elements? For what ideals I does this ring have finitely many elements?

3. Consider the ideal $\langle 1+i\rangle$ in $\mathbb{Z}[i]$.

   (a) Make use of the description of this ideal provided in Exercise 13.5 to show that for all $a + bi \in \mathbb{Z}[i]$, there exists $c \in \mathbb{Z}$, such that
   $$\langle 1+i\rangle + (a+bi) = \langle 1+i\rangle + c.$$

   (b) Use part a to prove that $\mathbb{Z}[i]/\langle 1+i\rangle$ has only two elements.

4. Consider the ideal
   $$I = \{f \in C[0,1] : f(1/4) = 0\}$$
   in the commutative ring $C[0,1]$; we considered this ideal in Example 17.5.

   (a) Prove that
   $$I + f = I + g \text{ if and only if } f(1/4) = g(1/4).$$

   (b) Prove that $C[0,1]/I$ is a field.

5. Consider the ideal $I = \langle(3,4)\rangle$ of the ring $\mathbb{Z}\times\mathbb{Z}$. Prove that $(\mathbb{Z}\times\mathbb{Z})/I$ is not a domain.

6. Prove that $J = \{(x,0) : x \in \mathbb{Z}\}$ is an ideal in $\mathbb{Z}\times\mathbb{Z}$. Prove that $(\mathbb{Z}\times\mathbb{Z})/J$ is a domain (even though $\mathbb{Z}\times\mathbb{Z}$ is *not* a domain).

7. Consider the ring $\mathbb{Z}[\alpha]$ described in Exercise 7.3.

(a) Prove that any element

$$\langle 1+\alpha\rangle + (a+b\alpha+c\alpha^2) \in \mathbb{Z}[\alpha]/\langle 1+\alpha\rangle$$

can be written in the form $\langle 1+\alpha\rangle + m$, where m is an integer.

(b) Show that $1+\alpha$ divides 6.

(c) Use (a) and (b) to show that

$$\mathbb{Z}[\alpha]/\langle 1+\alpha\rangle = \{\langle 1+\alpha\rangle + 0, \langle 1+\alpha\rangle + 1, \langle 1+\alpha\rangle + 2, \\ \cdots, \langle 1+\alpha\rangle + 5\}.$$

(d) Use c to show that $\mathbb{Z}[\alpha]/\langle 1+\alpha\rangle$ is not a domain.

8. The ring $\mathbb{Z}$ is a subring of $\mathbb{Q}$ but is not an ideal. Therefore, it makes no sense to speak of the ring of cosets $\mathbb{Q}/\mathbb{Z}$. Show by explicit example that multiplication makes no sense for $\mathbb{Q}/\mathbb{Z}$. (We will see in Chapter 32 that addition *does* make sense. This means that $\mathbb{Q}/\mathbb{Z}$ is an additive group, but not a ring.)

9. Consider the ring S of real-valued sequences, and Σ the ideal of S, considered in Exercise 12.8.

   (a) Give a nice description of the elements of the coset $\Sigma + (1,1,1,\ldots)$.

   (b) Show by explicit computation that S/Σ is not a domain.

   (c) Show that the ring S/Σ has infinitely many distinct idempotents. (Recall from Exercise 7.25 that an idempotent is an element e for which $e^2 = e$.)

10. Let R be a commutative ring; recall from Exercise 7.15 that the nilradical $N(R)$ of R is a subring.

    (a) Prove that $N(R)$ is an ideal of R.

    (b) Prove that the ring $R/N(R)$ has no non-zero nilpotent elements.

    (c) Check explicitly that part b is true for the ring $R = \mathbb{Z}_8$. To what ring is $\mathbb{Z}_8/N(\mathbb{Z}_8)$ isomorphic?

11. Consider the following alternate definition of addition of the cosets of an ideal I of a commutative ring R.

$$(I+a)+(I+b) = \{x+y : x \in I+a, y \in I+b\}.$$

Prove that this definition is equivalent to the definition of addition on R/I given in the text. (Compare this exercise to Exercise 3.11.)

Chapter 19

The Isomorphism Theorem for Rings

What we now must do is make careful sense of what we mean by saying that two rings 'are essentially the same'. Clearly there should be a one-to-one correspondence between the elements of the two rings, and this one-to-one correspondence should preserve the ring structure. Preserving the ring structure just means that we have a homomorphism.

19.1 Isomorphism

We make this formal in the following definition: Let R and S be rings. If there exists a one-to-one onto homomorphism $\varphi : R \to S$, we say that R and S are **isomorphic**; the function φ we call an **isomorphism.**

Recall (the Corollary 17.4) that a homomorphism is one-to-one if and only if its kernel is $\{0\}$; this fact then applies to isomorphisms and in practice is the way to check that part of their definition.

Let's now look at some examples of isomorphisms.

Example 19.1

Recall the identity homomorphism, defined on a ring R; it is the map $\iota : R \to R$ defined by $\iota(r) = r$. This is obviously an isomorphism. Here, the domain and range rings are *exactly* the same, rather than only 'essentially' the same.

Example 19.2

For a more interesting example, we claim that the function

$$\lambda : \mathbb{Q} \to \mathbb{Q}[x]/\langle x - 2\rangle$$

defined by $\lambda(r) = \langle x - 2\rangle + r$ is an isomorphism. Because of the way addition and multiplication are defined on the ring of cosets, this function clearly preserves them.

▷ **Quick Exercise.** Verify that the function λ preserves addition and multiplication. ◁

We next claim that λ is one-to-one. To do this, we show that its kernel is $\{0\}$. For that purpose, suppose that $r \in \mathbb{Q}$ and $\lambda(r) = 0$; we must show that $r = 0$. But $\langle x-2\rangle + r = \langle x-2\rangle + 0$ means that $r \in \langle x-2\rangle$; that is, the rational number r is a multiple of the polynomial $x - 2$. But $\deg(r) < \deg(x-2)$ forces us to conclude that $r = 0$.

Finally, we claim that this function is onto. Suppose that

$$\langle x-2\rangle + f \in \mathbb{Q}[x]/\langle x-2\rangle.$$

We must find an element of $\mathbb{Q}$ whose value is $\langle x-2\rangle + f$. By applying the Division Theorem 4.2 we have that $f = (x-2)q + r$, where the degree of the remainder r is less than the degree of $x - 2$; this means that r is a constant. But then

$$\lambda(r) = \langle x-2\rangle + r = \langle x-2\rangle + (x-2)q + r = \langle x-2\rangle + f,$$

as required.

We can certainly have functions between rings that satisfy several of the requirements for being an isomorphism without meeting them all. Here are a number of examples of this phenomenon:

Example 19.3

Suppose that R is a proper subring of the ring S; recall the inclusion homomorphism $\iota : R \to S$, where $\iota(r) = r$. This homomorphism is one-to-one but not onto.

Example 19.4

Consider the rings $\mathbb{Z}_4$ and $\mathbb{Z}_2 \times \mathbb{Z}_2$; they both have 4 elements, and so there exists a one-to-one correspondence between them. Is there any such correspondence that preserves addition and multiplication? We claim something stronger: There isn't even a one-to-one correspondence that preserves addition. If there were, 0 would have to go to $(0,0)$. (See Theorem 16.1a.) There are then three possibilities for where 1 might be mapped: $(1,0), (0,1), (1,1)$. But if the correspondence preserves addition, then the images of

$$2 = 1 + 1 \text{ and } 3 = 1 + 1 + 1$$

under the correspondence are forced by choice of the image of 1. It is easy to see that none of the three choices listed above allow a one-to-one function.

▷ **Quick Exercise.** Check that each of the three choices above yield a function that is not one-to-one. ◁

Example 19.5

Consider the rings $\mathbb{Z}$ and $2\mathbb{Z}$. There certainly is a one-to-one onto function between these rings that preserves addition; namely, $\varphi(n) = 2n$. But there can be no such onto map that also preserves multiplication, because $\mathbb{Z}$ has unity, while $2\mathbb{Z}$ doesn't. (See Theorem 16.1c.)

Suppose now that $\varphi : R \to S$ is an isomorphism. Let's consider the **inverse function** φ^{-1}. How is φ^{-1} defined? Given $s \in S$, there exists $r \in R$ such that $\varphi(r) = s$ (because φ is onto). But there is *only one* such r (because φ is one-to-one). It thus makes sense to define $\varphi^{-1}(s) = r$. This new function is also clearly one-to-one and onto. But is it an isomorphism? Given $s, t \in S$, suppose that x, y are the unique elements of R for which $\varphi(x) = s$ and $\varphi(y) = t$. Then

$$\begin{aligned}\varphi^{-1}(s+t) &= \varphi^{-1}(\varphi(x) + \varphi(y)) = \varphi^{-1}(\varphi(x+y)) \\ &= x + y = \varphi^{-1}(s) + \varphi^{-1}(t),\end{aligned}$$

and similarly φ^{-1} also preserves multiplication. To summarize, we have just shown that the inverse function of an isomorphism is an isomorphism. Thus, to say that R and S are isomorphic really is a symmetric relationship; the order in which we state them doesn't matter, because the existence of an isomorphism in one direction implies the existence of an isomorphism in the other direction.

19.2 The Fundamental Isomorphism Theorem

We're now ready to show the true equivalence of the notions of homomorphism and ideal by proving the theorem we've been leading up to for some time. Given an onto homomorphism $\varphi : R \to S$ between commutative rings R and S, we know that $\ker(\varphi)$ is an ideal, and so we have another homomorphism $\nu : R \to R/\ker(\varphi)$. We claim that the ranges of these homomorphisms are 'essentially the same', and what's more, the functions themselves are 'essentially the same'. We use the language of isomorphism to state this formally.

Theorem 19.1 The Fundamental Isomorphism Theorem for Commutative Rings *Let $\varphi : R \to S$ be an onto homomorphism between rings, and let $\nu : R \to R/\ker(\varphi)$ be the usual natural homomorphism. Then there exists an isomorphism $\mu : R/\ker(\varphi) \to S$ such that $\mu \circ \nu = \varphi$.*

We exhibit this situation in the following diagram:

$$\begin{tikzcd}
R \arrow[rr, "\varphi"] \arrow[dr, "\nu"'] & & S \\
& R/\ker(\varphi) \arrow[ur, "\mu"'] &
\end{tikzcd}$$

For simplicity's sake, in this book we are restricting our attention in this theorem to *commutative* rings, because we've really only looked at the concept of ideal in this case. However, at the expense of only slightly more work, a more general such theorem actually remains true in the case of arbitrary rings; we won't pursue this matter here.

Proof: Suppose that R and S are commutative rings, $\varphi : R \to S$ is an onto homomorphism, and $\nu : R \to R/\ker(\varphi)$ is the natural homomorphism. Clearly what we need to do is define a function μ, and prove that it has the desired properties. Once we've obtained the appropriate definition, the rest of the proof will flow smoothly (if a bit lengthily, because there are a lot of properties for μ we have to verify).

Choose an arbitrary element of $R/\ker(\varphi)$; it is a coset of the form $\ker(\varphi) + r$, where $r \in R$. What element of S should it correspond to? If the composition of functions required in the theorem is to work, we must have that $\mu(\ker(\varphi) + r) = \varphi(r)$; this is our definition of μ.

There is an immediate problem we must solve: This definition apparently *depends on the particular coset representative* r. That is, our function does not appear to be unambiguously defined. But suppose that $\ker(\varphi) + r = \ker(\varphi) + s$. Then $r - s \in \ker(\varphi)$, and so $\varphi(r) = \varphi(s)$. Thus, it really didn't matter what coset representative we chose, and so our function is well defined. Furthermore, it has been defined precisely so that $\mu \circ \nu(r) = \mu(\ker(\varphi) + r) = \varphi(r)$.

We must now check that μ is an isomorphism; we take each of the required properties in turn:

μ is a homomorphism: But

$$\begin{aligned}
\mu((\ker(\varphi) + r)(\ker(\varphi) + s)) &= \mu(\ker(\varphi) + rs) \\
&= \varphi(rs) = \varphi(r)\varphi(s) \\
&= \mu(\ker(\varphi) + r)\mu(\ker(\varphi) + s),
\end{aligned}$$

and similarly for addition.

μ is onto: But φ is onto, and so for any $s \in S$, there exists $r \in R$ such that $\varphi(r) = s$. Then $\mu(\ker(\varphi) + r) = \varphi(r) = s$.

μ is *one-to-one*: Suppose that $\mu(\ker(\varphi) + r) = \mu(\ker(\varphi) + s)$; then $\varphi(r) = \varphi(s)$, and so $r - s \in \ker(\varphi)$. But then $\ker(\varphi) + r = \ker(\varphi) + s$, as required. (You should note that this argument is just the reverse of the argument proving that μ is well defined.)

▷ **Quick Exercise.** Think about why the parenthetic remark is true. ◁

We thus have the isomorphism required by the theorem. □

19.3 Examples

Example 19.6

Let's look yet again at Example 16.2 where $\varphi : \mathbb{Q}[x] \to \mathbb{Q}$ is the function that evaluates a polynomial at 2. Because this homomorphism is onto and its kernel is $\langle x - 2 \rangle$, Theorem 19.1 asserts that the rings $\mathbb{Q}[x]/\langle x - 2 \rangle$ and $\mathbb{Q}$ are isomorphic, via the map μ which is defined by $\mu(\langle x - 2 \rangle + f) = f(2)$. In Example 19.2 we proved directly that these two rings are isomorphic, via the map $\lambda : \mathbb{Q} \to \mathbb{Q}[x]/\langle x - 2 \rangle$ defined by $\lambda(q) = \langle x - 2 \rangle + q$. What is the relationship between the functions λ and μ? They are inverse functions. For

$$\mu(\lambda(q)) = \mu(\langle x - 2 \rangle + q) = \mu \circ \nu(q) = \varphi(q) = q(2) = q.$$

(Notice this last equality holds because we are thinking of q as a constant polynomial.)

Example 19.7

If we apply the theorem to the residue homomorphism $\varphi : \mathbb{Z} \to \mathbb{Z}_4$, we obtain what we already suspected: $\mathbb{Z}/\langle 4\rangle$ and $\mathbb{Z}_4$ are isomorphic rings.

Example 19.8

Consider the function $\varphi : \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$ defined by $\varphi(x, y) = y$; you can easily check that this is an onto homomorphism. (Indeed, this was verified in a more general context as Exercise 16.4.) The kernel of this homomorphism is clearly $\{(x, 0) : x \in \mathbb{Z}\}$, which we might write as $\mathbb{Z} \times \{0\}$. (Note that in Section 17.6 we asked you for a homomorphism with this ideal as kernel.) Our theorem thus asserts that $(\mathbb{Z} \times \mathbb{Z})/(\mathbb{Z} \times \{0\})$ is isomorphic to $\mathbb{Z}$. Notice also that in Exercise 18.6 you proved that this quotient ring is a domain; because we've now shown that it is isomorphic to $\mathbb{Z}$, this result should not be surprising.

Example 19.9

Let's now consider two general but trivial examples of the theorem. Given any commutative ring R, it always possesses two ideals; namely, the trivial ideal $\{0\}$ and the improper ideal R itself. These are certainly the kernels of the identity isomorphism $\iota : R \to R$ and of the zero homomorphism $\zeta : R \to \{0\}$, respectively. Theorem 19.1 then asserts in the one case that R is isomorphic to $R/\{0\}$, and in the other that R/R is isomorphic to $\{0\}$. Speaking informally, this says that if we mod zero out of a ring, we have left it unaffected, while if we mod out the entire ring, we are left with the zero ring.

Example 19.10

As another example of the theorem, consider the function $\varphi : \mathbb{R}[x] \to \mathbb{C}$ defined by $\varphi(f) = f(i)$. Notice that evaluating a polynomial with real coefficients at the complex number i will certainly give a complex number. As usual, an evaluation map of this sort is a homomorphism. What is the kernel of this function? It consists of $\{f \in \mathbb{R}[x] : f(i) = 0\}$, but these are precisely those polynomials which have $x^2 + 1$ as a factor. That is, the kernel here is again a principal ideal, this time $\langle x^2 + 1\rangle$. Thus, we have that the field of complex numbers is isomorphic to $\mathbb{R}[x]/\langle x^2 + 1\rangle$.

The previous example is actually an algebraically elegant way of describing how we obtain the complex numbers from the real numbers. The goal in obtaining the complex numbers is after all to be able to solve more equations (such as $x^2 + 1 = 0$). We will inquire more carefully into this example soon, but we will first have to characterize those ideals that lead to fields as their rings of cosets. This characterization is one of the goals of the next chapter.

Example 19.11

For another more sophisticated example of a homomorphism onto a field, consider the function

$$\varphi : \mathbb{Z}[\sqrt{-5}] \to \mathbb{Z}_3$$

defined as $\varphi(a + b\sqrt{-5}) = [a - b]_3$. (Recall Exercise 7.2 and our discussion in Section 10.3 for a description of rings of the form $\mathbb{Z}[\sqrt{n}]$.)

Let's first check that this is a homomorphism. We see that φ preserves addition:

$$\varphi((a + b\sqrt{-5}) + (c + d\sqrt{-5})) = [(a + c) - (b + d)] =$$

$$[a - b] + [c - d] = \varphi(a + b\sqrt{-5}) + \varphi(c + d\sqrt{-5}).$$

Similarly, φ preserves multiplication:

$$\begin{aligned}\varphi((a + b\sqrt{-5})(c + d\sqrt{-5})) &= \varphi((ac - 5bd) + (ad + bc)\sqrt{-5}) \\ &= [ac - 5bd - ad - bc] \\ &= [ac - ad + bd - bc] \\ &= [(a - b)(c - d)] \\ &= \varphi(a + b\sqrt{-5})\varphi(c + d\sqrt{-5})\end{aligned}$$

(where we've used the fact that $[-5] = [1]$). The kernel of this function is $\{a + b\sqrt{-5} : 3|(a - b)\}$. But in Example 12.3 we showed that this is the smallest ideal of $\mathbb{Z}[\sqrt{-5}]$ that contains the elements 3 and $1 + \sqrt{-5}$; we denoted this ideal by $\langle 3, 1 + \sqrt{-5}\rangle$. We have thus concluded that the ring of cosets

$$\mathbb{Z}[\sqrt{-5}]/\langle 3, 1 + \sqrt{-5}\rangle$$

is isomorphic to $\mathbb{Z}_3$. In the next chapter we will discover the property that the ideals $\langle x^2+1\rangle$ in $\mathbb{R}[x]$ and $\langle 3, 1+\sqrt{-5}\rangle$ in $\mathbb{Z}[\sqrt{-5}]$ have in common.

Historical Remarks

The equivalence between homomorphisms and kernels, as expressed in the Fundamental Isomorphism Theorem for Commutative Rings, is a crucial idea in abstract algebra; there are corresponding theorems for many other algebraic structures. Indeed, we will encounter the corresponding theorem for groups in Chapter 33. The crisp, abstract formulation of this theorem contained in this chapter is very much an artifact of the 20th century axiomatic approach to algebra. The ideas behind this theory were encountered in many specific situations in the 19th century, but the sort of formulation we give here was not possible until the now accepted axiomatics for such algebraic structures as rings, groups, and fields were firmly in place. A lot of work and thought by many mathematicians went into these definitions. The French mathematician Camille Jordan was probably the first to consider clearly the notion of a quotient structure such as our ring of cosets R/I; he was working in the specific context of permutation groups. (See Chapters 29 and 30.) It is important for you to keep in mind that the abstract and efficient structure of modern algebra was not born overnight, but rather was the result of painstaking study of examples, from all over mathematics. The lesson of this history is clear: We should appreciate the generality of our theorems but must always return to specific examples to understand their purpose and application.

Chapter Summary

In this chapter we proved the *Fundamental Isomorphism Theorem for Commutative Rings*, which asserts that every onto homomorphism can be viewed as a natural homomorphism onto the ring modulo the kernel.

Warm-up Exercises

a. Why aren't $\mathbb{Z}_5$ and $\mathbb{Z}_7$ isomorphic?

b. Suppose that the field F is isomorphic to a ring R. Is R a field?

c. Suppose that the domain D is isomorphic to a ring R. Is R a domain?

d. Is every commutative ring isomorphic to a ring of cosets?

e. Suppose that $\varphi : R \to S$ is a ring homomorphism between commutative rings R and S. Is $R/\ker(\varphi)$ isomorphic to S? Be careful! What *is* true in this situation?

f. In what sense are the ideas of 'ideal' and 'homomorphism' equivalent?

g. Suppose that the elements of a commutative ring R are matrices, and R is isomorphic to the ring S. Are the elements of the ring S necessarily matrices? (If necessary, you should have a look at Exercise 16.7.)

h. Explain why $\varphi : \mathbb{Z}_5 \to \mathbb{Z}_5$, defined by $\varphi([a]_5) = [3a]_5$, is a one-to-one onto function that is not a ring isomorphism.

Exercises

1. Check that the homomorphisms given in Exercises 16.7 and 16.9 are in fact isomorphisms.

2. Show that the rings

$$\mathbb{Z}_8,\ \mathbb{Z}_4 \times \mathbb{Z}_2, \text{ and } \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$$

are all non-isomorphic, even though each of these rings has the same number of elements.

3. Prove that $\mathbb{Z}[i]/\langle 1+i\rangle$ is isomorphic to $\mathbb{Z}_2$, by defining a homomorphism from $\mathbb{Z}[i]$ onto $\mathbb{Z}_2$ whose kernel is $\langle 1+i\rangle$, and using the Fundamental Theorem 19.1. (Compare this to Exercise 18.3.)

4. Let $I = \{f \in C[0,1] : f(1/4) = 0\}$. Prove that $C[0,1]/I$ is isomorphic to $\mathbb{R}$ by defining the appropriate homomorphism from $C[0,1]$ onto $\mathbb{R}$ and using the Fundamental Theorem 19.1. (Compare this to Exercise 18.4.)

5. Let $I = \langle (3,4) \rangle \subseteq \mathbb{Z} \times \mathbb{Z}$. Prove that $(\mathbb{Z} \times \mathbb{Z})/I$ is isomorphic to $\mathbb{Z}_{12}$, by using a homomorphism of the form $\varphi(a,b) = [ax + by]_{12}$ (for some fixed x, y). (Compare this to Exercise 18.5.)

6. Use the Fundamental Theorem to prove that
$$(\mathbb{Q}[x] \times \mathbb{Z})/\langle (x,2) \rangle$$
is isomorphic to $\mathbb{Q} \times \mathbb{Z}_2$.

7. Let R be the set $\mathbb{Q} \times \mathbb{Q}$. We will equip this set with operations, other than the usual ones for the direct product, as described in Example 6.10. Namely, define the operations
$$(a,b) + (c,d) = (a + c, b + d)$$
(this is the usual addition), and
$$(a,b)(c,d) = (ac, ad + bc).$$
   (a) Prove from first principles that R is a ring.
   (b) Use the Fundamental Theorem to prove that R is isomorphic to $\mathbb{Q}[x]/\langle x^2 \rangle$.

8. Prove that $\mathbb{Z}_2 \times \mathbb{Z}_3$ is isomorphic to $\mathbb{Z}_6$ by showing that the homomorphism $\varphi([a]_2, [b]_3) = [3a + 4b]_6$ is onto and has zero kernel. (Or, simply that it is onto and $\mathbb{Z}_2 \times \mathbb{Z}_3$ and $\mathbb{Z}_6$ both have 6 elements.) What gets mapped to $[1]_6$? To $[2]_6$? $[3]_6$? $[4]_6$? $[5]_6$?

9. Let X be a set with n elements; consider the power set ring $P(X)$ of subsets of X (described in Exercise 6.20). Consider also the ring
$$\mathbb{Z}_2 \times \mathbb{Z}_2 \times \cdots \times \mathbb{Z}_2$$
of n-tuples whose entries are taken from $\mathbb{Z}_2$. Prove that these two rings are isomorphic.

10. Consider the ring $\mathbb{Z}[\alpha]$, described in Exercise 7.3. Prove that
$$\mathbb{Z}[\alpha]/\langle 1 + \alpha \rangle$$
is isomorphic to the ring $\mathbb{Z}_6$, by defining a homomorphism $\varphi : \mathbb{Z}[\alpha] \to \mathbb{Z}_6$ with the appropriate kernel. (Compare this to Exercise 18.7.)

11. Consider the homomorphism $\varphi : C \to \mathbb{R}$ given in Exercise 16.30. What two rings are isomorphic, according to the Fundamental Theorem?

12. Use the Fundamental Theorem to prove that $\mathbb{Q}[x]/\langle x^2 \rangle$ and
$$\left\{ \begin{pmatrix} a & b \\ 0 & a \end{pmatrix} \in M_2(\mathbb{Q}) : a,\ b \in \mathbb{Q} \right\}$$
are isomorphic. (Compare this to Exercise 7 above.)

13. In this problem you will prove that the fields $\mathbb{R}$ and $\mathbb{C}$ are *not* isomorphic. Suppose by way of contradiction that there exists a ring isomorphism $\varphi : \mathbb{C} \to \mathbb{R}$. Now answer the following questions (justifying your answers, of course): What is $\varphi(1)$? What is $\varphi(-1)$? What does this tell you about $\varphi(i)$?

14. In this problem you will prove that the fields $\mathbb{R}$ and $\mathbb{Q}$ are *not* isomorphic. Suppose by way of contradiction that there exists a ring isomorphism $\varphi : \mathbb{R} \to \mathbb{Q}$. Now answer the following questions (justifying your answers, of course): What is $\varphi(1)$? What is $\varphi(2)$? What does this tell you about $\varphi(\sqrt{2})$?

15. Use the ideas of Exercise 13 (or Exercise 14) to prove that $\mathbb{Q}$ and $\mathbb{C}$ are not isomorphic.

16. If you have encountered the ideas of *countably infinite* and *uncountably infinite* sets, another proof of Exercise 14 (and 15) is possible. What is it?

17. Suppose that rings R and S are isomorphic, via the isomorphism $\varphi : R \to S$.
    (a) Show that the rings $R[x]$ and $S[x]$ are isomorphic, by defining a ring isomorphism $\bar{\varphi} : R[x] \to S[x]$ which extends φ; by this we mean that if $r \in R$, then $\varphi(r) = \bar{\varphi}(r)$.
    (b) Suppose now that the rings R and S are fields. Prove that $r \in R$ is a root of $f \in R[x]$ if and only if $\varphi(r)$ is a root of $\bar{\varphi}(f)$.
    (c) Suppose again that R and S are fields. Prove that $f \in R[x]$ is irreducible if and only if $\bar{\varphi}(f) \in S[x]$ is irreducible.

18. Suppose that the commutative rings R and S are isomorphic, via the isomorphism $\varphi : R \to S$. Let $r \in R$. Use the Fundamental Isomorphism Theorem to prove that the rings $R/\langle r\rangle$ and $S/\langle\varphi(r)\rangle$ are isomorphic.

19. Extend Exercise 18. Suppose that the commutative rings R and S are isomorphic via the isomorphism $\varphi : R \to S$. Let I be an ideal of R; by Exercise 16.31 we know that $\varphi(I)$ is an ideal of S. Prove that the rings R/I and $S/\varphi(I)$ are isomorphic.

20. Suppose that the commutative rings R and S are isomorphic via the isomorphism $\varphi : R \to S$. Combine Exercises 17 and 19 to conclude that if $f \in R[x]$, then the rings $R[x]/\langle f\rangle$ and $S[x]/\langle\bar{\varphi}(f)\rangle$ are isomorphic.

21. Suppose that R is a commutative ring, with ideals A and I, where $I \subseteq A$.

    (a) Prove that A/I is an ideal in the ring R/I.

    (b) Prove that $(R/I)/(A/I)$ is isomorphic to R/A.

22. Let $R = C[0,1]$, $I = \{f \in C[0,1] : f(0) = f(1) = 1\}$ and $A = \{f \in C[0,1] : f(0) = 1\}$. Check that R, A, I satisfy the hypotheses of Exercise 21. Then exhibit explicitly the isomorphism given by that exercise.

23. Let R be a commutative ring with ideals I and J. Then $I \cap J$ and $I + J$ are also ideals of R. (See Exercises 11.12 and 11.14.) Furthermore, $I \cap J$ is an ideal of I and J is an ideal of $I + J$. (See Exercise 11.13.) Prove that the rings $I/(I \cap J)$ and $(I + J)/J$ are isomorphic.

24. Let R be $\mathbb{Z}$, $I = \langle 12\rangle$ and $J = \langle 8\rangle$. Check that R, I, J satisfy the hypotheses of Exercise 23. Then exhibit explicitly the isomorphism given by that exercise.

Chapter 20

Maximal and Prime Ideals

Let's return to Example 19.10. We saw that the ring homomorphism (the evaluation homomorphism) $\varphi : \mathbb{R}[x] \to \mathbb{C}$ given by $\varphi(f) = f(i)$ has kernel $\langle x^2 + 1\rangle$. This homomorphism was onto, and so $\mathbb{R}[x]/\langle x^2 + 1\rangle$ is isomorphic to the *field* $\mathbb{C}$. We would like to find out what sort of ideal leads to a ring of cosets that is a field.

20.1 Maximal Ideals

Now $x^2 + 1$ is an irreducible element of $\mathbb{R}[x]$ (that is, it is a polynomial that cannot be further factored over the reals). Readers of Chapter 13 may recall that in a principal ideal domain, an element is irreducible if and only if the corresponding principal ideal is *maximal* (Theorem 13.3c). Thus, we conclude that $\langle x^2 + 1\rangle$ is a maximal ideal of $\mathbb{R}[x]$. For your convenience we review the definition of maximal ideal as follows: An ideal I of a ring R is **maximal** if the only ideal of R properly containing I is R itself.

Let's prove again that $\langle x^2 + 1\rangle$ is a maximal ideal, without using results from Chapter 13. We assume that I is an ideal of $\mathbb{R}[x]$ that properly contains $\langle x^2 + 1\rangle$; we must show that in fact $I = \mathbb{R}[x]$. Because I properly contains $\langle x^2 + 1\rangle$, I contains a polynomial p that is not a multiple of $x^2 + 1$. Because p is not a multiple of $x^2 + 1$, we have that

$$\langle x^2 + 1\rangle \neq \langle x^2 + 1\rangle + p.$$

But the ring of cosets of $\langle x^2 + 1\rangle$ is isomorphic to $\mathbb{C}$, which is a field. So, $\langle x^2 + 1\rangle + p$ has a multiplicative inverse, and so there exists a polynomial q in $\mathbb{R}[x]$ such that

$$(\langle x^2 + 1\rangle + p)(\langle x^2 + 1\rangle + q) = \langle x^2 + 1\rangle + pq = \langle x^2 + 1\rangle + 1.$$

But then

$$\langle x^2+1\rangle = \langle x^2+1\rangle + 1 - pq$$

and so $1-pq=r$ for some $r \in \langle x^2+1\rangle$. That is, $1 = pq+r$. Now $p \in I$ and $r \in \langle x^2+1\rangle \subseteq I$, and so $1 = pq + r \in I$ (because I is an ideal). But an ideal that contains 1 is the entire ring; thus, we have shown that the only ideal that properly contains $\langle x^2+1\rangle$ is $\mathbb{R}[x]$. We have thus proved (again) that $\langle x^2+1\rangle$ is a maximal ideal of $\mathbb{R}[x]$.

Example 20.1

For another example, consider the residue homomorphism from $\mathbb{Z}$ onto $\mathbb{Z}_p$, where p is prime. Once again, the homomorphism is onto a field, and the kernel (in this case, $\langle p\rangle$) is a maximal ideal. (See Exercise 20.3.) Of course, we could also conclude this by noting that p is irreducible and use Theorem 13.3c.

These examples suggest the following general theorem:

Theorem 20.1 *Let R be a commutative ring with unity. Then M is a maximal ideal of R if and only if R/M is a field.*

Proof: Let R be a commutative ring with unity, and suppose first that M is a maximal ideal. To show that R/M is a field, we consider any $M + a \in R/M$, with $M + a \neq M + 0$; we must show that $M + a$ has a multiplicative inverse. But because $M + a \neq M$, we know that $a \notin M$. Consider

$$\langle M, a\rangle = \{m + ab : m \in M, b \in R\};$$

we are using the notation $\langle M, a\rangle$ to suggest that this is in fact the smallest ideal of R that contains both M and a. In Exercise 20.1, you prove exactly this.

Furthermore, $\langle M, a\rangle$ properly contains M, because $a \in \langle M, a\rangle$ but $a \notin M$. Thus, because M is maximal, $\langle M, a\rangle = R$. So there exist $b \in R$ and $m \in M$ such that $m + ab = 1$. But then

$$M + 1 = M + (m + ab) = M + ab = (M + a)(M + b),$$

and so $M + b$ is the required multiplicative inverse.

Conversely, suppose that R/M is a field. Let I be an ideal with $R \supseteq I \supset M$; we must show that $I = R$. Choose r, an element in I but not in M. Because R/M is a field, there exists $s \in R$ such that $M + rs = M + 1$. That is, $M = M + 1 - rs$ or $1 - rs = m$ for some m in M. But $1 = rs + m$ implies $1 \in I$, because $r \in I$ and $m \in I$. But $1 \in I$ implies that $I = R$. Hence, M is maximal. □

Example 20.2

An important if relatively trivial example of this theorem occurs in case R itself is a field. Because all non-zero elements are units, we've observed before that the only proper ideal a field has is $\{0\}$; hence, $\{0\}$ is (practically by default) a *maximal* ideal, and so the theorem applies. That is, the ring of cosets $R/\{0\}$ is a field; but of course this ring is just (isomorphic to) R itself. But we can also apply the converse: If R is a commutative ring with unity without ideals (other than $\{0\}$ and R), then R is necessarily a field. (Of course, we've seen this result before as the Corollary 11.4.)

Example 20.3

In Example 19.10 we proved that $\mathbb{R}[x]/\langle x^2+1\rangle$ is isomorphic to the field $\mathbb{C}$. This means that $\langle x^2+1\rangle$ is a maximal ideal in $\mathbb{R}[x]$.

Example 20.4

In Example 19.11 we proved that $\mathbb{Z}[\sqrt{-5}]/\langle 3, 1+\sqrt{-5}\rangle$ is isomorphic to the field $\mathbb{Z}_3$. This means that $\langle 3, 1+\sqrt{-5}\rangle$ is a maximal ideal of $\mathbb{Z}[\sqrt{-5}]$. Note we proved the maximality of this ideal directly in Section 13.4.

Example 20.5

Consider the irreducible polynomial x^2-2 in $\mathbb{Q}[x]$ (note that this is *not* irreducible in $\mathbb{R}[x]$). Readers of Chapter 13 know that $\langle x^2-2\rangle$ is a maximal ideal, because $\mathbb{Q}[x]$ is a PID and by Theorem 13.3 principal ideals generated by irreducibles are maximal. (Alternatively, it is possible to prove this directly, using an argument like that we provided earlier in this chapter, for $\langle x^2+1\rangle$ in $\mathbb{R}[x]$; see Exercise 20.13.) Thus, by Theorem 20.1, we know that $\mathbb{Q}[x]/\langle x^2-2\rangle$ is a field. But what is this field? Consider the evaluation homomorphism $\varphi : \mathbb{Q}[x] \to \mathbb{R}$, given by

$\varphi(f) = f(\sqrt{2})$. Quite evidently, the kernel of this homomorphism is exactly $\langle x^2 - 2\rangle$. But it seems clear that φ is not onto $\mathbb{R}$. For example, $\sqrt{3} \notin \varphi(\mathbb{Q}[x])$: We cannot obtain $\sqrt{3}$ by plugging $\sqrt{2}$ into a rational polynomial (this statement is probably plausible, but we won't prove it here — see Example 38.2). But if φ is not onto, we cannot directly apply the Fundamental Isomorphism Theorem 19.1. What the Fundamental Isomorphism Theorem does tell us is that $\mathbb{Q}[x]/\langle x^2 - 2\rangle$ is isomorphic to $\varphi(\mathbb{Q}[x])$, a proper subfield of the field $\mathbb{R}$.

Furthermore, this field $\mathbb{Q}[x]/\langle x^2 - 2\rangle$ contains a subring isomorphic to $\mathbb{Q}$: Consider the function

$$\iota : \mathbb{Q} \to \mathbb{Q}[x]/\langle x^2 - 2\rangle$$

defined by $\iota(q) = \langle x^2 - 2\rangle + q$.

▷ **Quick Exercise.** Check that this is a one-to-one homomorphism (which is *not* onto). ◁

Thus, $\iota(\mathbb{Q})$ is the isomorphic copy of $\mathbb{Q}$ contained in the field $\mathbb{Q}[x]/\langle x^2-2\rangle$. This suggests that we can use the previous theorem to build bigger fields. In this case we seem to have constructed a field strictly between $\mathbb{Q}$ and $\mathbb{R}$. We'll have a lot to say about this in Chapters 42 and 43.

20.2 Prime Ideals

We now consider a class of ideals closely related to the class of maximal ideals. A proper ideal I of a commutative ring with unity is a **prime ideal** if whenever $ab \in I$, then either $a \in I$ or $b \in I$. Let's look at an example of a prime ideal.

Example 20.6

Consider the ideal $\langle x\rangle$ in $\mathbb{Z}[x]$. We claim that it is prime: For if $pq \in \langle x\rangle$, then pq is a polynomial with no constant term. But the constant term of a product of polynomials is the product of their constant terms, and so this means that at least one of p and q must also lack a constant term; that is, at least one of p and q belongs to $\langle x\rangle$. This is what we wished to show.

Thus, $\langle x\rangle$ is a prime ideal in $\mathbb{Z}[x]$. Readers of Chapter 13 may recall that this is *not* a maximal ideal; we can use the theorem above to verify this. To do that, we need to find a homomorphism defined on $\mathbb{Z}[x]$ whose kernel is $\langle x\rangle$. Reflection on our earlier examples might suggest another evaluation homomorphism, this time at zero: If $\psi(f) = f(0)$, then $\ker(\psi) = \langle x\rangle$. Of course, ψ is here a homomorphism onto $\mathbb{Z}$, and so $\mathbb{Z}[x]/\langle x\rangle$ is isomorphic to $\mathbb{Z}$. Thus, we conclude that $\langle x\rangle$ is *not* a maximal ideal (because $\mathbb{Z}$ is *not* a field). More directly, we can conclude that $\langle x\rangle$ is not a maximal ideal, by noting that

$$\langle x\rangle \subset \langle 2, x\rangle \subset \mathbb{Z}[x].$$

However, the homomorphic image $\mathbb{Z}$ is at least a *domain*; this observation suggests the following theorem:

Theorem 20.2 *Let R be a commutative ring with unity. Then P is a prime ideal if and only if R/P is an integral domain.*

Proof: Let R be a commutative ring with unity, and suppose first that P is a prime ideal. To show that R/P is a domain (because it obviously has unity) we need only show that it has no zero divisors. So suppose that $(P + a)(P + b) = P + 0$. But then $ab \in P$, and so because P is prime a or b belongs to P, and so $P + a = P + 0$ or $P + b = P + 0$, as required. In the Quick Exercise below you will do the converse argument, which is essentially the reverse of what we've just done. □

▷ **Quick Exercise.** Prove the converse of the above theorem. ◁

Note in particular the special case when R itself is a domain. We conclude that a ring with unity is a domain if and only if $\{0\}$ is a prime ideal.

Because all fields are domains, we can conclude that maximal ideals are necessarily prime.

For principal ideals, determination of whether an ideal is prime reduces to determining whether the generator itself is a prime element (see Chapter 13 for a definition and discussion of prime elements):

Theorem 20.3 *Let R be a commutative ring and $0 \neq a \in R$. Then $\langle a\rangle$ is a prime ideal if and only if a is a prime element.*

Proof: We have left this easy proof as Exercise 20.2. □

Example 20.7

> As an example of Theorem 20.2, consider (from Example 16.3) the ring $\mathbb{Z} \times \mathbb{Z}$ and the homomorphism $\pi(x, y) = y$, which is onto $\mathbb{Z}$. Because $\mathbb{Z}$ is a domain, the kernel $\mathbb{Z} \times \{0\}$ is a prime ideal of $\mathbb{Z} \times \mathbb{Z}$ (that is also not maximal).

This example is worth generalizing. Suppose

$$R_1, R_2, \cdots, R_n$$

are rings and $\prod R_i = R_1 \times R_2 \times \cdots \times R_n$ is their n-fold direct product. Recall that

$$\prod R_i = \{(r_1, r_2, \cdots, r_n) : r_i \in R_i, \text{for } i = 1, 2, \cdots, n\}.$$

(See Exercise 6.15.) Now for each j consider the function

$$\pi_j : \prod R_i \to R_j$$

defined by $\pi_j(r_1, r_2, \cdots, r_n) = r_j$. This map is called the *jth projection*. Because of the definition of the ring operations on $\prod R_i$ (which are componentwise) it is obvious that π_j is an onto homomorphism; furthermore, its kernel is

$$R_1 \times R_2 \times \cdots \times R_{j-1} \times \{0\} \times R_{j+1} \times \cdots \times R_n.$$

Note that if R_j is a domain, then this kernel is a prime ideal. Similarly, if R_j is a field, then this kernel is a maximal ideal.

$\triangleright$ **Quick Exercise.** Why? $\triangleleft$

Even if all the R_i are domains, then $\prod R_i$ certainly is not: For

$$(a_1, a_2, \cdots, a_n)(b_1, b_2, \cdots, b_n) = (0, 0, \cdots, 0)$$

if and only if $a_i b_i = 0$ for each i. But because the R_i are domains, either a_i or b_i is 0. We thus have the product in $\prod R_i$ being zero if and only if the set $\{i : a_i \neq 0\}$ is disjoint from the set $\{j : b_j \neq 0\}$. For example, $(3, 1, 0, 0)(0, 0, 0, 5) = (0, 0, 0, 0)$. Thus, although $\prod R_i$ is not a domain, determination of which elements are zero divisors is certainly an easy matter. This means that representing a given ring as (isomorphic) to a product of rings is often a real help toward understanding the arithmetic of the ring. We will see a dramatic and famous example of this in the next chapter.

Chapter Summary

In this chapter we discussed *maximal ideals* and showed that an ideal I of a commutative ring R is maximal if and only if R/I is a field. Similarly, we discussed *prime ideals* and showed that an ideal I is prime if and only if R/I is a domain.

Warm-up Exercises

a. Give examples of a commutative ring with unity and an ideal satisfying the following (or argue that no such example exists):

   (a) the ideal is prime but not maximal.

   (b) the ideal is maximal but not prime.

   (c) the ideal is prime and the ring has zero divisors.

b. Explain why $\langle x^2 - 2 \rangle$ is a maximal ideal of $\mathbb{Q}[x]$, but not of $\mathbb{R}[x]$.

c. Explain why $\langle x - 2 \rangle$ is a prime ideal of $\mathbb{Z}[x]$ but is not maximal.

d. Give the quickest possible proof that a maximal ideal is prime.

e. What are the zero divisors of $\mathbb{Z} \times \mathbb{Z}$? What about of $\mathbb{Z}_4 \times \mathbb{Z}$?

f. Let R be a commutative ring with unity. Under what circumstances is $\{0\}$ a prime ideal? A maximal ideal?

g. Let R be a commutative ring with unity. Why isn't the ideal R maximal or prime?

Exercises

1. Let R be a commutative ring with unity, I an ideal of R, and $a \in R$. Define
$$\langle I, a \rangle = \{k + ab : \ k \in I, \ b \in R\}.$$
Prove that $\langle I, a \rangle$ is an ideal. Then show that $I \subseteq \langle I, a \rangle$ and $a \in \langle I, a \rangle$. Furthermore, show that $\langle I, a \rangle$ is the smallest ideal of R that contains both I and a. (This result is needed in the proof of Theorem 20.1.)

2. Prove Theorem 20.3. That is, suppose the R is a commutative ring, and $0 \neq a \in R$. Prove that $\langle a \rangle$ is a prime ideal if and only if a is a prime element.

3. Let p be a prime integer in $\mathbb{Z}$. Prove directly that $\langle p \rangle$ is a maximal ideal.

4. Find all prime ideals of $\mathbb{Z}$. Find all maximal ideals of $\mathbb{Z}$.

5. Find all prime ideals of $\mathbb{Z}_{20}$. Find all maximal ideals of $\mathbb{Z}_{20}$.

6. Find all prime ideals of $\mathbb{Z} \times \mathbb{Z}$. Find all maximal ideals of $\mathbb{Z} \times \mathbb{Z}$.

7. Find all prime ideals of $\mathbb{Z}_2 \times \mathbb{Z}_3$; find all maximal ideals of $\mathbb{Z}_2 \times \mathbb{Z}_3$. Now do the same for $\mathbb{Z}_2 \times \mathbb{Z}_4$.

8. Find all prime ideals of $\mathbb{Q}$ and $\mathbb{Q} \times \mathbb{Q}$; do likewise for all maximal ideals.

9. Consider $\mathbb{Q}[x, y]$, the set of all polynomials with coefficients from $\mathbb{Q}$, in the two indeterminates x and y. In Exercise 12.12, you considered this ring and showed that it is not a PID.

   (a) Consider the function
$$\psi : \mathbb{Q}[x, y] \to \mathbb{Q}$$
defined by $\psi(f) = f(0, 0)$. Why is this a ring homomorphism? Why is the kernel of ψ a maximal ideal? Determine explicitly a nice description of the kernel.

   (b) Consider the function
$$\rho : \mathbb{Q}[x, y] \to \mathbb{Q}[x]$$
defined by $\rho(f) = f(x, 0)$. Why is this a ring homomorphism? Why is the kernel of ρ a prime ideal? Determine explicitly a nice description of the kernel. Find a larger proper ideal, thus showing concretely that this ideal is not maximal.

   (c) Consider the function
$$\varphi : \mathbb{Q}[x, y] \to \mathbb{Q}[x]$$
defined by $\varphi(f) = f(x, x)$. Prove that this is a ring homomorphism. Why is the kernel of φ prime but not maximal? Determine explicitly a nice description of the kernel. Find a larger proper ideal, thus showing concretely that this ideal is not maximal.

10. Consider the function
$$\varphi : \mathbb{Z} \to \mathbb{Z} \times \mathbb{Z}$$
defined by $\varphi(n) = (n, n)$. Check that this is a homomorphism. Now, $\ker(\varphi) = \{0\}$. Argue directly from the definition that $\{0\}$ is a prime ideal of $\mathbb{Z}$. Now $\mathbb{Z} \times \mathbb{Z}$ is obviously not a domain; why does this not contradict the theorem in the text characterizing prime ideals?

11. Consider the function
$$\varphi : \mathbb{Q}[x] \to \mathbb{R}$$
defined by $\varphi(f) = f(\sqrt{2})$. Check that this is a homomorphism.

   (a) Describe the elements of the ring $\varphi(\mathbb{Q}[x])$.

   (b) This ring is clearly a domain (because it is a subring of $\mathbb{R}$). Is it a field?

   (c) What is the kernel of this homomorphism? Is it prime or maximal?

   (d) How could we reach the conclusion for part c, using an argument from Chapter 13?

12. Consider the function
$$\varphi : \mathbb{Z}[x] \to \mathbb{Z}_3[x]$$
defined by
$$\varphi(a_0 + a_1x + \cdots + a_nx^n) = [a_0] + [a_1]x + \cdots + [a_n]x^n.$$
   (a) Prove that this is an onto ring homomorphism.
   (b) Why is the kernel of φ a prime ideal? Give an explicit description of its kernel (it is a principal ideal).

13. Consider the ideal $\langle x^2 - 2\rangle$ in $\mathbb{Q}[x]$, discussed in Example 20.5. Prove that this is a maximal ideal, without using either Chapter 13 or Theorem 20.1.

14. Consider the function $\varphi : \mathbb{Z}[x] \to \mathbb{Q}$ defined by $\varphi(f) = f(-1/2)$.
   (a) What is the kernel of φ?
   (b) We know that $\mathbb{Q}$ is a field. Does this mean that $\ker(\varphi)$ is a maximal ideal? Be careful!
   (c) Give a nice description of the elements $\varphi(\mathbb{Z}[x])$.

15. Consider the homomorphism discussed in Exercise 16.30 and Exercise 19.11. What ideal do we now know is maximal? Prove that this ideal is maximal directly, from the definition.

16. Consider the homomorphism discussed in Exercise 16.10 and Exercise 19.9. What ideal do we now know is maximal? Prove that this ideal is maximal directly, from the definition.

17. Let R be a commutative ring with unity, and I a prime ideal of R. Note that $I[x]$ is a subring of $R[x]$. Prove that $I[x]$ is a prime ideal of $R[x]$. Now suppose that I is a maximal ideal; show by example that $I[x]$ need not be a maximal ideal.

18. Let R and S be commutative rings. Prove that if the direct product $R \times S$ is a domain, then exactly one of R and S is the zero ring.

19. It turns out that in a commutative ring with unity, every proper ideal is a subset of a maximal ideal. (The proof of this theorem requires more sophisticated set theory than we wish to enquire into here.) Use this theorem to establish the following:

   (a) Prove that every commutative ring with unity has a field as a homomorphic image.
   (b) If the commutative ring R with unity has a unique maximal ideal M, then M consists of exactly the set of non-units.

20. Let P be an ideal of a commutative ring with unity. Prove that P is a prime ideal if and only if, whenever $P \supseteq I \cdot J$, where I and J are ideals, then $P \supseteq I$ or $P \supseteq J$. (For the definition of the product of two ideals, see Exercise 11.15.)

21. Prove that in a PID, a non-zero proper ideal is prime if and only if it is maximal.

22. Prove that in a finite commutative ring with unity, a proper ideal is prime if and only if it is maximal.

23. Let R be a commutative ring with unity, for which every element is an idempotent; such a ring is called **Boolean**. (See Exercise 7.25 for more about idempotents.) Prove that in R, a proper ideal is prime if and only if it is maximal.

Chapter 21

The Chinese Remainder Theorem

We can cast considerable light on the arithmetic of the rings $\mathbb{Z}_m$, by making use of the theory of the previous chapters. We will in the process encounter a famous and ancient result known as the Chinese Remainder Theorem.

21.1 Direct Products of Domains

Let's begin by looking at an example.

Example 21.1

Consider the ring $\mathbb{Z}_6$. The function $\varphi : \mathbb{Z}_6 \to \mathbb{Z}_3$ defined by $\varphi([a]_6) = [a]_3$ obviously preserves addition and multiplication and hence is a homomorphism, if it is well defined. But if $[a]_6 = [b]_6$, then $6|(a-b)$, and so $3|(a-b)$; thus, $\varphi([a]_6) = [a]_3 = [b]_3 = \varphi([b]_6)$ and so φ is well defined. Note also that φ is onto. Because $\mathbb{Z}_3$ is a field, we have that $\ker(\varphi) = \langle[3]\rangle$ is a maximal ideal.

Now the corresponding function $\psi : \mathbb{Z}_6 \to \mathbb{Z}_2$ is also a homomorphism, and its kernel $\langle[2]\rangle$ is also a maximal ideal.

Let's put these two homomorphisms together, using the idea of direct product. Namely, define $\mu : \mathbb{Z}_6 \to \mathbb{Z}_3 \times \mathbb{Z}_2$ by setting

$$\mu([a]_6) = (\varphi([a]_6), \psi([a]_6) = ([a]_3, [a]_2).$$

It is easy to see that this is a homomorphism, because φ and ψ are.

▷ **Quick Exercise.** Check this. ◁

Because the zero element of the direct product $\mathbb{Z}_3 \times \mathbb{Z}_2$ is the element $([0],[0])$, an element of $\mathbb{Z}_6$ belongs to the kernel of μ only if it belongs to the kernels of *both* φ and ψ. That is, the kernel of μ is

$$\langle[3]\rangle \cap \langle[2]\rangle = \{[0],[3]\} \cap \{[0],[2],[4]\} = \{[0]\}.$$

Thus, μ is actually a one-to-one homomorphism (Corollary 17.4). Let's write out the homomorphism μ explicitly:

$$\begin{aligned}
[0]_6 &\longmapsto ([0]_2, [0]_3)\\
[1]_6 &\longmapsto ([1]_2, [1]_3)\\
[2]_6 &\longmapsto ([0]_2, [2]_3)\\
[3]_6 &\longmapsto ([1]_2, [0]_3)\\
[4]_6 &\longmapsto ([0]_2, [1]_3)\\
[5]_6 &\longmapsto ([1]_2, [2]_3)
\end{aligned}$$

Notice that this function is actually onto, and so $\mathbb{Z}_6$ is isomorphic to $\mathbb{Z}_3 \times \mathbb{Z}_2$. We have represented $\mathbb{Z}_6$ as a direct product of simpler pieces. And because those simpler pieces $\mathbb{Z}_3$ and $\mathbb{Z}_2$ are domains (and in fact fields), the zero divisors in this ring are easy to determine: They are those elements with a zero in at least one component.

▷ **Quick Exercise.** Check that this analysis coincides with what you already know about which elements of $\mathbb{Z}_6$ are zero divisors. ◁

In Example 21.1, our ability to obtain an isomorphism between $\mathbb{Z}_6$ and a direct product of two domains (and, in fact, fields), depended on the fact that this ring has two prime (and, in fact, maximal) ideals, whose intersection is zero. Let's examine this situation in a more general context.

Suppose that the commutative ring R is isomorphic to a direct product of n domains $D_1, D_2, \cdots D_n$, via the isomorphism

$$\mu : R \to D_1 \times D_2 \times \cdots \times D_n.$$

Consider the n projection homomorphisms

$$\pi_i : D_1 \times D_2 \times \cdots \times D_n \to D_i,$$

as discussed in Section 20.2. The composition $\pi_i \circ \mu$ is a ring homomorphism from R onto the domain D_i, and as such its kernel is a prime ideal (Theorem 20.2). Notice that if an element of R belongs to the kernels of all these maps, then the element is sent by μ to the zero element of the direct product. That is, such an element belongs to the kernel of μ. But μ is one-to-one, and so its kernel is the zero ideal. This means that R possesses a collection of n prime ideals whose intersection is zero. We record the result of this argument in the following theorem:

Theorem 21.1 *If a commutative ring with unity is isomorphic to a direct product of finitely many integral domains, then it has a finite collection of prime ideals, whose intersection is the zero ideal.*

In Exercises 21.9-21.11 we will inquire into the possibility of a converse for this theorem.

Example 21.2

Example 21.1 provides an example for this theorem, because $\mathbb{Z}_6$ is isomorphic to a product of two domains and had two prime ideals whose intersection is the zero ideal.

Example 21.3

Consider the direct product of domains $\mathbb{Z} \times \mathbb{Z}_3 \times \mathbb{Q}$. It has three prime ideals whose intersection is zero, namely,

$$\{0\} \times \mathbb{Z}_3 \times \mathbb{Q}, \ \mathbb{Z} \times \{0\} \times \mathbb{Q}, \text{ and } \mathbb{Z} \times \mathbb{Z}_3 \times \{0\}.$$

Example 21.4

Now consider $\mathbb{Z}_{12}$. Here, our representation is not so nice. In Exercise 21.3 you will show that the only prime ideals of $\mathbb{Z}_{12}$ are

$$\langle[2]\rangle = \{[0],[2],[4],[6],[8],[10]\} \text{ and } \langle[3]\rangle = \{[0],[3],[6],[9]\}.$$

Notice that

$$\langle[2]\rangle \cap \langle[3]\rangle = \{[0],[6]\} \supset \{[0]\}.$$

This means that it is *not* possible to find a homomorphism from $\mathbb{Z}_{12}$ onto a direct product of domains, as we were able to for $\mathbb{Z}_6$ in Example 21.1. The difference turns out to be this: 6 does not have any repeated prime factors, while 12 does. We can show that $\mathbb{Z}_{12}$ is isomorphic to $\mathbb{Z}_4 \times \mathbb{Z}_3$ (of course, $\mathbb{Z}_4$ is not a field). Consider the obvious residue homomorphisms $\mathbb{Z}_{12} \to \mathbb{Z}_4$ and $\mathbb{Z}_{12} \to \mathbb{Z}_3$: They have kernels $\{[0],[4],[8]\}$ and $\{[0],[3],[6],[9]\}$, and the intersection of these ideals is zero.

▷ **Quick Exercise.** Write down the two residue homomorphisms explicitly, verify that their kernels are as stated, and then write out the isomorphism between $\mathbb{Z}_{12}$ and $\mathbb{Z}_4 \times \mathbb{Z}_3$ explicitly. ◁

Now, what are the zero divisors of $\mathbb{Z}_4 \times \mathbb{Z}_3$? They must consist of elements that are either zero in exactly one component, or else zero in the second component and a zero divisor of $\mathbb{Z}_4$ in the first component. These elements must correspond under our isomorphism to the zero divisors of $\mathbb{Z}_{12}$.

▷ **Quick Exercise.** Check this explicitly. ◁

21.2 Chinese Remainder Theorem

We now prove the general theorem that explains the representation of $\mathbb{Z}_6$ and $\mathbb{Z}_{12}$ as direct products, which we obtained in Examples 21.1 and 21.4 above.

Theorem 21.2 *Let $p_1, p_2, \cdots, p_n$ be distinct prime integers and $m = p_1^{k_1} p_2^{k_2} \cdots p_n^{k_n}$. Then $\mathbb{Z}_m$ is isomorphic to*

$$\mathbb{Z}_{p_1^{k_1}} \times \mathbb{Z}_{p_2^{k_2}} \times \cdots \times \mathbb{Z}_{p_n^{k_n}}.$$

Proof: Consider the function $\varphi_i : \mathbb{Z}_m \to \mathbb{Z}_{p_i^{k_i}}$ defined by $\varphi_i([a]_m) = [a]_{p_i^{k_i}}$; this is well-defined because if $[a]_m = [b]_m$, then $m|(a-b)$, and so $p_i^{k_i}|(a-b)$. It is clearly a homomorphism with kernel $\langle [p_i^{k_i}] \rangle$. Now define the function

$$\mu : \mathbb{Z}_m \to \mathbb{Z}_{p_1^{k_1}} \times \mathbb{Z}_{p_2^{k_2}} \times \cdots \times \mathbb{Z}_{p_n^{k_n}}$$

by setting

$$\mu([a]) = (\varphi_1([a]), \cdots, \varphi_n([a])).$$

This is evidently a ring homomorphism.

▷ **Quick Exercise.** Check this. ◁

Now, its kernel is

$$\bigcap_{i=1}^{n} \langle [p_i^{k_i}] \rangle = \{[a] : p_i^{k_i}|a, \text{for all } i\} = \{[0]\}.$$

This means that the homomorphism μ is one-to-one.

To show that μ is an isomorphism, it remains to check that it is onto. We will actually do this twice, because both proofs are illuminating:

Existential Proof: Now $\mathbb{Z}_m$ and $\mathbb{Z}_{p_1^{k_1}} \times \mathbb{Z}_{p_2^{k_2}} \times \cdots \times \mathbb{Z}_{p_n^{k_n}}$ are both finite sets with m elements; thus, because μ is one-to-one, it *must* be onto.

Constructive Proof: Given

$$([a_1], [a_2], \cdots, [a_n]) \in \mathbb{Z}_{p_1^{k_1}} \times \mathbb{Z}_{p_2^{k_2}} \times \cdots \times \mathbb{Z}_{p_n^{k_n}},$$

we must find $[a] \in \mathbb{Z}_m$ such that $\varphi_i([a]) = [a_i]$, for all i. Let

$$m_i = \frac{m}{p_i^{k_i}};$$

because m_i and p_i are relatively prime, by the GCD identity there exist integers x_i and y_i with $x_i m_i + y_i p_i^{k_i} = 1$. But then

$$[x_i m_i]_{p_i^{k_i}} = [1].$$

Now let

$$a = x_1 m_1 a_1 + x_2 m_2 a_2 + \cdots + x_n m_n a_n.$$

We claim that $\varphi_i([a]) = [a_i]$, for all i. But $p_i^{k_i}|m_j$ for all $j \neq i$, and so

$$\varphi_i([a]) = [a]_{p_i^{k_i}} = [x_i m_i a_i] = [x_i m_i][a_i] = [1][a_i] = [a_i],$$

as required. □

The number theory version of this theorem is known as the Chinese Remainder Theorem, because an example of the sort of number theory problem it solves first appears in a work by the Chinese mathematician Sun Tsu, in about the third century AD. We now restate the theorem in its number theoretic guise:

Theorem 21.3 The Chinese Remainder Theorem *Let $p_1, p_2, \cdots, p_n$ be distinct prime integers and $m = p_1^{k_1} p_2^{k_2} \cdots p_n^{k_n}$. Then the set of congruences*

$$\begin{aligned} x &\equiv a_1 \pmod{p_1^{k_1}} \\ x &\equiv a_2 \pmod{p_2^{k_2}} \\ &\cdots \\ x &\equiv a_n \pmod{p_n^{k_n}} \end{aligned}$$

has a simultaneous solution, which is unique modulo m.

Proof: Consider the element

$$([a_1], [a_2], \cdots, [a_n]) \in \mathbb{Z}_{p_1^{k_1}} \times \mathbb{Z}_{p_2^{k_2}} \times \cdots \times \mathbb{Z}_{p_n^{k_n}}.$$

Because the function μ in the previous proof is onto, there exists $[a] \in \mathbb{Z}_m$ so that $\varphi_i([a]) = [a_i]$, for all i. Then a simultaneously satisfies the n congruences in the theorem.

▷ **Quick Exercise.** Why? ◁ □

Example 21.5

Consider the three congruences $x \equiv 4 \pmod 7$, $x \equiv 1 \pmod 2$, and $x \equiv 3 \pmod 5$; the theorem asserts that they have a simultaneous solution modulo 70. In fact, we can construct the solution by following the constructive proof above. In this case $m_1 = 10$, $m_2 = 35$, and $m_3 = 14$. Then $x_1 = 5$ because

$$[5]_7[10]_7 = [5]_7[3]_7 = [1]_7.$$

In a similar fashion $x_2 = 1$ and $x_3 = 4$. Then

$$a = (5)(10)(4) + (1)(35)(1) + (4)(14)(3) = 403.$$

But $403 \equiv 53 \pmod{70}$, and so our simultaneous solution of the three congruences is 53.

▷ **Quick Exercise.** Check directly that 53 satisfies these three congruences. ◁

Alternatively, we can view this example in light of our first version of the Chinese Remainder Theorem (Theorem 21.2). From that version of the theorem, we know that $\mathbb{Z}_7 \times \mathbb{Z}_2 \times \mathbb{Z}_5$ is isomorphic to $\mathbb{Z}_{70}$. The isomorphism takes the element $(4, 1, 3)$ to 53.

Note that if all the primes in the factorization of m occur only once, we then have that $\mathbb{Z}_m$ is a direct product of finitely many domains (in fact, fields). But suppose that at least one prime divisor of m occurs with degree at least two. We claim that $\mathbb{Z}_m$ cannot be a direct product of domains; this amounts to claiming that the intersection of all the prime ideals of $\mathbb{Z}_m$ is not zero.

To show that this is true, we must first convince ourselves that $\mathbb{Z}_m$ even has any prime ideals. But because $\mathbb{Z}_m$ is a finite ring, it has only finitely many ideals; consequently, it certainly has at least one maximal ideal (which is necessarily a prime).

To show that the intersection of all the prime ideals of $\mathbb{Z}_m$ is not zero, choose an arbitrary prime ideal P. Now $\mathbb{Z}_m$ is isomorphic to $\mathbb{Z}_{p_1^{k_1}} \times \mathbb{Z}_{p_2^{k_2}} \times \cdots \times \mathbb{Z}_{p_n^{k_n}}$ under the "usual" isomorphism given in the proof of Theorem 21.2, where we may as well assume that $k_1 > 1$. Let $[x]_m$ be the element of $\mathbb{Z}_m$ that gets mapped to the element of $\mathbb{Z}_{p_1^{k_1}} \times \mathbb{Z}_{p_2^{k_2}} \times \cdots \times \mathbb{Z}_{p_n^{k_n}}$ that has $[p_1]_{p_1^{k_1}}$ in the first component and $[0]$ in all the other components. Because $k_1 > 1$, the element $[x]_m$ is non-zero. Furthermore,

$$[x]_m^{k_1} = [0]_m \in P.$$

But because P is prime, this implies that $[x] \in P$.

▷ **Quick Exercise.** Why? ◁

Hence, the intersection of all prime ideals of $\mathbb{Z}_m$ contains $[x]_m$, and so cannot be $\{[0]_m\}$; thus, $\mathbb{Z}_m$ in this case cannot be a direct product of domains. This is the general explanation of the case $m = 12$.

▷ **Quick Exercise.** What would be the value for $[x]_m$ in the case $m = 12$? ◁

Historical Remarks

Sun Tsu's problem was phrased in this way: "We have things of which we do not know the number. If we count them by threes, the remainder is 2; if we count them by fives, the remainder is 3; if we count them by sevens, the remainder is 2. How many things are there?" This is clearly an example of the Chinese Remainder Theorem, which you will solve in Exercise e below. Because this is the only problem treated by him, it is unclear whether he had available a general method of solving a system of linear congruences. However, in the 13th-century, another Chinese mathematician named Qin Jiushao published a mathematical text that includes a number of examples of such problems. These examples provide in essence our algorithmic proof of the Chinese Remainder Theorem. In particular, to solve such congruences as $[x_i m_i] = [1]$, he uses a version of Euclid's Algorithm, where operations are carried out on a counting board. This medieval Chinese mathematics is much more sophisticated than anything happening in Europe at the same time.

Chapter Summary

In this chapter we proved that any ring of the form $\mathbb{Z}_m$ is isomorphic to a direct product of rings of the form $\mathbb{Z}_{p^k}$, where p is prime. In its number-theoretic guise, this is known as the *Chinese Remainder Theorem.*

Warm-up Exercises

a. According to the Chinese Remainder Theorem, what direct product of rings is $\mathbb{Z}_{24}$ isomorphic to? What about $\mathbb{Z}_{60}$? $\mathbb{Z}_{11}$? $\mathbb{Z}_9$?

b. Note that $[2]^4 = [0]$ in $\mathbb{Z}_{16}$. Why does this mean that $[2]$ belongs to *all* prime ideals of this ring?

c. What's the relationship between m and n, if there exists an onto ring homomorphism $\varphi : \mathbb{Z}_m \to \mathbb{Z}_n$?

d. Solve the simultaneous congruences

$$\begin{aligned} x &\equiv 1 \,(\text{mod } 2) \\ x &\equiv 2 \,(\text{mod } 3). \end{aligned}$$

e. Express Sun Zi's problem in modern notation, and solve it.

Exercises

1. Show directly that $\mathbb{Z}_{12}$ is isomorphic to $\mathbb{Z}_3 \times \mathbb{Z}_4$ by defining a homomorphism from $\mathbb{Z}_{12}$ onto $\mathbb{Z}_3 \times \mathbb{Z}_4$ as μ was defined in the proof of Theorem 21.2. Write out an explicit element-by-element description of this isomorphism, as we did in Section 21.1 for $\mathbb{Z}_6$.

2. Repeat Exercise 1 for the ring $\mathbb{Z}_{30}$. (First of all you must determine the direct product to which it is isomorphic.)

3. Determine the prime ideals of $\mathbb{Z}_{12}$; determine the maximal ideals of $\mathbb{Z}_{12}$.

4. Repeat Exercise 3 for $\mathbb{Z}_{30}$.

5. Solve the simultaneous congruences

$$\begin{aligned} x &\equiv 1 \,(\text{mod } 4) \\ x &\equiv 5 \,(\text{mod } 7) \\ x &\equiv 3 \,(\text{mod } 9). \end{aligned}$$

6. Solve the simultaneous congruences

$$\begin{aligned} x &\equiv 1 \,(\text{mod } 2) \\ x &\equiv 6 \,(\text{mod } 7) \\ x &\equiv 2 \,(\text{mod } 27) \\ x &\equiv 6 \,(\text{mod } 11). \end{aligned}$$

7. Find $a, b \in \mathbb{Z}$ so that there is no simultaneous solution to

$$x \equiv a(\text{mod } 6) \text{ and } x \equiv b(\text{mod } 4).$$

Why does this not contradict the Chinese Remainder Theorem?

8. The rings $\mathbb{Z}_9$ and $\mathbb{Z}_3 \times \mathbb{Z}_3$ both have nine elements; indeed, a careless reading of Theorem 21.2 might lead one to suppose that these rings are isomorphic. Show that this is false.

9. In Exercise 21.11 you will discover that the converse of Theorem 21.1 is false. However, the following theorem is true: A commutative ring R with unity is isomorphic to a direct product of finitely many integral domains if and only if it has a finite collection of prime ideals $\{P_i\}$ whose intersection is zero, and for each $r \in R$, there exist $r_i \in \cap\{P_j : j \neq i\}$ for which $r = \Sigma r_i$. Prove this.

10. In this exercise we consider exactly the condition described in Theorem 21.1. We call a ring R a **finite subdirect product of domains** if there exist finitely many integral domains $D_1, D_2, \cdots, D_n$ and a one-to-one homomorphism

$$\mu : R \to \prod D_i$$

such that $\pi_i(\mu(R)) = D_i$, for all i, where π_i is the projection homomorphism. Prove that R is a finite subdirect product of domains if and only if it has a finite collection of prime ideals whose intersection is zero.

11. Consider
$$R \;=\; \{(n, n+3m) : n, m \in \mathbb{Z}\} \;\subseteq\; \mathbb{Z} \times \mathbb{Z}.$$
In this exercise you will show that R is a finite subdirect product of domains but it is not a direct product of domains.

    (a) Prove that R is a subring of $\mathbb{Z} \times \mathbb{Z}$. Show that R is not an integral domain.

    (b) Show by specific example that R does not include all elements of $\mathbb{Z} \times \mathbb{Z}$. (That is, R is not the entire direct product.)

    (c) Consider the kernels P_1 and P_2 of the projection homomorphisms π_1 and π_2 that project onto the first and second coordinates, respectively. Prove that $P_1 = \langle (0,3) \rangle$ and that $P_2 = \langle (3,0) \rangle$.

    (d) Show from the definition given in Exercise 21.10 that R is a subdirect product of the domains $\mathbb{Z}$.

    (e) Suppose that P is any other prime ideal of R. Show that P necessarily contains either P_1 or P_2.

    (f) Now use the previous result and Exercise 9 to argue that R cannot be a finite direct product of domains.

12. Show that $\mathbb{Z}_6[x]$ is a finite subdirect product of domains.

13. Make a definition for a *finite subdirect product of fields*, on the model of the definition in Exercise 21.10. Then state and prove the theorem analogous to the result proved in Exercise 21.10.

14. In this exercise we construct a ring that is a domain (and so is trivially a finite subdirect product of domains) but is not a finite subdirect product of fields; to do this exercise you need to understand Exercises 21.10 and 21.13. Let p be a prime integer, and define
$$\mathbb{Z}_{\langle p \rangle} = \left\{ q \in \mathbb{Q} : q = \frac{a}{b},\ a, b \in \mathbb{Z}, \quad p \text{ does not divide } b \right\}.$$

    (a) Show that $\mathbb{Z}_{\langle p \rangle}$ is a subring of $\mathbb{Q}$, and so is an integral domain.

    (b) Define
    $$\varphi : \mathbb{Z}_{\langle p \rangle} \longrightarrow \mathbb{Z}_p$$
    by setting $\varphi(\frac{a}{b}) = [a]_p[b]_p^{-1}$. Prove that φ is an onto ring homomorphism.

    (c) Prove that the kernel of the homomorphism φ from part b is $\langle p \rangle$. This means that $\langle p \rangle$ is a maximal ideal of $\mathbb{Z}_{\langle p \rangle}$.

    (d) Prove that $\frac{a}{b} \in \mathbb{Z}_{\langle p \rangle}$ is a unit if and only if $\frac{a}{b} \notin \langle p \rangle$.

    (e) Use part d to argue that $\langle p \rangle$ contains *every* proper ideal of $\mathbb{Z}_{\langle p \rangle}$, and so this ring has a unique maximal ideal.

    (f) Why does part e mean that $\mathbb{Z}_{\langle p \rangle}$ is not a finite subdirect product of fields?

Section IV in a Nutshell

This section considers functions from one ring R to another ring S that preserve certain algebraic properties: consider $\varphi : R \to S$ such that

$$\varphi(a+b) = \varphi(a) + \varphi(b) \quad \text{and} \quad \varphi(ab) = \varphi(a)\varphi(b),$$

for all $a, b \in R$. We call φ a *ring homomorphism.*

A ring homomorphism always preserves the zero of the ring, additive inverses, unity and multiplicative inverses (Theorem 16.1). While a ring homomorphism $\varphi : R \to S$ need not be onto, it is onto the image of R in S ($\varphi(R)$) which is itself a subring of S (Theorem 16.2).

The *kernel* of φ is defined by

$$\ker(\varphi) = \varphi^{-1}(0) = \{r \in R : \varphi(r) = 0\}.$$

The kernel is always an ideal of R (Theorem 17.1). Furthermore, if $s \in \varphi(R) \subseteq S$, then $\varphi^{-1}(s)$ is the coset $\ker(\varphi) + r$, for any $r \in \varphi^{-1}(s)$. $\ker(\varphi) = 0$ if and only if φ is one-to-one (Theorem 17.4).

The cosets of any ideal I of R partitions R into pairwise disjoint sets (Theorem 18.1), called *cosets.* The set of cosets R/I is itself a ring, called the *ring of cosets* or the *quotient ring of R mod I.* There is a *natural homomorphism* from R onto R/I given by

$$\nu(a) = I + a.$$

The kernel of ν is I (Theorem 18.3).

If $\varphi : R \to S$ is a one-to-one onto homomorphism, we call φ an *isomorphism*; in this case we say the R and S are *isomorphic.* The *Fundamental Isomorphism Theorem* (Theorem 19.1) states that if $\varphi : R \to S$ is an onto homomorphism and $\nu : R \to R/\ker(\varphi)$ is the natural homomorphism, then $R/\ker(\varphi)$ is isomorphic to S. Furthermore, if we define $\mu : R/\ker(\varphi) \to S$ by $\mu(\ker(\varphi) + r) = \varphi(r)$, then μ is an isomorphism and $\mu \circ \nu = \varphi$. The essential content of this important theorem is that the output of a ring homomorphism can be obtained by forming the ring of cosets of the appropriate ideal. Chapter 19 gives many examples of this theorem.

An ideal I of R is *maximal* if the only ideal of R properly containing I is R itself. An ideal I is maximal if and only if R/I is a field (Theorem 20.1).

An ideal I of R is *prime* if whenever $ab \in I$ then either $a \in I$ or $b \in I$. An ideal I is prime if and only if R/I is a domain (Theorem 20.2). It follows that every maximal ideal is a prime ideal, since every field is a domain.

Finally, the section closes with a chapter on the famous *Chinese Remainder Theorem.* This is presented in two forms, of which the latter is the more usual:

(Theorem 21.2) Let $p_1, p_2, \cdots, p_n$ be distinct prime integers and $m = p_1^{k_1} p_2^{k_2} \cdots p_n^{k_n}$. Then $\mathbb{Z}_m$ is isomorphic to

$$\mathbb{Z}_{p_1^{k_1}} \times \mathbb{Z}_{p_2^{k_2}} \times \cdots \times \mathbb{Z}_{p_n^{k_n}}.$$

(Theorem 21.3) Let $p_1, p_2, \cdots, p_n$ be distinct prime integers and $m = p_1^{k_1} p_2^{k_2} \cdots p_n^{k_n}$. Then the set of congruences

$$\begin{gathered} x \equiv a_1 \,(\text{mod } p_1^{k_1}) \\ x \equiv a_2 \,(\text{mod } p_2^{k_2}) \\ \cdots \\ x \equiv a_n \,(\text{mod } p_n^{k_n}) \end{gathered}$$

has a simultaneous solution, which is unique modulo m.

V

Groups

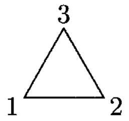
Chapter 22

Symmetries of Figures in the Plane

Many geometric objects possess a large amount of symmetry. Roughly speaking, this means that a change of the viewer's perspective does not change what is seen. Equivalently, we can move the object instead. In this case we want to consider motions of the object that leave it apparently unchanged.

22.1 Symmetries of the Equilateral Triangle

For example, consider an equilateral triangle with vertices labelled 1, 2, and 3:



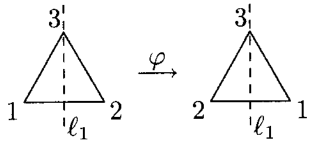
Notice that a counterclockwise rotation through 120° moves vertex 1 to the location vertex 2 has just vacated, and moves vertex 2 to location 3, and vertex 3 to location 1. We shall denote this rotation by ρ. Notice that if we ignore the labels of the vertices, after applying ρ the triangle is in the same position as it was before the motion. Note that if we apply ρ twice, that is, we rotate the triangle through 240°, the triangle again appears unchanged. We shall denote this rotation by $\rho\rho$, or ρ^2 for short.

Perhaps we need a more precise definition: a **rigid motion** of the plane is a one-to-one function from the plane onto itself that preserves distance. We call these rigid *motions* because they can be realized by moving the plane in three-dimensional space. If S is a subset of the plane (that is, a figure in the plane like our equilateral triangle), a **symmetry** of S is a rigid motion of the plane that takes S onto itself.

So, when we talk of a rotation of the equilateral triangle, we can just as well think of rotating the entire plane. Thus, ρ and ρ^2 are symmetries of the triangle.

Are there any other symmetries of the triangle? First of all, there is certainly the identity: the one-to-one onto distance-preserving function that takes each point to itself. This corresponds to *no motion at all*, and we will denote it by ι.

But are there any more interesting symmetries? Yes! Consider the **reflection** through the line ℓ_1 which assigns to each point P the point P' which is the same perpendicular distance from ℓ_1 but on the other side. (If P is actually on the line, it is sent to itself.)



We could think of this as rotating the plane on axis ℓ_1 through 180°. (Notice this motion happens in three-dimensional space.) This rigid motion takes the triangle onto itself, and so is a symmetry of the triangle that we will call φ (for *flip*).

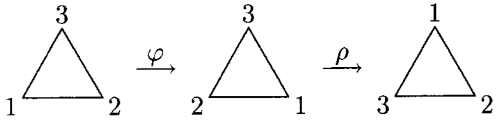
Now, obviously one symmetry followed by another is still a symmetry; note that 'followed by' here means *functional composition*, if we think of these rigid motions as functions from the plane onto itself.

So we have

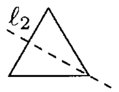
$$\rho, \rho^2, \rho^3 = \text{ no motion at all } = \iota.$$

$$\varphi, \varphi^2 = \iota$$

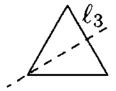
in our list of symmetries of the triangle. But we also have $\rho\varphi$ (where we mean by this juxtaposition the composition of these two functions, first φ and then ρ).



Now this is just a reflection about the line ℓ_2:



Similarly, $\varphi\rho$ looks like a reflection about the line ℓ_3:



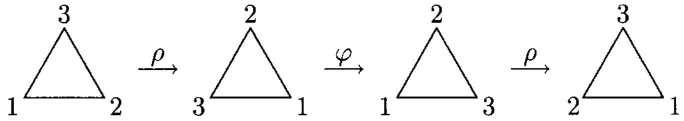
▷ **Quick Exercise.** Verify that $\varphi\rho$ is indeed equivalent to the reflection about the line ℓ_3. ◁

Note that the order in which these compositions are done makes a big difference! These six turn out to be a list of *all* the symmetries of the equilateral triangle (we'll prove this later).

We can now obtain a *multiplication table* of these symmetries as follows. The entry in the α row and β column consists of the composition $\alpha\beta$:

	ι	ρ	ρ^2	φ	$\rho\varphi$	$\varphi\rho$
ι	ι	ρ	ρ^2	φ	$\rho\varphi$	$\varphi\rho$
ρ	ρ	ρ^2	ι	$\rho\varphi$	$\varphi\rho$	φ
ρ^2	ρ^2	ι	ρ	$\varphi\rho$	φ	$\rho\varphi$
φ	φ	$\varphi\rho$	$\rho\varphi$	ι	ρ^2	ρ
$\rho\varphi$	$\rho\varphi$	φ	$\varphi\rho$	ρ	ι	ρ^2
$\varphi\rho$	$\varphi\rho$	$\rho\varphi$	φ	ρ^2	ρ	ι

For instance, let's compute $\rho\varphi\rho$. We can picture this composition as follows:



We see that this is indeed equivalent to φ.

▷ **Quick Exercise.** Try generating the remainder of this multiplication table yourself, thinking geometrically about composing movements of the triangle. ◁

Notice that this 'multiplication' obeys the associative law because it is simply the composition of functions, which is associative. Note also that there seems to be a block substructure to this table. Namely, if we denote the three rotations by R and the three flips by F, we have

	R	F
R	R	F
F	F	R

We will return to this substructure of the table in Example 27.5.

22.2 Permutation Notation

Now, how do we know that this comprises the complete list of the symmetries of the equilateral triangle?

To answer this question, we make the following observation: If we specify where each of the vertices goes, then we have determined the symmetry. We introduce some notation here, to describe what a symmetry does to the vertices of the triangle. For example, to describe ρ, we put beneath the integers 1, 2, and 3 the names of the vertex locations they're taken to. Thus, we should put 2 under 1 to indicate that the counterclockwise rotation of ρ takes vertex 1 to the location vertex 2 has just vacated. This is an example of a general notation for *permutations*, which we will study later. Thus, the permutation of the vertices $1, 2, 3$ accomplished by the rotation ρ is just

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

This then means that the vertex in location 1 is moved to location 2, and likewise for the other two columns.

This latter notation gives us a function from $\{1,2,3\}$ onto $\{1,2,3\}$. For example,

$$\begin{pmatrix} 1\,2\,3 \\ 2\,3\,1 \end{pmatrix}(2) = 3.$$

Explicitly, we have the following correspondence between symmetries of the triangle and permutations of their vertex locations:

$$\iota \longleftrightarrow \begin{pmatrix} 1\,2\,3 \\ 1\,2\,3 \end{pmatrix}$$

$$\rho \longleftrightarrow \begin{pmatrix} 1\,2\,3 \\ 2\,3\,1 \end{pmatrix}$$

$$\rho^2 \longleftrightarrow \begin{pmatrix} 1\,2\,3 \\ 3\,1\,2 \end{pmatrix}$$

$$\varphi \longleftrightarrow \begin{pmatrix} 1\,2\,3 \\ 2\,1\,3 \end{pmatrix}$$

$$\varphi\rho \longleftrightarrow \begin{pmatrix} 1\,2\,3 \\ 1\,3\,2 \end{pmatrix}$$

$$\rho\varphi \longleftrightarrow \begin{pmatrix} 1\,2\,3 \\ 3\,2\,1 \end{pmatrix}$$

Now, there are only six ways to rearrange the list $1, 2, 3$: We have three choices for the first element in the list, two remaining choices for the second element, and only one remaining choice for the third element, making $3 \cdot 2 \cdot 1 = 3! = 6$ altogether. So we have the complete list of the symmetries of the triangle.

We call the set of these six symmetries the **group of symmetries** of the equilateral triangle; the table above is called its **group table** or **multiplication table**. The group of symmetries of the equilateral triangle is sometimes called the **3rd dihedral group** and denoted D_3. In general, the group of symmetries of a regular n-sided polygon is called the **nth dihedral group** and is denoted D_n. In the future, we will call these groups either dihedral groups or the groups of symmetries of appropriate regular polygons. In Exercise 22.8, you will show that D_n has $2n$ elements.

From the group table, we know that the composition of symmetries $\rho(\varphi\rho)$ gives us the symmetry φ. But consider the two corresponding permutations of the vertices:

$$\rho = \begin{pmatrix} 1\,2\,3 \\ 2\,3\,1 \end{pmatrix} \quad \text{and} \quad \varphi\rho = \begin{pmatrix} 1\,2\,3 \\ 1\,3\,2 \end{pmatrix}.$$

These are both functions, which can be composed together. Let's see what $\rho(\varphi\rho)$ does to vertex 1. First, the symmetry $\varphi\rho$ sends 1 to 1, then ρ sends 1 to 2. Thus, the composition function sends 1 to 2.

▷ **Quick Exercise.** Compute what this composition does to 2 and 3. ◁

We thus have

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

(Remember, this composition is done right-to-left.) This is of course exactly the permutation describing what the corresponding symmetry φ does to the vertex locations! Thus, $\rho(\varphi\rho) = \varphi$.

22.3 Matrix Notation

If you've had a bit of linear algebra, we can describe the symmetries of the equilateral triangle using matrix notation. Let's now look again at the rigid motions of the plane we've used in describing the symmetries of the equilateral triangle. Remember a rigid motion of the plane is really a one-to-one onto function from the plane onto the plane, which preserves distance.

Now the plane can be considered algebraically as the set $\mathbb{R}^2$ of all ordered pairs of real numbers. We denote this by writing $P(x, y)$ for the point P in the plane with coordinates (x, y).

How then can we represent rotation (about the origin, say) through angle θ? To answer this, suppose that $P(x, y)$ is rotated through angle θ to point $P'(x', y')$. If we represent P by the polar coordinates (r, φ) then

$$x = r\cos\varphi \quad \text{and} \quad y = r\sin\varphi,$$

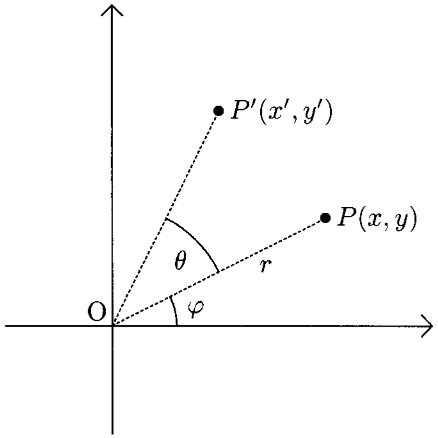
and so

$$x' = r\cos(\theta + \varphi) \quad \text{and} \quad y' = r\sin(\theta + \varphi).$$

By using the trigonometric sum formulas, we obtain

$$\begin{aligned} x' &= r\cos\varphi\cos\theta - r\sin\varphi\sin\theta = x\cos\theta - y\sin\theta \\ y' &= r\cos\varphi\sin\theta + r\sin\varphi\cos\theta = x\sin\theta + y\cos\theta. \end{aligned}$$

These two equations describe the transformation $P(x, y) \rightarrow P'(x', y')$, which is the rotation through the angle θ about the origin.



A particularly nice way to look at this transformation is as a matrix multiplication. Consider the 2×2 matrix

$$R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

and the 2×1 arrays $P = \begin{pmatrix} x \\ y \end{pmatrix}$ and $P' = \begin{pmatrix} x' \\ y' \end{pmatrix}$. Then we have that $RP = P'$. Note that we think of P and P' as columns in order to make this matrix multiplication work. Thus, we can represent rotation of the plane by means of a *matrix multiplication.* (See Exercise 6.7.)

What about the reflection through lines? Consider the matrix

$$F = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Note that $FP = \begin{pmatrix} -x \\ y \end{pmatrix}$, and so we see that multiplication by F exactly describes reflection of the plane through the y-axis.

So now consider the equilateral triangle centered about the origin $(0, 0)$, so that vertex 1 is $(-\frac{\sqrt{3}}{2}, -\frac{1}{2})$, vertex 2 is $(\frac{\sqrt{3}}{2}, -\frac{1}{2})$, and vertex 3 is $(0, 1)$. (Note that the vertices of this triangle are all distance 1 from the origin.) We can then explicitly describe the symmetries ρ and φ by multiplication by R and F, respectively. We then obtain the following correspondence between symmetries and matrices:

$$\iota \longleftrightarrow I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\rho \longleftrightarrow R = \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}$$

$$\rho^2 \longleftrightarrow R^2 = \begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}$$

$$\varphi \longleftrightarrow F = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\varphi\rho \longleftrightarrow FR = \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}$$

$$\rho\varphi \longleftrightarrow RF = \begin{pmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}$$

Note the interconnection among the composition of symmetries, the composition of permutations of the vertex locations, and the matrix multiplication! In other words, the matrix corresponding to the composition of two symmetries is exactly the product of the matrices corresponding to those symmetries.

▷ **Quick Exercise.** Pick a couple of symmetries and verify that the product of the symmetries has as its corresponding matrix exactly the product of the corresponding matrices of the original symmetries. ◁

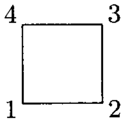
(We will explore this correspondence more carefully later.)

22.4 Symmetries of the Square

Let's now examine the symmetries of the square. We consider the square pictured below, with vertices at

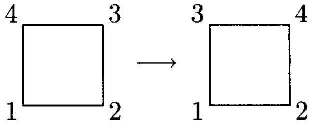
$$(-1,-1), \quad (1,-1), \quad (1,1), \quad (-1,1).$$

(Call these vertices 1, 2, 3, and 4.)



Because vertices must be moved to vertices, there can be no more than 4! = 24 symmetries of the square. However, consider the permutation of the vertex locations

$$\begin{pmatrix} 1\,2\,3\,4 \\ 1\,2\,4\,3 \end{pmatrix}.$$



There is no way to move the square in this way, because vertex 1 (wherever it is moved) must stay adjacent to vertices 2 and 4. Let's count how many such permutations are possible. We can move vertex 1 to any of four places. But because vertex 2 must be adjacent, this means that there remain only two possibilities for 2. And once we've established where the 12 edge goes, the entire square's motion is accounted for. This means there are at most 8 symmetries of the square. In Exercise 22.1, you will find all 8, and analyze them as we have analyzed the symmetries of the equilateral triangle. As mentioned before, we denote the group of the symmetries of the square by D_4, the 4th dihedral group. In Exercise 22.8 you actually argue that the nth dihedral group D_n has $2n$ elements.

Chapter Summary

In this chapter we explored the notion of symmetry of figures in the plane. We obtained the *group of symmetries* for the equilateral triangle. This group is called the 3rd *dihedral group.*

Warm-up Exercises

a. Explain geometrically why ι appears in each row and in each column of the group table for D_3.

b. We pointed out a block substructure in the group table for D_3. What does this substructure mean geometrically?

c. Give the matrices that accomplish rotations through 45° and 30°. Check that these matrices work for $(1,0)$ and $(0,1)$.

d. Compute $\rho\varphi\rho$ three times: **(1)** by making an equilateral triangle out of cardboard, and actually performing the motions; **(2)** by composing the permutation functions which describe what locations the vertices are taken to; **(3)** by multiplying the corresponding matrices. Did you get the same answer all three times?

e. Consider an *isosceles* triangle, which is not equilateral. How many symmetries does it have?

f. Consider a *scalene* triangle (all sides have different lengths). How many symmetries does it have?

Exercises

1. Complete the analysis of the symmetries of the square, which we began in the text. Some will be rotations, and some will be flips. Determine matrix and permutation representations for them, draw a table of correspondence, and compute the group table for your symmetries.

2. Repeat Exercise 1 for D_5, the group of symmetries of a regular pentagon.

3. Determine all symmetries of a non-square rectangle, and represent them with matrices and permutations. How many are rotations, and how many are flips?

4. Repeat Exercise 3 for a rhombus (that is, an equilateral parallelogram, which is not a square).

5. Show algebraically that the rotation transformation preserves distance: Consider the points

$$P_1(x_1, y_1) \text{ and } P_2(x_2, y_2).$$

   (a) What is the square of the distance between P_1 and P_2?

   (b) Now rotate through the angle θ, by multiplying by the appropriate matrix, to obtain the points

$$P_1'(x_1', y_1') \text{ and } P_2'(x_2', y_2').$$

   Compute the square of the distance between these points. Use trig identities to show that this is the same as in part a.

6. Verify by multiplying two matrices together that a rotation through angle θ, followed by a rotation through angle φ, gives a rotation through angle $\theta + \varphi$.

7. How many symmetries can you find for the unit circle? Which rotations are possible? Which flips?

8. Find out how many elements there are in D_n, the group of symmetries of a regular n-sided polygon.

9. You can check that all of the matrices of the symmetries of the equilateral triangle and the square have the property that their determinants are always ± 1. (See Exercise 8.2 for a definition of the determinant of a 2×2 matrix.) In this exercise you will show that if a matrix preserves distance, then its determinant must be 1.

   (a) Suppose that $A \in M_2(\mathbb{R})$, and $\det(A) = 0$. Show that multiplication by A cannot preserve distance. Do this by showing that multiplication by A takes some point in the plane to the origin, and hence cannot preserve distance.

   (b) Suppose next that

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = A \in M_2(\mathbb{R}),$$

   but $\det(A) \neq 0$. Suppose that multiplication by A does preserve distance, and consider successively what happens to

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} d \\ -c \end{pmatrix}, \begin{pmatrix} -b \\ a \end{pmatrix}.$$

   You will be able to infer that $\det(A) = \pm 1$.

10. Our description of the symmetries of the equilateral triangle can be elegantly rephrased using the arithmetic of the complex numbers $\mathbb{C}$, described in Chapter 8.

    (a) Argue that the three vertices of the triangle can be thought of as numbers of the form $e^{i\alpha_i}$ in the complex plane, for appropriate angles α_i.

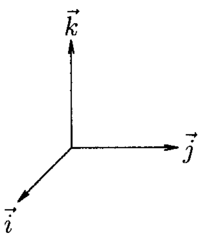
    (b) Show that you can represent the rotations of the triangle in the symmetry group by complex multiplication by a number of the form $e^{i\theta}$, for an appropriate choice of θ.

    (c) What operation on the complex numbers performs the flip ϕ?

Chapter 23

Symmetries of Figures in Space

We'd now like to turn to symmetries of three-dimensional objects, in particular of the regular tetrahedron and the cube. By way of analogy with plane figures, it should be clear that we should be concerned with rigid motions of three-dimensional space ($\mathbb{R}^3$) taking the given object onto itself. However, we should make clear what is meant by a rigid motion of $\mathbb{R}^3$. We mean more than just a one-to-one onto distance-preserving function. Consider $\mathbb{R}^3$ equipped with a coordinate system with the usual right-hand rule orientation as depicted below (where $\vec{i} \times \vec{j} = \vec{k}$):

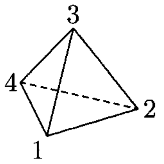


The function $\beta : \mathbb{R}^3 \to \mathbb{R}^3$ defined by $\beta(x, y, z) = (x, -y, z)$ (that is, reflection through the xz-plane) is a one-to-one onto distance-preserving function that *cannot* be accomplished by moving $\mathbb{R}^3$ *in* $\mathbb{R}^3$. We can see this because it changes the set of vectors $\vec{i}, \vec{j}, \vec{k}$, which has *right-handed* orientation, into the set $\vec{i}, -\vec{j}, \vec{k}$, which has *left-handed* orientation.

The pertinent comparison is to the reflection $(x, y) \mapsto (x, -y)$ in the plane, which can't be accomplished in the plane. It is an accident of physics that we inhabit three-dimensional space. Hence, we are happy with reflections through a line in the plane, because they can be accomplished in three-dimensional space. We are not so happy with reflections through a plane in space (even though it turns out they can be accomplished as a motion in four dimensions). The key here is to restrict ourselves to one-to-one onto distance-preserving functions that preserve the right-handed orientation of our coordinate system; we call such functions **rigid motions of space**. (This should jibe with your intuition of rigid motion.) This definition of symmetry is quite restrictive because it excludes functions like β, which can be realized as mirror reflections. Many people would call such functions symmetries too, but we will not do so here.

23.1 Symmetries of the Regular Tetrahedron

We're now ready to determine the group of symmetries of a regular tetrahedron. Because it has four vertices, there are no more than $4! = 24$ such symmetries.



Let's first consider symmetries leaving one vertex (say, number 1) fixed. We have the rotations

$$\rho_1 \longmapsto \begin{pmatrix} 1\,2\,3\,4 \\ 1\,3\,4\,2 \end{pmatrix}$$

and

$$\rho_1^2 \longmapsto \begin{pmatrix} 1\,2\,3\,4 \\ 1\,4\,2\,3 \end{pmatrix},$$

where, of course, $\rho_1^3 = \iota$, the identity.

But consider another permutation of the vertices leaving number 1 fixed, say,

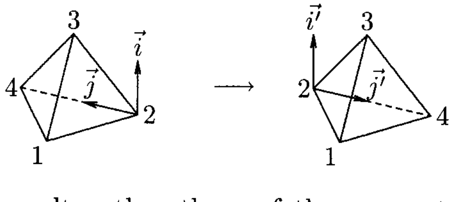
$$\begin{pmatrix} 1\,2\,3\,4 \\ 1\,4\,3\,2 \end{pmatrix}.$$

This corresponds to a reflection through a plane.

▷ **Quick Exercise.** What plane? ◁

But this *can't* be accomplished with a rigid motion! To see this, paint the exterior of face 234 white and the interior of face 234 red. Now apply the above permutation. Clearly, for this to be a symmetry of the regular tetrahedron, a face must be mapped to another face. Here, because vertex 1 is fixed, face 234 is clearly mapped to itself. But after the permutation, what color are the exterior and interior of face 234? They've switched colors! This permutation has caused the tetrahedron to turn 'inside itself'. This doesn't jibe with our intuition of rigid motion. Indeed, this is the sort of thing that happens when a distance-preserving permutation is applied that changes the right-handed orientation.

To see this change in orientation another way, place vectors $\vec{i}$ and $\vec{j}$ in the plane of the face 234, with $\vec{j}$ parallel to the edge 24. Then $\vec{i} \times \vec{j}$ points toward the tetrahedron. Suppose $\vec{i}'$ and $\vec{j}'$ are the images of $\vec{i}$ and $\vec{j}$, respectively. Note that $\vec{i}' \times \vec{j}'$ now points away from the tetrahedron. That is, the side of the face that was formerly in the interior of the tetrahedron is now on the exterior.



So we lose altogether three of the symmetries of the triangle 234: namely,

$$\begin{pmatrix} 1\,2\,3\,4 \\ 1\,4\,3\,2 \end{pmatrix}, \begin{pmatrix} 1\,2\,3\,4 \\ 1\,2\,4\,3 \end{pmatrix}, \text{ and } \begin{pmatrix} 1\,2\,3\,4 \\ 1\,3\,2\,4 \end{pmatrix}.$$

Similarly, we have rotations corresponding to the other three vertices:

$$\begin{array}{ll} \rho_2 \to \begin{pmatrix} 1\,2\,3\,4 \\ 4\,2\,1\,3 \end{pmatrix} & \rho_2^2 \to \begin{pmatrix} 1\,2\,3\,4 \\ 3\,2\,4\,1 \end{pmatrix} \\ \rho_3 \to \begin{pmatrix} 1\,2\,3\,4 \\ 2\,4\,3\,1 \end{pmatrix} & \rho_3^2 \to \begin{pmatrix} 1\,2\,3\,4 \\ 4\,1\,3\,2 \end{pmatrix} \\ \rho_4 \to \begin{pmatrix} 1\,2\,3\,4 \\ 3\,1\,2\,4 \end{pmatrix} & \rho_4^2 \to \begin{pmatrix} 1\,2\,3\,4 \\ 2\,3\,1\,4 \end{pmatrix} \end{array}$$

These give us 8 symmetries fixing exactly one vertex, together with one fixing all four, while we have banned 6.

▷ **Quick Exercise.** For each fixed vertex, we found 3 permutations that are not allowed. There are 4 vertices, so why are there only 6 banned permutations and not 12? ◁

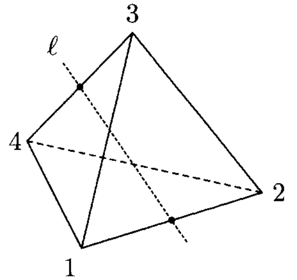
What about the $24 - 15 = 9$ permutations that leave no vertex fixed? How many of these lead to symmetries? Let's try to consider a typical one. We may suppose without loss of generality that vertex 1 is sent to vertex 2. So, the permutation looks like this:

$$\begin{pmatrix} 1\,2\,3\,4 \\ 2\;?\;?\;? \end{pmatrix}.$$

If 2 ↦ 1, then (because we've assumed that it fixes no vertex) it must look like this:

$$\begin{pmatrix} 1\,2\,3\,4 \\ 2\,1\,4\,3 \end{pmatrix}.$$

This is a symmetry! Just let ℓ be the line connecting the midpoints of the segments 12 and 34, respectively. Then this symmetry can be accomplished by rotating 180° about the line ℓ.



We will call this symmetry φ_1; note that $\varphi_1^2 = \iota$. In a similar fashion we have

$$\varphi_2 \to \begin{pmatrix} 1\,2\,3\,4 \\ 3\,4\,1\,2 \end{pmatrix}, \qquad \varphi_3 \to \begin{pmatrix} 1\,2\,3\,4 \\ 4\,3\,2\,1 \end{pmatrix}.$$

But what if 2 is not mapped to 1? Suppose without loss of generality that 2 ↦ 3. Then, because we have no fixed points, we must have

$$\begin{pmatrix} 1\,2\,3\,4 \\ 2\,3\,4\,1 \end{pmatrix}.$$

We claim this *can't* be accomplished by a rigid motion of $\mathbb{R}^3$. Why? Once again, this motion would not preserve the orientation of our coordinate system. Consider any face of the tetrahedron. Note that the permutation will flip this face around—the exterior side becomes the interior side. This cannot be accomplished by a rigid motion.

▷ **Quick Exercise.** Pick a particular face of the tetrahedron and check explicitly that this permutation interchanges the interior and exterior sides of the face. ◁

Thus, there are six more permutations that are forbidden:

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

We thus have 12 symmetries of the regular tetrahedron: 8 that fix exactly one vertex, 3 that fix no vertex, and the one that fixes all four. We then obtain the following group table for the symmetries of the regular tetrahedron:

	ι	φ_1	φ_2	φ_3	ρ_1	ρ_2	ρ_3	ρ_4	ρ_1^2	ρ_2^2	ρ_3^2	ρ_4^2
ι	ι	φ_1	φ_2	φ_3	ρ_1	ρ_2	ρ_3	ρ_4	ρ_1^2	ρ_2^2	ρ_3^2	ρ_4^2
φ_1	φ_1	ι	φ_3	φ_2	ρ_3	ρ_4	ρ_1	ρ_2	ρ_4^2	ρ_3^2	ρ_2^2	ρ_1^2
φ_2	φ_2	φ_3	ι	φ_1	ρ_4	ρ_3	ρ_2	ρ_1	ρ_2^2	ρ_1^2	ρ_4^2	ρ_3^2
φ_3	φ_3	φ_2	φ_1	ι	ρ_2	ρ_1	ρ_4	ρ_3	ρ_3^2	ρ_4^2	ρ_1^2	ρ_2^2
ρ_1	ρ_1	ρ_4	ρ_2	ρ_3	ρ_1^2	ρ_4^2	ρ_2^2	ρ_3^2	ι	φ_3	φ_1	φ_2
ρ_2	ρ_2	ρ_3	ρ_1	ρ_4	ρ_3^2	ρ_2^2	ρ_4^2	ρ_1^2	φ_3	ι	φ_2	φ_1
ρ_3	ρ_3	ρ_2	ρ_4	ρ_1	ρ_4^2	ρ_1^2	ρ_3^2	ρ_2^2	φ_1	φ_2	ι	φ_3
ρ_4	ρ_4	ρ_1	ρ_3	ρ_2	ρ_2^2	ρ_3^2	ρ_1^2	ρ_4^2	φ_2	φ_1	φ_3	ι
ρ_1^2	ρ_1^2	ρ_3^2	ρ_4^2	ρ_2^2	ι	φ_2	φ_3	φ_1	ρ_1	ρ_3	ρ_4	ρ_2
ρ_2^2	ρ_2^2	ρ_4^2	ρ_3^2	ρ_1^2	φ_2	ι	φ_1	φ_3	ρ_4	ρ_2	ρ_1	ρ_3
ρ_3^2	ρ_3^2	ρ_1^2	ρ_2^2	ρ_4^2	φ_3	φ_1	ι	φ_2	ρ_2	ρ_4	ρ_3	ρ_1
ρ_4^2	ρ_4^2	ρ_2^2	ρ_1^2	ρ_3^2	φ_1	φ_3	φ_2	ι	ρ_3	ρ_1	ρ_2	ρ_4

In examining this group table we can detect a block substructure, similar to that we discovered in the group table for D_3. We can highlight this by labelling the symmetries

$$\{\iota, \varphi_1, \varphi_2, \varphi_3\}$$

by F, the symmetries

$$\{\rho_1, \rho_2, \rho_3, \rho_4\}$$

by R, and the symmetries

$$\{\rho_1^2, \rho_2^2, \rho_3^2, \rho_4^2\}$$

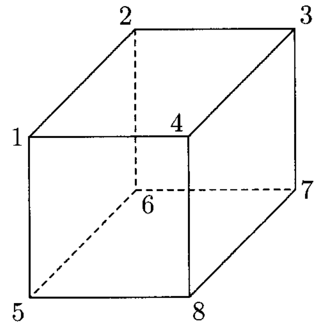
by R^2. Then the block structure looks like this:

	F	R	R^2
F	F	R	R^2
R	R	R^2	F
R^2	R^2	F	R

We will examine this block structure further in Example 27.6.

23.2 Symmetries of the Cube

We shall now determine the symmetries of a cube. We shall identify the vertices and the faces of the cube as follows:



face 1485	F	(front)
face 3267	B	(back)
face 4378	R	(right)
face 2156	L	(left)
face 2341	U	(up)
face 5876	D	(down)

As we've seen before, each symmetry corresponds to a distinct permutation of the vertices. However, there are

$$8! = 40,320$$

of these permutations! We'd like a more clever approach than this. Notice that each symmetry corresponds to a distinct permutation of the faces. However, there are still $6! = 720$ of these permutations.

To improve our geometric intuition a bit, let's paint the faces of the cube as follows:

front	— W (white)	left	— O (orange)
back	— Y (yellow)	up	— B (black)
right	— R (red)	down	— G (green)

This allows us to distinguish easily among the moving faces of the cube (the colors), and the unchanging locations to which the faces travel—such locations as front or right. (At this stage you might find it worthwhile to get a cube—a die, say—to manipulate.)

Now, notice that a symmetry is entirely determined once we know the colors of the front and right faces after the symmetry has been performed. (This would of course be true with any other pair of adjacent faces.)

This observation makes it easy to count all the symmetries of the cube, by counting all possible adjacent color pairs:

WR
WB RB
WO RY BY OY GY
WG RG BO OG

Of course, the symmetry making the front red and right green is distinct from the symmetry making the front green and right red, and so this gives us 24 symmetries of the cube.

But 24 is the number of permutations of 4 objects. Perhaps there is some way of describing the symmetries of the cube as permutations of 4 geometric objects. Notice that the cube has exactly 4 diagonals, and each symmetry takes the diagonals onto diagonals. Let's label the diagonals as follows:

$$a(1-7), \quad b(2-8), \quad c(3-5), \quad d(4-6).$$

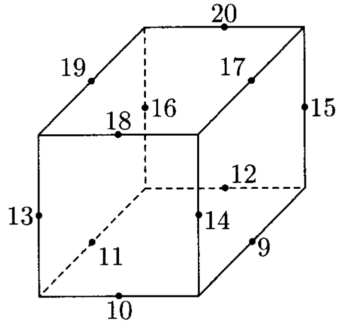
Each symmetry can be realized as a permutation of these four diagonals.

Example 23.1

Consider the symmetry of the cube that places R on the front face and W on the right face. Note that this symmetry leaves diagonals a and c fixed while interchanging diagonals b and d.

▷ **Quick Exercise.** Check that symmetry in Example 23.1 does indeed move the diagonals of the cube in the way described there. ◁

We can even make explicit the geometric motion necessary for each of the symmetries if we introduce some more fixed locations, so that we can specify certain axes of rotation:



Example 23.2

Return to the example above. Notice that this symmetry is realized by a 180° rotation about the line through the points 14 and 16.

▷ **Quick Exercise.** Verify this. ◁

We perform a similar analysis for all 24 symmetries of the cube, obtaining the following table. In the first column, we have given names to the symmetries. In the second and third columns, we have specified the color of the front and right faces, after the symmetry has been accomplished. In the fourth column, we describe the geometric motion necessary to accomplish the symmetry; note that all motions (except the identity) are rotations of the cube. The axes of rotation are

1. lines through opposite faces,

2. diagonals, or

3. lines through midpoints of opposite edges.

The fifth column specifies the permutation of the diagonals that results.

Symmetries of the Cube

	Front	Right	Geometric	
Element	Color	Color	Motion	Permutation
ι	W	R	no motion	abcd
ρ_1	W	B	90° on F,B	dcab
ρ_1^2	W	O	180° on F,B	badc
ρ_1^3	W	G	270° on F,B	cdba
ρ_2	R	Y	90° on U,D	bcda
ρ_2^2	Y	O	180° on U,D	cdab
ρ_2^3	O	W	270° on U,D	dabc
ρ_3	B	R	90° on L,R	bdac
ρ_3^2	Y	R	180° on L,R	dcba
ρ_3^3	G	R	270° on L,R	cadb
μ_1	O	G	120° about a	adbc
μ_1^2	B	Y	240° about a	acdb
μ_2	G	W	120° about b	dbac
μ_2^2	R	G	240° about b	cbda
μ_3	O	B	120° about c	dacb
μ_3^2	G	Y	240° about c	bdca
μ_4	B	W	120° about d	cabd
μ_4^2	R	B	120° about d	bcad
α_1	O	Y	180° about 13 & 15	cbad
α_2	R	W	180° about 14 & 16	adcb
α_3	G	O	180° about 10 & 20	acbd
α_4	Y	G	180° about 9 & 19	bacd
α_5	Y	B	180° about 11 & 17	abdc
α_6	B	O	180° about 12 & 18	dbca

▷ **Quick Exercise.** Verify the internal consistency of several lines of this table. It's easier if you have a cube to handle, preferably painted appropriately! ◁

By means of some exceedingly tedious computation, we can then obtain a group table for the group of symmetries of the cube:

	ι	ρ_1	ρ_1^2	ρ_1^3	ρ_2	ρ_2^2	ρ_2^3	ρ_3	ρ_3^2	ρ_3^3	μ_1	μ_1^2	μ_2	μ_2^2	μ_3	μ_3^2	μ_4	μ_4^2	α_1	α_2	α_3	α_4	α_5	α_6
ι	ι	ρ_1	ρ_1^2	ρ_1^3	ρ_2	ρ_2^2	ρ_2^3	ρ_3	ρ_3^2	ρ_3^3	μ_1	μ_1^2	μ_2	μ_2^2	μ_3	μ_3^2	μ_4	μ_4^2	α_1	α_2	α_3	α_4	α_5	α_6
ρ_1	ρ_1	ρ_1^2	ρ_1^3	ι	μ_4^2	α_5	μ_3	μ_1^2	α_4	μ_2	ρ_2^3	α_6	α_3	ρ_2	α_1	ρ_3^3	ρ_3	α_2	μ_1	μ_2^2	μ_3^2	ρ_2^2	ρ_3^2	μ_4
ρ_1^2	ρ_1^2	ρ_1^3	ι	ρ_1	α_2	ρ_3^2	α_1	α_6	ρ_2^2	α_3	μ_3	μ_4	μ_3^2	μ_4^2	μ_1	μ_2	μ_1^2	μ_2^2	ρ_2^3	ρ_2	ρ_3^3	α_5	α_4	ρ_3
ρ_1^3	ρ_1^3	ι	ρ_1	ρ_1^2	μ_2^2	α_4	μ_1	μ_4	α_5	μ_3^2	α_1	ρ_3	ρ_3^3	α_2	ρ_2^3	α_3	α_6	ρ_2	μ_3	μ_4^2	μ_2	ρ_3^2	ρ_2^2	μ_1^2
ρ_2	ρ_2^2	μ_1^2	α_1	μ_3^2	ρ_2^2	ρ_2^3	ι	μ_2^2	α_2	μ_4^2	ρ_3^3	α_4	ρ_1	α_3	ρ_3	α_5	ρ_1^3	α_6	ρ_3^2	ρ_1^2	μ_3	μ_2	μ_4	μ_1
ρ_2^2	ρ_2^2	α_4	ρ_3^2	α_5	ρ_2^3	ι	ρ_2	α_3	ρ_1^2	α_6	μ_4^2	μ_2	μ_1^2	μ_3	μ_2^2	μ_4	μ_3^2	μ_1	α_2	α_1	ρ_3	ρ_1	ρ_1^3	ρ_3^3
ρ_2^3	ρ_2^3	μ_2	α_2	μ_4	ι	ρ_2	ρ_2^2	μ_3	α_1	μ_1	α_6	ρ_1	α_4	ρ_3	α_3	ρ_1^3	α_5	ρ_3^3	ρ_1^2	ρ_3^2	μ_2^2	μ_1^2	μ_3^2	μ_4^2
ρ_3	ρ_3	μ_3	α_3	μ_2^2	μ_1^2	α_6	μ_4	ρ_3^2	ρ_3^3	ι	ρ_1^3	α_1	ρ_2^3	α_4	α_5	ρ_2	α_2	ρ_1	μ_3^2	μ_2	ρ_2^2	μ_1	μ_4^2	ρ_1^2
ρ_3^2	ρ_3^2	α_5	ρ_2^2	α_4	α_1	ρ_1^2	α_2	ρ_3^3	ι	ρ_3	μ_2^2	μ_3^2	μ_4	μ_1	μ_4^2	μ_1^2	μ_2	μ_3	ρ_2	ρ_2^3	α_6	ρ_1^3	ρ_1	α_3
ρ_3^3	ρ_3^3	μ_4^2	α_6	μ_1	μ_3^2	α_3	μ_2	ι	ρ_3	ρ_3^2	α_4	ρ_2	α_2	ρ_1^3	ρ_1	α_1	ρ_2^3	α_5	μ_1^2	μ_4	ρ_1^2	μ_2^2	μ_3	ρ_2^2
μ_1	μ_1	ρ_3^3	μ_4^2	α_6	ρ_1^3	μ_2^2	α_4	ρ_2^3	μ_3	α_1	μ_1^2	ι	ρ_3^2	μ_4	μ_2	ρ_1^2	ρ_2^2	μ_3^2	ρ_1	α_5	α_2	ρ_3	α_3	ρ_2
μ_1^2	μ_1^2	α_1	μ_3^2	ρ_2	α_6	μ_4	ρ_3	α_4	μ_2	ρ_1	ι	μ_1	μ_3	ρ_2^2	ρ_3^2	μ_4^2	μ_2^2	ρ_1^2	ρ_3^3	α_3	α_5	ρ_2^3	α_2	ρ_1^3
μ_2	μ_2	α_2	μ_4	ρ_2^3	ρ_3^3	μ_3^2	α_3	ρ_1	μ_1^2	α_4	ρ_2^2	μ_4^2	μ_2^2	ι	ρ_1^2	μ_1	μ_3	ρ_3^2	α_6	ρ_3	ρ_1^3	ρ_2	α_1	α_5
μ_2^2	μ_2^2	ρ_3	μ_3	α_3	α_4	μ_1	ρ_1^3	α_2	μ_4^2	ρ_2	μ_3^3	ρ_3^2	ι	μ_2	μ_4	ρ_2^2	ρ_1^2	μ_1^2	α_5	ρ_1	ρ_2^3	ρ_3^3	α_6	α_1
μ_3	μ_3	α_3	μ_2^2	ρ_3	ρ_1	μ_4^2	α_5	α_1	μ_1	ρ_2^3	μ_4	ρ_1^2	ρ_2^2	μ_1^2	μ_3^2	ι	ρ_3^2	μ_2	ρ_1^3	α_4	ρ_2	α_1	ρ_3^3	α_2
μ_3^2	μ_3^2	ρ_2	μ_1^2	α_1	α_3	μ_2	ρ_3^3	ρ_1^3	μ_4	α_5	ρ_3^2	μ_2^2	μ_4^2	ρ_1^2	ι	μ_3	μ_1	ρ_2^2	ρ_3	α_6	ρ_1	α_2	ρ_2^3	α_4
μ_4	μ_4	ρ_2^3	μ_2	α_2	ρ_3	μ_1^2	α_6	α_5	μ_3^2	ρ_1^3	ρ_1^2	μ_3	μ_1	ρ_3^2	ρ_2^2	μ_2^2	μ_4^2	ι	α_3	ρ_3^3	α_4	α_1	ρ_2	ρ_1
μ_4^2	μ_4^2	α_6	μ_1	ρ_3^3	α_5	μ_3	ρ_1	ρ_2	μ_2^2	α_2	μ_2	ρ_2^2	ρ_1^2	μ_3^2	μ_1^2	ρ_3^2	ι	μ_4	α_4	ρ_1^3	α_1	α_3	ρ_3	ρ_2^3
α_1	α_1	μ_3^2	ρ_2	μ_1^2	ρ_1^2	α_2	ρ_3^2	μ_1	ρ_2^3	μ_3	ρ_3	ρ_1^3	α_5	α_6	ρ_3^3	ρ_1	α_4	α_3	ι	ρ_2^2	μ_4^2	μ_4	μ_2	μ_2^2
α_2	α_2	μ_4	ρ_2^3	μ_2	ρ_3^2	α_1	ρ_1^2	μ_4^2	ρ_2	μ_2^2	α_3	α_5	ρ_1^3	ρ_3^3	α_6	α_4	ρ_1	ρ_3	ρ_2^2	ι	μ_1	μ_3^2	μ_1^2	μ_3
α_3	α_3	μ_2^2	ρ_3	μ_3	μ_2	ρ_3^3	μ_3^2	ρ_1^2	α_6	ρ_2^2	α_5	α_2	ρ_2	ρ_1	ρ_1^3	ρ_2^3	α_1	α_4	μ_4	μ_1^2	ι	μ_4^2	μ_1	ρ_3^2
α_4	α_4	ρ_3^2	α_5	ρ_2^2	μ_1	ρ_1^3	μ_2^2	μ_2	ρ_1	μ_1^2	ρ_2	ρ_3^3	ρ_3	ρ_2^3	α_2	α_6	α_3	α_1	μ_4^2	μ_3	μ_4	ι	ρ_1^2	μ_3^2
α_5	α_5	ρ_2^2	α_4	ρ_3^2	μ_3	ρ_1	μ_4^2	μ_3^2	ρ_1^3	μ_4	α_2	α_3	α_6	α_1	ρ_2	ρ_3	ρ_3^3	ρ_2^3	μ_2^2	μ_1	μ_1^2	ρ_1^2	ι	μ_2
α_6	α_6	μ_1	ρ_3^3	μ_4^2	μ_4	ρ_3	μ_1^2	ρ_2^2	α_3	ρ_1^2	ρ_1	ρ_2^3	α_1	α_5	α_4	α_2	ρ_2	ρ_1^3	μ_2	μ_3^2	ρ_3^2	μ_3	μ_2^2	ι

▷ **Quick Exercise.** Check several interesting computations in the group table. Having a cube to manipulate will help here. ◁

We have seen that all rigid motions in the symmetry groups of the tetrahedron and the cube are rotations about some line. In fact, all rigid motions of three-dimensional space are such rotations; we will not prove this interesting fact here.

Chapter Summary

In this chapter we extended our notion of symmetry to symmetry of objects in three-dimensional space. We then applied these ideas to obtain the symmetries of the regular tetrahedron and the cube.

Warm-up Exercises

a. We noted that a symmetry of the cube is determined if we specify the color of the front and right faces, after the motion. Is the symmetry determined when we specify the color of some different pair of faces? Does this work for *any* pair of faces?

b. Compute the following elements twice in the group of symmetries of the tetrahedron, once using the group table, and once using the representation as permutations:

$$\varphi_1{\rho_2}^2\varphi_2, \quad \varphi_2\varphi_1\varphi_2, \quad \rho_1\rho_3.$$

c. Explain the geometric meaning of the block substructure in the group table for the symmetries of the regular tetrahedron.

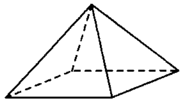
d. Compute the following elements twice in the group of symmetries of the cube, once using the group table, and once using the representation as permutations:

$$\alpha_3\rho_2{\mu_1}^2, \quad \rho_2\rho_1, \quad \alpha_1{\rho_1}^3\alpha_1.$$

e. Describe any block substructure that you detect in the group table for the cube.

Exercises

1. Find all symmetries of a pyramid as drawn below (the base is a square, and the four sides are congruent):



2. Show that ρ_1 and φ_1 *generate* the group of symmetries of the tetrahedron. This means that you can find a formula for each of the other symmetries, in terms only of ρ_1 and φ_1. (By all means, use our group table!)

3. Consider the *flatlanders*, who live in the plane and consequently cannot conceive of motion in three dimensions. Formulate a definition for symmetry in $\mathbb{R}^2$ for flatlanders, and then determine for them the group of symmetries of the equilateral triangle and the square.

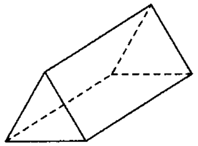
4. In our discussion of the symmetries of the tetrahedron, we excluded the symmetry corresponding to the permutation

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

because it could not be accomplished by a rigid motion of space. However, we did observe that it could be accomplished by a reflection through a plane; this is a called a *mirror reflection.* Show that *all* of the excluded permutations for the group of symmetries of the tetrahedron are mirror reflections. Specify the plane for each of these twelve permutations.

5. Are there any mirror reflections that leave the cube fixed? (See the previous exercise for a discussion of mirror reflections.) Can you determine how many such there are?

6. Find all symmetries of the 'tent':



The ends are equilateral triangles, and the sides are congruent non-square rectangles. Give the group table. Comment on its relationship to the group table for the triangle.

7. (For those comfortable with multiplying 3×3 matrices.) Consider the tetrahedron in $\mathbb{R}^3$ with vertices at

$$(1,1,-1), (-1,-1,-1), (1,-1,1), \text{ and } (-1,1,1).$$

If we label these vertices 1, 2, 3, 4, respectively, then this corresponds to the labeling of the tetrahedron we used in the text.

Notice that the matrix

$$F_1 = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

corresponds to the symmetry φ_1. Find a matrix which corresponds to the symmetry ρ_1, and then determine matrices for all the other symmetries, using Exercise 2 above.

8. Consider a cube with a special down face: This face is different, and consequently, a symmetry must leave it alone; the face might be rotated, but must remain the down face. (You might imagine a circular dot in the center of this face.) How many symmetries of the unmarked cube are still symmetries of this cube?

Chapter 24

Abstract Groups

In this chapter we intend to generalize the examples of the previous two chapters, to obtain an abstract notion of *group*, quite similar in flavor to our abstract definition of *ring* in Chapter 6. We have a set of objects (which up to now have been symmetries), equipped with a binary operation (which up to now has been the composition of symmetries). What abstract properties should this binary operation satisfy?

Because the composition of symmetries can be viewed as a composition of functions, this binary operation is clearly *associative.* We will include this in our abstract definition.

Notice that we always included the *identity* symmetry: the symmetry consisting of no motion at all. When composed with any other symmetry, we obtain the second symmetry. This clearly seems similar to our definitions of additive and multiplicative identities. We will include an identity element in our abstract definition.

Any symmetry can be undone; that is, there is a motion that restores the object in question to its original orientation. What does this mean in terms of the group table? For example, in the group D_3 of symmetries of the equilateral triangle, the motion that undoes the rotation ρ is simply ρ^2, and this fact is reflected in the group table by the fact that

$$\rho\rho^2 = \iota \quad \text{and} \quad \rho^2\rho = \iota.$$

So, the existence of *inverses* is the third requirement of our definition.

▷ **Quick Exercise.** Choose several elements in the groups of symmetries for the cube and tetrahedron. What are their inverses? Does this make sense geometrically? ◁

24.1 Definition of Group

We now state our definition formally. A **group** G is a set of elements on which one binary operation ($\circ$) is defined that satisfies the following properties: (The symbols g, h, and k represent any elements from G.)

(Rule 1) $(g \circ h) \circ k = g \circ (h \circ k)$

(Rule 2) There exists an element e in G such that

$$g \circ e = e \circ g = g$$

(Rule 3) For each element g in G there exists an element x so that

$$g \circ x = x \circ g = e$$

We introduce some terminology to describe these rules (which will seem familiar after our experience with rings): Rule 1 says that the operation $\circ$ is **associative**, Rule 2 says that an **identity** exists, and Rule 3 says that each element of the group has an **inverse**. We can thus paraphrase our definition by saying this: A group is a set with an associative binary operation with an identity, where all elements have inverses.

24.2 Examples of Groups

What are some examples of groups?

Example 24.1

The groups of symmetries for the triangle, square, tetrahedron, and cube, where the operation $\circ$ is functional composition. The operation (being functional composition) is associative, and the identity is the symmetry consisting of no motion (the *identity function*). The inverse of each element is also clearly a symmetry and consists geometrically of 'undoing' the symmetry. Function-theoretically, it is exactly the *inverse function* of the given symmetry.

Now let R be a ring, equipped with operations $+$ and $\cdot$. Let's forget the multiplication. We claim that R equipped with $+$ is a group. Obviously, $+$ is associative (Rule 2 in the definition of ring). Clearly, 0 plays the role of the identity for $+$ (Rule 3 in the definition of ring). And every element in a ring has an (additive) inverse, by Rule 4 in the definition of a ring.

We have thus provided ourselves with an extremely large fund of group examples, because we know about so many rings. It's probably worthwhile listing a few of these explicitly, so that you can better appreciate the ground covered by this observation.

Example 24.2

The integers $\mathbb{Z}$, equipped with $+$.

Example 24.3

The integers modulo m (that is, $\mathbb{Z}_m$), equipped with $+$, as defined in Chapter 3.

Example 24.4

The set $M_2(\mathbb{Z})$ of 2×2 matrices, with integer entries, equipped with matrix addition.

Example 24.5

The Gaussian integers $\mathbb{Z}[i]$, under addition.

When we look at the additive structure of a ring, and consider it as a group, we call it the **additive group of a ring**. If the only groups around occurred as the additive groups of rings, the abstract concept of group would clearly be superfluous. But by Rule 1 from the definition of ring, the binary operation addition for a ring is always *commutative*. That is, it satisfies the rule

$$a + b = b + a.$$

Note that we do *not* include such a requirement on the abstract operation $\circ$ in our definition of group above, because we wish to include those groups in Example 24.1.

Thus, the notion of group is a genuine generalization of the notion of additive group of a ring because it includes more structures. As a more general concept, we will in the next chapters discuss quite a distinct algebraic theory regarding groups.

If a group satisfies the additional rule

Rule 4 $a \circ b = b \circ a$

we say that the group is **abelian**. That is, a group with a commutative operation is abelian. The word 'abelian' honors the early 19th-century mathematician Abel, whom we mentioned in Chapter 9. (Although 'commutative group' might seem a reasonable term to use, we will follow long-standing practice and say 'abelian' instead.)

24.3 Multiplicative Groups

We need some more examples:

Example 24.6

The set $\mathbb{Q}^*$ of non-zero rational numbers, under multiplication, is a group. Multiplication is clearly an associative operation, and its identity is 1. Note that every such rational number has a (multiplicative) inverse: the multiplicative inverse of rational number a/b (where $a, b \neq 0$) is b/a.

Examples 24.2 through 24.5 might have led you to infer that we should invariably associate 'group operation' with addition, rather than multiplication. This is certainly false, as Example 24.6 and several examples below show.

We can generalize Example 24.6. Let F be any field. Then let F^* be the set of non-zero elements of F. Equip this set with its usual multiplication. Then this is a group.

▷ **Quick Exercise.** Why? ◁

Example 24.7

Let's create this group, for a particular field. Consider the field $\mathbb{Z}_7$. We are then claiming that the set

$$\mathbb{Z}_7^* = \{1, 2, 3, 4, 5, 6\}$$

of residue classes modulo 7 forms a group under multiplication. Because $\mathbb{Z}_7$ is a field, we know that all these elements have (multiplicative) inverses, but let's compute them explicitly anyway: Because it is the identity, clearly the inverse of 1 is 1. Because $2 \cdot 4 = 1$, 2 and 4 are inverses of one another; because $3 \cdot 5 = 1$, 3 and 5 are inverses of one another. And because $6^2 = 1$, 6 is its own inverse.

▷ **Quick Exercise.** Compute the inverses in the multiplicative group $\mathbb{Z}_{11}^*$. ◁

We can generalize even further. Suppose that R is any ring with unity. In Chapter 8 we defined $U(R)$, the set of those elements of R that are *units*; that is, $U(R)$ is the set of those elements of R that have multiplicative inverses. For a field F, $U(F)$ is just F^*, its set of non-zero elements.

But we claim that for *any* ring with unity, $U(R)$ is a group under multiplication. Because R has unity, the set $U(R)$ evidently possesses an identity. And clearly (by definition) every element of $U(R)$ has a (multiplicative) inverse. There remains a subtle point to verify, before we can claim that $U(R)$ is a group: Is in fact $U(R)$ *closed* under multiplication? That is, is multiplication on $U(R)$ a binary operation? The answer is yes, and we proved exactly this in Chapter 8.

▷ **Quick Exercise.** Carefully re-read this proof in Section 8.2. ◁

We will consequently call $U(R)$ the **group of units** for the ring (with unity) R.

▷ **Quick Exercise.** Why are we restricting ourselves to rings *with unity*? ◁

Thus, each ring with unity has associated with it *two* groups: its additive group, and its group of units. Let's look concretely at a few groups of units:

Example 24.8

The set $U(\mathbb{Z}_6) = \{1, 5\}$ is a group under multiplication. The multiplicative inverse of 5 is itself.

Example 24.9

Consider the group of units of the Gaussian integers:

$$U(\mathbb{Z}[i]) = \{1, -1, i, -i\}.$$

Note that -1 is its own inverse and i and $-i$ are inverses of each other.

Example 24.10

Consider the sets $U(M_2(\mathbb{Z}))$ and $U(M_2(\mathbb{R}))$; they are both groups, where the operation is matrix multiplication. Recall that the units of $M_2(\mathbb{Z})$ and $M_2(\mathbb{R})$ are those matrices with non-zero determinant. (See Exercise 8.2.) Both groups have infinitely many elements. For example,

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

are both elements of $U(M_2(\mathbb{Z}))$.

▷ **Quick Exercise.** What are their inverses? ◁

▷ **Quick Exercise.** Show that $U(M_2(\mathbb{Z}))$ is a non-abelian group. (The examples we've just given will do!) ◁

Example 24.11

Consider $U(\mathbb{Z}[\sqrt{2}])$. In Example 10.12 we computed infinitely many elements of this multiplicative group. We in fact gave a complete list of its elements (although we were unable to prove the completeness of this list).

▷ **Quick Exercise.** Verify that $7 + 5\sqrt{2}$ is an element of this group. (That is, compute its inverse.) ◁

We can have multiplicative groups living inside previously studied rings, which do not consist of the entire group of units. For a couple of such examples, consider the following.

Example 24.12

Let's work inside the field $\mathbb{Z}_{13}$. Take the element 4. We now compute the (multiplicative) powers of this element; to make our computations clear, we will this time make explicit use of modular notation:

$$[4]^1 = [4], \quad [4]^2 = [16] = [3], \quad [4]^3 = [4][3] = [12],$$
$$[4]^4 = [4][12] = [48] = [9], \quad [4]^5 = [4][9] = [10],$$
$$[4]^6 = [4][10] = [40] = [1].$$

We have stopped here, because if we continue, we will repeat elements we have already obtained. We now claim that

$$\{[1], [4], [3], [12], [9], [10]\} =$$
$$\{[1], [4]^1, [4]^2, [4]^3, [4]^4, [4]^5\}$$

is a group under multiplication. You can check this by brute force, but the fact that all elements are powers of [4], and that $[4]^6 = [1]$, means that computation of products and inverses is easy! For example,

$$[4]^3[4]^4 = [4]^7 = [4]^6[4] = [4].$$

And what's the inverse of $[4]^2$? We claim it is $[4]^{6-2}$; the computation

$$[4]^2[4]^4 = [4]^6 = [1]$$

makes this evident. We will return to this example (and generalizations) in Chapter 26.

Example 24.13

Let's work inside the field $\mathbb{C}$. Consider the set of all complex numbers of modulus 1. We call this set $\mathbb{S}$:

$$\mathbb{S} = \{\alpha \in \mathbb{C} : |\alpha| = 1\} = \{a + bi \in \mathbb{C} : a^2 + b^2 = 1\}.$$

Considered within the complex plane, this consists of exactly the points on the circle centered at the origin, of radius one. (You should certainly review our discussion of complex numbers in Chapter 8, if necessary.) We consequently call $\mathbb{S}$ the **unit circle**. We claim that $\mathbb{S}$ is a group, under complex multiplication. This set is certainly closed under multiplication because

$$|\alpha\beta| = |\alpha||\beta| = 1 \cdot 1 = 1.$$

And clearly, the multiplicative identity 1 belongs to S. But what about multiplicative inverses? If we express an element of $\mathbb{S}$ in trigonometric form, it looks like

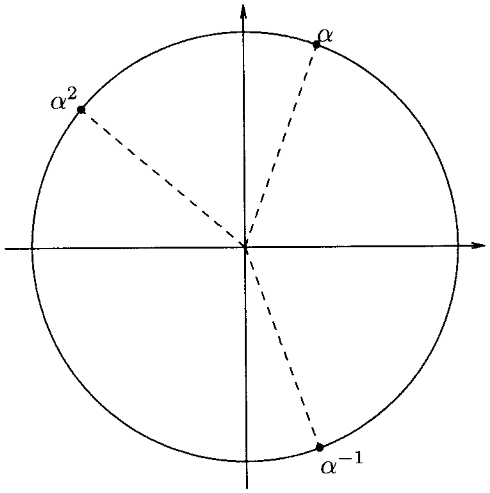
$$e^{i\theta} = \cos\theta + i\sin\theta.$$

And by DeMoivre's Theorem 8.4, its multiplicative inverse is exactly

$$e^{-i\theta} = \cos(-\theta) + i\sin(-\theta) = \cos\theta - i\sin\theta,$$

which still belongs to $\mathbb{S}$, as you can easily verify.

The diagram below shows graphically the values of α, α^2, and α^{-1} for an arbitrary complex number $\alpha \in \mathbb{S}$.



Example 24.14

We now work inside $M_2(\mathbb{R})$. Consider the following set of 2×2 matrices:

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix} \quad \begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}$$

$$\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \quad \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix} \quad \begin{pmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}$$

Then this is a group under matrix multiplication. To verify this explicitly, we would need to make sure each element has an inverse in the set (this is not too difficult), and also check that this set is closed under multiplication. If we were to do this by brute force, there would be a lot of multiplications to check! (How many?) However, this set of matrices should look familiar: It was a geometrically explicit realization of D_3, the group of symmetries of an equilateral triangle. (See Chapter 22.)

▷ **Quick Exercise.** Recall from Chapter 22 that multiplication of these matrices corresponds to composition of symmetries. With this observation, why must this set of matrices be closed under multiplication? ◁

Example 24.15

We now work inside $M_2(\mathbb{C})$, the ring of 2×2 matrices with complex entries. Consider the following set of 2×2 matrices:

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}, \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}$$

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

Then this is a group under matrix multiplication. It is a tedious matter to verify that this set is closed under multiplication, and we will not carry out this project. However, it would be worthwhile for you to try a few sample multiplications, to get a feel for how this group works:

▷ **Quick Exercise.** Do this. ◁

It is important to note that every element in this set has an inverse; you will verify that this is the case in Exercise 24.3. This group is non-abelian; it is known as the group of **quaternions**. We will denote by Q_8. In Exercise 28.14 we will introduce a more compact notation for the elements of Q_8.

It's worth noting that many sets equipped with a single operation do *not* satisfy the group axioms.

Example 24.16

> Consider the integers $\mathbb{Z}$ equipped with multiplication. This is not a group. Although the operation is associative, and there is an identity (namely, 1), almost none of the elements have inverses (with respect to this operation).

More generally, if we take a ring R and forget its addition, we will never obtain a group! To begin with, we wouldn't get a group unless we had a multiplicative identity (unity). But even if R has unity, we would then require that every element of R have a multiplicative inverse; yet 0 never has a multiplicative inverse! (The only exception here is the trivial example of the *zero ring*.)

Of course, this isn't the only way a set with a binary operation can fail to be a group.

Example 24.17

> The non-negative integers $\{0, 1, 2, \cdots\}$ is not a group under addition: Although it has an additive identity, all elements (except 0 itself) lack an inverse.

Example 24.18

> Let R be an arbitrary ring. We consider the set $\mathrm{Aut}(R)$ of all ring isomorphisms from R onto itself; such isomorphisms are called **automorphisms**. We claim that $\mathrm{Aut}(R)$ is a group, under functional composition. We need to check that this operation is well defined; that is, is the composition of two such isomorphisms itself an isomorphism? You will do this in Exercise 24.10, where you check that such a composition still preserves addition and multiplication and is one-to-one and onto. Next, note that for any ring R, the identity function $\iota : R \to R$ defined by $\iota(r) = r$ is in fact a ring isomorphism (see Example 19.1). This element will be the identity element in this group, because when we compose it with any other automorphism, we get the same function back. And in the discussion following Example 19.5, we described how the inverse function of a ring isomorphism is itself a ring isomorphism, and so each element of $\mathrm{Aut}(R)$ has a inverse. In Exercises 24.11–24.14 you will determine the elements of $\mathrm{Aut}(R)$, for certain specific rings.

Chapter Summary

In this chapter we have generalized the notion of group of symmetries, to obtain the concept of abstract *group*. We looked at many examples of groups, including two important classes arising from rings: the *additive group of a ring*, and the *group of units* of a ring with unity.

Warm-up Exercises

a. Give examples of the following:

   (a) An abelian group with infinitely many elements.

   (b) An abelian group with finitely many elements.

   (c) A non-abelian group with infinitely many elements.

   (d) A non-abelian group with finitely many elements.

b. Determine the inverses of the following elements, specifying the group operation in each case:

   (a) $[3] \in \mathbb{Z}_5$.

   (b) $[3] \in \mathbb{Z}_5^*$.

   (c) $\begin{pmatrix} 3 & 4 \\ -4 & -5 \end{pmatrix} \in U(M_2(\mathbb{Z}))$.

   (d) $\varphi\rho \in D_3$.

   (e) $(2, -4) \in \mathbb{Q} \times \mathbb{Q}$.

   (f) $(2, -4) \in U(\mathbb{Q} \times \mathbb{Q})$.

   (g) $\frac{3}{5} + \frac{4}{5}i \in \mathbb{S}$.

c. Give examples of the following (you need to specify both the group in which you are computing, as well as specific elements):

   (a) A non-identity element that is its own inverse.

   (b) Two elements a and b, so that $(a \circ b)^{-1} \neq a^{-1} \circ b^{-1}$.

   (c) A non-identity element a so that $a \circ a \circ a \circ a = 1$.

d. Let R be a commutative ring with unity. What two groups are associated with R?

e. Consider the element

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

in the group $U(M_2(\mathbb{R}))$. Why is it true that $AB = BA$, for all other elements B in this group? Does this mean that this group is abelian?

f. Consider the following sets, equipped with an operation. Are they groups, or not?

(a) $\mathbb{Z}$, with subtraction.

(b) $\mathbb{Z}$, with multiplication.

(c) $M_2(\mathbb{Z})$, with matrix addition.

(d) $M_2(\mathbb{Z})$, with matrix multiplication.

(e) $\mathbb{R}^*$, with multiplication.

(f) $\mathbb{R}^*$, with division.

(g) $\mathbb{R}^*$, with addition.

(h) $\mathbb{Z} \times \mathbb{Z}$, with the operation $*$, defined by $(a,b) * (c,d) = (a,d)$.

(i) The set of vectors in $\mathbb{R}^3$, with cross product.

(j) The set $\mathbb{R}^+$ of strictly positive real numbers, with multiplication.

g. The following table represents a binary operation on the set $\{a, b, c, d\}$. Argue that this set with this operation is not a group. (This fails to be a group for more than one reason.)

	a	b	c	d
a	a	b	d	c
b	b	a	c	d
c	c	d	b	a
d	d	c	a	b

Exercises

1. Suppose you have a set of n elements with a binary operation that you think might be a group. You easily check that there is an identity and that every element has an inverse, but you are now faced with showing the operation is associative. So, you check the equation $a \circ (b \circ c) = (a \circ b) \circ c$ for all possible a, b, and c. How many equations must you check? If $n = 5$, how many equations is this? If $n = 10$, how many?

2. Specify the elements of the groups of units of the following commutative rings:

$$\mathbb{Z}_{15}, \quad \mathbb{Q}[x], \quad \mathbb{Z}[x], \quad \mathbb{Z} \times \mathbb{Q}.$$

3. Consider the group of quaternions, described in Example 24.15. Determine explicitly which of these eight elements are inverses of one another. Also, show by example that this is not an abelian group.

4. Consider the set $\mathbb{R}\backslash\{-1\}$ of all real numbers except -1. Define the operation $*$ by

$$a * b = a + b + ab.$$

Prove that this is a group.

5. In this problem we consider permutations of the set $\mathbb{R}$.

(a) Let $S(\mathbb{R})$ denote the set of all real-valued functions $f : \mathbb{R} \to \mathbb{R}$, such that f is *one-to-one* and *onto*. Prove that $S(\mathbb{R})$ is a group, where the operation is functional composition.

(b) Now let $A(\mathbb{R})$ be the set of functions from $S(\mathbb{R})$ that are also **order-preserving**: By this we mean that if $x < y$, then $f(x) < f(y)$. Prove that $A(\mathbb{R})$ is a group under functional composition.

6. Consider the set of matrices of the form

$$\begin{pmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{pmatrix},$$

where a, b, c are arbitrary real numbers. Show that this set forms a group under matrix multiplication.

7. Let n be a positive integer and $\mathcal{C}$ be a circle. Now for $i = 0, 1, \ldots, n-1$, let ρ_i be the rotation of $\mathcal{C}$ counterclockwise through the angle $2\pi i/n$ radians. Show that this set of rotations is a group under the operation of composition. How many elements are in this group?

8. Let G be a group with operation $\circ$. Suppose that $x \circ x = 1$, for all $x \in G$. Prove that G is abelian.

9. Suppose that G is a group with operation $\circ$; suppose that $x, y \in G$. Show that if

$$(x \circ y) \circ (x \circ y) = (x \circ x) \circ (y \circ y),$$

then $x \circ y = y \circ x$.

10. Let R be any ring, and suppose that $\phi, \psi \in \mathrm{Aut}(R)$. Show that the composition $\phi\psi \in \mathrm{Aut}(R)$, by checking that this function has the appropriate domain and range, is one-to-one, onto, and preserves addition and multiplication. (This exercise verifies that $\mathrm{Aut}(R)$ is closed under functional composition; in Example 24.18 we complete the verification that $\mathrm{Aut}(R)$ is a group under this operation.)

11. Prove that $\mathrm{Aut}(\mathbb{Z})$ is a group with only a single element.

12. Show that $\mathrm{Aut}(\mathbb{Q})$ is a group with only a single element.

13. In this problem you will sketch the proof that $\mathrm{Aut}(\mathbb{R})$ is a group with only a single element. You will use the fact that all positive real numbers have exactly two square roots.

    (a) Let $a, b \in \mathbb{R}$. Show that $a \geq b$ if and only if $a - b = x^2$, for some $x \in \mathbb{R}$.

    (b) Use part a to show that if $\varphi \in \mathrm{Aut}(\mathbb{R})$, then $a \geq b$ if and only if $\varphi(a) \geq \varphi(b)$.

    (c) Argue that any automorphism of $\mathbb{R}$ is fixed on the rational numbers $\mathbb{Q}$. (See Exercise 12.)

    (d) You may assume that between any two real numbers is a rational number. Use this to prove that any automorphism of $\mathbb{R}$ is fixed on all real numbers, and so $\mathrm{Aut}(\mathbb{R})$ has only a single element.

14. Consider the field of complex numbers $\mathbb{C}$, and its group of automorphisms $\mathrm{Aut}(\mathbb{C})$. Show that this group has only two elements, namely the identity automorphism ι, and the complex conjugation map ϕ defined by $\phi(a + bi) = a - bi$. (See Example 16.4).

15. Let n be a positive integer; consider the set

$$G_n = \{1, -1\} \times \mathbb{Z}_n.$$

We define an operation on this set by

$$(i, [k]) * (j, [m]) = (ij, [m + jk]).$$

Prove that this makes G_n a group. *Note:* Do not neglect to show that this operation is associative! In Exercise 28.12 we will provide a more compact notation for this group, and show rigorously that it is 'essentially the same' as the dihedral group D_n we introduced in Chapter 22—the group of symmetries of a regular n-sided polygon.

Chapter 25

Subgroups

In this chapter we will follow closely our exposition of abstract rings, in Chapters 6 and 7. We first prove a theorem about arithmetic in an abstract group, quite similar to Theorem 6.1. We then introduce the idea of *subgroup*, which is quite analogous to the corresponding idea *subring.*

25.1 Arithmetic in an Abstract Group

Theorem 25.1 *Suppose that G is a group with operation $\circ$, and $g, h, k \in G$.*

a. (Cancellation on the right) If $g \circ k = h \circ k$, then $g = h$.

b. (Cancellation on the left) If $k \circ g = k \circ h$, then $g = h$.

c. (Solution of Equations) The equation $g \circ x = h$ always has a unique solution in G; likewise, the equation $x \circ g = h$ always has a unique solution in G.

d. (Uniqueness of inverse) Every element of G has exactly one inverse.

e. (Uniqueness of identity) There is only one element of G which satisfies the equations $z \circ g = g \circ z = g$ for all g; namely, the element e.

f. (Inverse of a product) The inverse of a product is the product of the inverses, in reversed order: $(g \circ h)^{-1} = h^{-1} \circ g^{-1}$.

Notice that we need to state both parts a and b, because in an arbitrary group the operation is not commutative: Hence, one of these statements does not immediately follow from the other. Likewise, the solutions guaranteed for the two equations in part c need not be the same.

▷ **Quick Exercise.** Consider these equations in the group of symmetries of the tetrahedron:

$$\varphi_1 x = \rho_3 \quad \text{and} \quad x\varphi_1 = \rho_3.$$

Discover the solutions to these equations (by examining the group table in Section 23.1); note that they are not the same. ◁

Notice that we have already encountered the rule in part f for computing inverses of products, when we discussed the multiplicative inverse of a product in a non-commutative ring; it might be worthwhile to re-read our discussion of this rule in Section 8.2.

Proof: (a) & (b): Merely operate on the appropriate side of the equation by an inverse of k.

(c): This is Exercise 25.2; your proof will be similar to part b of Theorem 6.1.

(d): This is Exercise 25.3.

(e): Suppose that e and z both serve as identities in G. Then $e = e \circ z$ (because z is an identity); but $e \circ z = z$, (because e is an identity). Thus $e = z$.

(f): Consider that

$$(g \circ h) \circ (h^{-1} \circ g^{-1}) = g \circ (h \circ h^{-1}) \circ g^{-1} = g \circ e \circ g^{-1} = g \circ g^{-1} = e.$$

▷ **Quick Exercise.** Similarly, show that $(h^{-1} \circ g^{-1}) \circ (g \circ h) = e$. ◁

Because inverses are unique, $(g \circ h)^{-1} = h^{-1} \circ g^{-1}$. □

25.2 Notation

Henceforth, when we discuss a generic abstract group, we will tend to denote its operation by juxtaposition, rather than by using the symbol $\circ$ for the operation; we will say that the group is being written *multiplicatively.* In this case, we will usually denote the (unique) identity by 1 and will denote the (unique) inverse of element g by g^{-1}. Note

that we have already used multiplicative notation for symmetry groups (this despite the fact that the group operation in symmetry groups is *functional composition*). As in those groups, $g^2 = g \circ g$, $g^3 = g \circ g \circ g$, and so on. In fact, when we use this multiplicative notation, we will often refer informally to the operation as multiplication.

If a generic abstract group is abelian, we will tend to denote its operation by $+$; we will say that the group is being written *additively.* In this case, we will usually denote the identity by 0 and will denote the inverse of element g by $-g$. We can in this situation talk about **subtraction**:

$$g - h = g + (-h).$$

However, for some abelian groups, it is still most natural to use multiplicative notation; for example, we will still use multiplicative notation for the abelian multiplicative group $\mathbb{Q}^*$ of non-zero rational numbers.

You should exercise extreme care in sorting out which sort of notation we use for a given group. General remarks, definitions, and theorems may be expressed multiplicatively, but they still apply for groups written additively. Note that for many sets equipped naturally with more than one operation (such as rings), only one of those operations makes the set a group. Be sure you know which!

25.3 Subgroups

It is often easier to check whether a given set is a group, if it is a subset of a larger group with the same operation. This is directly analogous to the ring theoretic situation, where we were led to a definition of *subring.* Similarly, we say that a subset H of a group G is a **subgroup** if H is itself a group under the operation induced from G.

We can immediately list some examples:

Example 25.1

> The additive group of integers $\mathbb{Z}$ is a subgroup of the additive group of reals $\mathbb{R}$.

We can generalize Example 25.1. If R is a subring of S, and we ignore the multiplication, then clearly the additive group R is a subgroup of the additive group S. This obviously provides us with a raft of additional examples of subgroups.

▷ **Quick Exercise.** List for yourself at least six interesting examples of subgroups arising in this way. ◁

Example 25.2

The unit circle $\mathbb{S}$ is a subgroup of the multiplicative group $\mathbb{C}^*$ of non-zero complex numbers. (See Example 24.13.)

Example 25.3

The set $\{1, 4, 3, 12, 9, 10\}$ is a subgroup of the multiplicative group of units $U(\mathbb{Z}_{13}) = \mathbb{Z}_{13}^*$. (See Example 24.12.)

Example 25.4

Let G be any group (written multiplicatively). Then $\{1\}$ and G itself are always subgroups of G. We call $\{1\}$ the **trivial subgroup** of G. G is the **improper subgroup** of G. All subgroups of G other than the improper subgroup G are **proper subgroups**.

Example 25.5

Note that because $\mathbb{Z}$ is an additive group, the trivial subgroup of this group is $\{0\}$.

▷ **Quick Exercise.** Is $\{1\}$ a subgroup of $\mathbb{Z}$? (Under what operation? Remember that when we say that H is a subgroup of G, we are implying that G is also a group.) ◁

Example 25.6

Consider the subset $\{\iota, \varphi\}$ of D_3. Because φ is its own inverse, it is easy to see that this is a subgroup.

25.4 Characterization of Subgroups

As with rings, we have a theorem that makes it easy to check whether a given set is a subgroup:

Theorem 25.2 The Subgroup Theorem *A non-empty subset H of a group G is a subgroup if and only if whenever $h, k \in H$, then $hk^{-1} \in H$.*

Note that in an abelian group (written additively), this says just that a non-empty subset of a group is a subgroup if and only if it is closed under *subtraction*.

Proof: This theorem is proved just like the corresponding theorem for rings (Theorem 7.1) and is Exercise 25.4. □

Note that Examples 7.6 through 7.9 are examples of this theorem in action (if you just ignore multiplication).

▷ **Quick Exercise.** Review these examples from Chapter 7. ◁

Here are some examples, which are not additive groups of rings:

Example 25.7

We work in the multiplicative group $\mathbb{Q}^*$. Let

$$H = \left\{ \frac{m}{n} : m, n \text{ are odd integers} \right\}.$$

We claim that H is a subgroup. Pick two typical elements of H, m/n and r/s, where m, n, r, and s are all odd integers. Then

$$\frac{m}{n}\left(\frac{r}{s}\right)^{-1} = \frac{m}{n}\frac{s}{r} = \frac{ms}{nr}.$$

But ms and nr are clearly odd integers, and so this is an element of H. By the theorem, H is a subgroup. (The most interesting thing about this conclusion is that H is a group at all!)

Example 25.8

We work in D_4, the group of symmetries of the square. Consider the set

$$\{\iota, \rho, \rho^2, \rho^3\}.$$

Note that $\rho^{-1} = \rho^3$, and $(\rho^2)^{-1} = \rho^2$. This means that any substitution into hk^{-1} by elements from this set will turn into a power of ρ and hence will belong to the set. We will generalize this example in the next chapter.

Chapter Summary

In this chapter we proved some elementary properties of the arithmetic of an abstract group. We then introduced the concept of *subgroup*, listed many examples, and proved a theorem which characterizes subgroups.

Warm-up Exercises

a. Which of the following are subgroups? (The operation on the larger group is always the obvious one.)

   (a) The subset of even integers in $\mathbb{Z}$.

   (b) The subset $\{0, 2, 4\}$ in $\mathbb{Z}_7$.

   (c) The subset $\{2^n : n \in \mathbb{Z}\}$ in $\mathbb{Q}$.

   (d) The subset $\{2^n : n \in \mathbb{Z}\}$ in $\mathbb{Q}^*$.

   (e) The subset $\{\iota, \rho\}$ in D_3.

   (f) The subset $\mathbb{N}$ of $\mathbb{Z}$.

b. Is every subring also a subgroup?

c. Is the group of units of a ring a subgroup of the ring?

d. How many identity elements can a group have? How many inverses can a given element of a group have?

e. Provide the (unique) solution to the following group equations:

   (a) $\mu_1 x = \rho_2$, in the group of symmetries of the cube.

   (b) $x\mu_1 = \rho_2$, in the group of symmetries of the cube.

   (c) $11 + x = 4$, in $\mathbb{Z}_{15}$.

   (d) $4x = 11$, in $U(\mathbb{Z}_{15})$.

Exercises

1. Consider the set
$$i\mathbb{R} = \{ai : a \in \mathbb{R}\} \subseteq \mathbb{C};$$
these are the **imaginary** numbers. Prove that this is a subgroup of the additive group of $\mathbb{C}$. Is I a subring of the ring $\mathbb{C}$? Similarly, show that $i\mathbb{Z} = \{ni : n \in \mathbb{Z}\}$ is a subgroup of the additive group of the Gaussian integers $\mathbb{Z}[i]$.

2. Prove Theorem 25.1c. That is, suppose that G is a group and $g, h \in G$. Prove that $gx = h$ has a unique solution; likewise, prove that $xg = h$ has a unique solution. (We have written the equations multiplicatively.)

3. Prove Theorem 25.1d. That is, prove that in a group, every element has exactly one inverse.

4. Prove the Subgroup Theorem 25.2: A non-empty subset H of a group G is a subgroup if and only if whenever $h, k \in H$, then $hk^{-1} \in H$.

5. Show that if H and K are subgroups of the group G, then $H \cap K$ is also a subgroup of G. Show by example that $H \cup K$ need not be a subgroup.(This exercise can and should be compared to Exercises 7.9 and 7.10.)

6. Suppose that G is a group, written multiplicatively. Let $g \in G$, and suppose that $g^2 = g$. Prove that g is the identity.

7. Let G be a group, and $a, b, c \in G$. Prove that the equation $axc = b$ has a unique solution in G.

8. Suppose that G is equipped with an associative operation $*$. Suppose that G has an element e so that $g * e = g$, for all $g \in G$; furthermore, for all $g \in G$, there exists an element $g' \in G$ so that $gg' = e$. Why are these assumptions apparently weaker than decreeing that G be a group? Prove, however, that these assumptions are sufficient to force G to be a group.

9. Show that if $(xy)^{-1} = x^{-1}y^{-1}$ for all x and y in the group G, then G is abelian.

10. Complete the following multiplication table so the following will be a group.

	a	b	c	d
a				
b				d
c			d	
d				

11. Find all subgroups of $U(\mathbb{Z}_8)$; of $\mathbb{Z}_7^*$; of $U(\mathbb{Z}_{15})$.

12. Show that $n\mathbb{Z}$ is a subgroup of the additive group of integers $\mathbb{Z}$, for all integers n.

13. Find all finite subgroups of the additive group $\mathbb{C}$. What can you say about all finite subgroups of the multiplicative group $\mathbb{C}^*$?

14. Argue *geometrically* that the dihedral group, D_n, has a subgroup of order n.

15. Let G be a group and $a \in G$. Define the **centralizer of** a to be

$$C(a) = \{g \in G : ga = ag\}.$$

That is, $C(a)$ consists of all the elements that commute with a.

(a) Find $C(\rho)$ in D_3.

(b) Find $C(4)$ in $\mathbb{Z}_7$.

(c) Show that $C(a)$ is a subgroup of G.

(d) Let H be a subgroup of G, and let

$$C(H) = \{g \in G : gh = hg \text{ for all } h \in H\};$$

call $C(H)$ the **centralizer of** H. Show that $C(H)$ is a subgroup of G.

16. Let $Z(G)$, the **center of** G, be the set of elements of G that commute with *all* elements of G.

(a) Find the center of the quaternions, defined in Example 24.15.

(b) Find the center of $\mathbb{Z}_5$.

(c) Show that $Z(G)$ is a subgroup of G.

(d) If $Z(G) = G$, what can you say about the group G?

17. If H is a subgroup of G, then show that $Z(G) \cap H$ is a subgroup of $Z(H)$.

18. Recall that the elements of $U(M_2(\mathbb{R}))$ are the real-valued 2×2 matrices with non-zero determinants. (See Exercise 8.2.) Show that the collection of such matrices with determinants equal to one is a subgroup of $U(M_2(\mathbb{R}))$.

19. Consider the elements of $U(M_2(\mathbb{R}))$ of the form

$$\begin{pmatrix} a & 0 \\ b & 1 \end{pmatrix},$$

where $a \neq 0$. Prove that this is a subgroup.

20. Show that the group given in Exercise 24.7 is a subgroup of $\mathbb{S}$, the group given in Example 24.13.

21. Suppose a and b are non-identity elements of a group G, that $ab = ba$ and $b^2 = 1$. Show that $\{a^n, ba^n : n \in \mathbb{Z}\}$ is a subgroup of G.

22. We generalize Exercise 21: Suppose that a and b are non-identity elements of a group G, that $ab = ba$ and $b^3 = 1$. Show that $\{a^n, ba^n, b^2a^n : n \in \mathbb{Z}\}$ is a subgroup.

23. Generalize the situation in the previous two exercises, replacing 2 and 3 by some positive integer m.

Chapter 26

Cyclic Groups

Suppose that G is a group (written multiplicatively), and $g \in G$. If we repeatedly multiply g by itself, we get the **powers** of g:

$$g^1 = g, \quad g^2 = gg, \quad g^3 = g(g^2), \quad g^4 = g(g^3), \quad \cdots.$$

Given an element g^n of this form, we call n an **exponent**; for now, we are restricting ourselves to exponents that are positive integers.

Example 26.1

> Choose the element 2 in $\mathbb{Q}^*$. Then the powers of 2 are the distinct elements
>
> $$2, \quad 4, \quad 8, \quad 16, \quad \cdots.$$
>
> Note that there are infinitely many distinct elements in this list.

Example 26.2

> Choose element ρ in D_3, the group of symmetries of the equilateral triangle. Then the powers of ρ are the repeating list of elements
>
> $$\rho, \quad \rho^2, \quad \rho^3 = \iota, \quad \rho^4 = \rho, \quad \cdots.$$
>
> Note that there are exactly three distinct elements in this list.

26.1 The Order of an Element

It turns out that all elements in a group behave in one or the other of the two ways illustrated by our examples above. We make this precise in the following theorem:

Theorem 26.1 *Suppose that G is a group and let $g \in G$. Then exactly one of the following two cases holds:*

a. *(The Torsion-free Case) All the powers*

$$g, \quad g^2, \quad g^3, \quad g^4, \quad \cdots$$

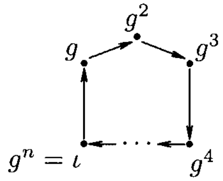
of g are distinct elements of G.

b. *(The Torsion Case) There exists a least positive integer n for which $g^n = 1$. In this case, the elements*

$$g, \quad g^2, \quad g^3, \quad \cdots, \quad g^{n-1}, \quad g^n = 1$$

are all distinct.

If an element falls into the second case, we say that it is a **torsion** element. The integer n we call its **order**. In this case, we denote the (finite) order of an element g by $o(g)$. Thus, in Example 26.2 above, $o(\rho) = 3$. If an element falls into the first case, we say that it is a **torsion-free** element and that it has **infinite order**. The word 'torsion' is intended to reflect the fact that in the second case the powers of the element cycle back on themselves:



Proof: Let G be a group, and $g \in G$. If all the powers of g are distinct, we obviously have the first case. Suppose, then, that not all powers of g are distinct. Then there exist positive integers r and s with $g^r = g^s$. We may as well assume that $r < s$. But now multiply both sides of this equation by r copies of g^{-1}:

$$1 = (g^{-1})^r(g^r) = (g^{-1})^r g^s = g^{s-r}.$$

This tells us that there exists some positive integer n (in this case, $s-r$) so that $1 = g^n$. By the Well-ordering Principle choose the smallest such n. We then claim that

$$g, \quad g^2, \quad g^3, \quad \cdots, \quad g^{n-1}, \quad g^n = 1$$

are all distinct elements. If not, then $g^r = g^s$, where $1 \leq r < s \leq n$. By the same argument as above, $g^{s-r} = 1$. But this is impossible because $s - r < n$, and we had chosen n to be the *least* such exponent. We have thus concluded that the second case holds, if the first fails. □

Example 26.3

The order of φ in the group of symmetries of the equilateral triangle is 2. Of course, we could then write $o(\varphi) = 2$.

Example 26.4

The order of the identity in any group is 1.

▷ **Quick Exercise.** What are the orders of the elements 1, 2, and 4 in $\mathbb{Z}_{13}^*$? ◁

Note that you have just illustrated the fact that non-identity elements in a given group need not have the same order.

▷ **Quick Exercise.** What is the order of the element -1 in the group $U(\mathbb{Z}[\sqrt{2}])$? What about the element $1 + \sqrt{2}$? ◁

The point of the previous Quick Exercise is this: It is possible that a group possess some elements with infinite order, and some elements (other than the identity) with finite order. The next two examples provide a somewhat more complicated illustration of this.

Example 26.5

The element $e^{\frac{\pi i}{5}} = \cos(\pi/5) + i\sin(\pi/5)$ in the unit circle $\mathbb{S}$ has order 10. (Use DeMoivre's Theorem 8.4.)

Example 26.6

Consider the element $\cos 1 + i \sin 1$ in the unit circle $\mathbb{S}$. (Here, the angle in question is 1 radian, or about 57 degrees.) We claim that this element is torsion-free. If not, there must exist a positive integer n for which

$$1 = (\cos 1 + i \sin 1)^n = \cos n + i \sin n.$$

Thus, n is an angle whose cosine is 1. Elementary trigonometry tells us that n must be an integer multiple of 2π: $n = 2\pi k$. But then $\pi = n/(2k)$. That is, π must be a rational number! This is certainly false, and is known as Lambert's theorem; we will discuss this fact further in Section 39.3. Thus, the element $\cos 1 + i \sin 1$ has infinite order.

26.2 Rule of Exponents

We will now define what we mean by **negative exponents** on an element in a multiplicative group. If n is a positive integer, we define g^{-n} to mean $(g^{-1})^n$. And by g^0, we mean the identity 1. The reason we make these definitions is exactly so that the ordinary rule of exponents works:

$$(g^r)(g^s) = g^{r+s},$$

for all integers positive, negative, or zero. In ordinary arithmetic, preserving the rule of exponents is exactly the reason why such computations as $3^{-2} = 1/9$ and $3^0 = 1$ are defined in the way they are.

Proof of Rule of Exponents: If r and s are both positive, this is clear, by the definition of positive exponents. If r and s are both negative, then

$$g^r g^s = (g^{-1})^{-r}(g^{-1})^{-s},$$

where the exponents $-r$ and $-s$ are both positive. This then reduces to the previous case. What if one exponent is positive and one is negative? Suppose that $r, s > 0$; then

$$g^r g^{-s} = g^r (g^{-1})^s.$$

We then cancel out terms until we run out of either g's or g^{-1}'s. In either case, we obtain g^{r-s}, as we require.

▷ **Quick Exercise.** There remains the case when r or s (or both) is 0; handle this. ◁

□

Suppose that g has finite order n; then

$$1 = g^n = g^{n-1}g.$$

Because the inverse of g is unique, $g^{-1} = g^{n-1}$. Consequently, any negative power of g can be expressed as a positive power, if g is a torsion element.

Example 26.7

Consider the element [3] in

$$U(\mathbb{Z}_{14}) = \{1, 3, 5, 9, 11, 13\}.$$

The order of this element is 6.

▷ **Quick Exercise.** Check that the order of [3] is 6. ◁

Thus, $[3]^{-1} = [3]^5$. We can check this directly:

$$[3]^5 = [243] = [14 \cdot 17 + 5] = [5],$$

and $[3][5] = [15] = [1]$.
Let's express $[3]^{-4}$ as a positive power of [3]:

$$[3]^{-4} = ([3]^{-1})^4 = ([3]^5)^4 = [3]^{20} = ([3]^6)^3[3]^2 = [1][3]^2 = [3]^2.$$

On the other hand, if an element g has infinite order, its inverse cannot be expressed as a positive power of g. Indeed, all the elements

$$\cdots,\ g^{-2},\ g^{-1},\ g^0 = 1,\ g^1 = g,\ g^2,\ g^3,\ \cdots$$

are distinct. The proof of this is a slight variation of the proof regarding the torsion-free case in Theorem 26.1; we leave it as Exercise 26.4.

▷ **Quick Exercise.** What is the complete list of all powers (positive, negative, or zero) of the element 2 in $\mathbb{Q}^*$? List all powers for $-4 < n < 4$ of the element $1 + \sqrt{2}$ in $U(\mathbb{Z}[\sqrt{2}])$. ◁

In case an element g has finite order, there is a simple relationship between $o(g)$ and *any* integer m for which $g^m = 1$. We use the Division Theorem 2.1 for $\mathbb{Z}$ to see this:

Theorem 26.2 *Suppose that g is an element of a group, with order n. Suppose also that $g^m = 1$, where m is some positive integer. Then n divides m.*

Proof: Let g be a group element, and $n = o(g)$. Suppose also that $g^m = 1$. Now n is the least positive integer so that $g^n = 1$. Thus, $n \leq m$. By the Division Theorem for the integers (Theorem 2.1), we can write $m = qn + r$, where q is a positive integer, and r is an integer with $0 \leq r < n$. But then

$$1 = g^m = g^{qn+r} = (g^n)^q g^r = g^r.$$

But because n is the least positive integer for which $g^n = 1$, we must have that $r = 0$. In other words, n divides m. □

In our discussion in this chapter we have up to now restricted ourselves to groups written multiplicatively. Of course, all our definitions and results apply to groups written additively as well, but the notation looks quite different.

To begin with, consider the positive powers of an element g of a group G, which is written additively. Our previous experience with rings supplies us with the appropriate notation:

$$g,\ g+g=2g,\ g+g+g=3g,\ \cdots,\ g+g+\cdots g=ng,\ \cdots.$$

If the order of g is infinite, then these elements are all distinct.

Example 26.8

The element 2 in the additive group $\mathbb{Z}$ has infinite order, because

$$2, \quad 4, \quad 6, \quad 8, \cdots$$

are all distinct.

▷ **Quick Exercise.** What is the order of -1 in the additive group $\mathbb{Z}[\sqrt{2}]$? Notice how different this question is than our earlier one about the order of -1 in the multiplicative group $U(\mathbb{Z}[\sqrt{2}])$. ◁

If the order of g is finite, adding g to itself finitely many times yields the identity, which in this case we denote by 0.

Example 26.9

The element 2 in the additive group $\mathbb{Z}_6$ has order 3, because the least n for which $n2 = 0$ is 3: $2 + 2 + 2 = 0$.

▷ **Quick Exercise.** Determine the orders of all elements in $\mathbb{Z}_6$. ◁

Note that if g has finite order in an additive group (say, $o(g) = n$), then $-g = (n-1)g$. (This, of course, is just the observation we made in multiplicative groups that $g^{-1} = g^{n-1}$.) That is, $-g$ is an integer multiple of g. For example, the order of [2] in $\mathbb{Z}_{24}$ is 12. Then $[(12-1)2] = [22] = -[2]$, the inverse of [2].

▷ **Quick Exercise.** What is the order of $3+2x$ in the additive group $\mathbb{Z}_6[x]$? Express its inverse as an integer multiple of itself, and check directly that this works. ◁

26.3 Cyclic Subgroups

Suppose now that g is an element of order n in the group G (written multiplicatively). Now consider the subset

$$\{1, g, g^2, \cdots, g^{n-1}\}$$

of G. Because all the negative powers of g can be expressed as a positive power, it is quite clear that this is a subgroup. Furthermore, it is evidently *the smallest subgroup of* G *that contains* g, because by closure of the operation all powers of g belong to any group containing g. We call this subgroup the **cyclic subgroup generated by** g, and denote it by $\langle g \rangle$.

Example 26.10

The cyclic subgroup generated by ρ in D_3 is $\{\iota, \rho, \rho^2\}$ and the cyclic subgroup generated by φ in this group is $\{\iota, \varphi\}$.

Example 26.11

The cyclic subgroup of $\mathbb{Z}_7^*$ generated by 2 is $\{1, 2, 4\}$ and that generated by 3 is the entire group.

Example 26.12

The cyclic subgroup of $\mathbb{C}^*$ generated by i is

$$\{1, i, -1, -i\}.$$

Example 26.13

The cyclic subgroup of $\mathbb{Z}_{12}$ generated by 4 is $\{0, 4, 8\}$ and that generated by 0 is the trivial subgroup $\{0\}$.

What if g is an element of infinite order in G? Then the smallest subgroup of G containing g must at least include all integer powers of g. In fact, in this case we set

$$\langle g \rangle = \{\cdots, g^{-2}, g^{-1}, 1, g, g^2, \cdots\}.$$

▷ **Quick Exercise.** Show that $\langle g \rangle$ as just defined is indeed a subgroup of G. ◁

Example 26.14

The cyclic subgroup generated by 2 in the group $\mathbb{Z}$ is the subgroup $2\mathbb{Z}$.

Example 26.15

The cyclic subgroup generated by π in the group $\mathbb{R}^*$ is the infinite set

$$\left\{\cdots, \frac{1}{\pi^2}, \frac{1}{\pi}, 1, \pi, \pi^2, \cdots\right\}.$$

We can subsume these two definitions of cyclic subgroup under a single expression: For any element g in the group G,

$$\langle g \rangle = \{g^m : m \in \mathbb{Z}\}.$$

Of course, if g is torsion-free, we obtain distinct elements for each choice of integer m. If g has finite order, each element of the cyclic subgroup is obtained by infinitely many choices of m.

Notice that we have used the same notation for cyclic subgroup that we have earlier used for principal ideal. This should cause no confusion, as long as you are certain whether we are working with groups or rings. This notational coincidence underlies the similarity of the concepts: The cyclic subgroup generated by an element is the smallest subgroup containing the element, while the principal ideal for an element is the smallest ideal containing the element.

26.4 Cyclic Groups

We say that a group G is itself **cyclic** if there exists some element $g \in G$ so that $\langle g \rangle = G$; that is, the cyclic subgroup of G generated by g is the entire group. A cyclic group with infinitely many elements is necessarily generated by an element of infinite order. A cyclic group with finitely many elements (say, n) is necessarily generated by an element with order n.

Example 26.16

The integers $\mathbb{Z}$ is a cyclic group, because $\mathbb{Z} = \langle 1 \rangle$.

Example 26.17

The group $\mathbb{Z}_m$ is cyclic, for any integer $m > 1$, because $\mathbb{Z}_m = \langle 1 \rangle$.

Example 26.18

The group $\mathbb{Z}_{13}^*$ is cyclic, with generator 2.

Note that in a cyclic group not all non-identity elements serve as generators.

▷ **Quick Exercise.** Give non-identity elements in the groups $\mathbb{Z}$ and in $\mathbb{Z}_{13}^*$ that do *not* generate the entire group. Can you do this in $\mathbb{Z}_5^*$? ◁

Example 26.19

The group D_3 is not cyclic because it has 6 elements, but it contains no element of order 6.

Example 26.20

> The group $U(\mathbb{Z}_8) = \{1, 3, 5, 7\}$ is not cyclic either; all non-identity elements have order 2. That is, there is no element of order 4.

Except for notational differences, it may by now seem that any cyclic group of a given order is essentially like any other of the same order and this is in fact the case. We will make this statement precise when we introduce the notion of *group isomorphism*, in the next chapter.

▷ **Quick Exercise.** Recall the definition of *ring isomorphism.* What do you suppose the definition of group isomorphism should be? ◁

Chapter Summary

In this chapter we analyzed the important notions of the *order* of an element of a group and saw that an element of a group can have either infinite or finite order. We also discussed the *cyclic subgroup* generated by an element.

Warm-up Exercises

a. What are the orders of the following elements?

(a) $5 \in \mathbb{Z}_{15}$.

(b) $7 \in U(\mathbb{Z}_{15})$.

(c) $\varphi\rho \in D_3$.

(d) $6 \in \mathbb{Z}$.

(e) $6 \in \mathbb{Q}^*$.

(f) $i \in \mathbb{S}$.

(g) $i \in \mathbb{C}$.

(h) $\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \in U(M_2(\mathbb{R}))$.

b. Are the following groups cyclic? Either explain why not or else specify a generator.

(a) $\mathbb{Z}$.

(b) $\mathbb{Q}$.

(c) The group of symmetries of the cube.

(d) $\mathbb{Z}_8$.

(e) $U(\mathbb{Z}_8)$.

(f) $M_2(\mathbb{Z})$.

(g) $\mathbb{Z}[i]$.

(h) $\{1, -1, i, -i\} \subseteq \mathbb{C}^*$.

(i) $\mathbb{Z} \times \mathbb{Z}$.

(j) $\mathbb{Z}_2 \times \mathbb{Z}_4$.

(k) $\mathbb{Z}_2 \times \mathbb{Z}_3$. (Be careful.)

c. Can a group possess (non-identity) elements of both infinite and finite order? Explain, or give an example.

d. Are all cyclic groups abelian?

e. Why do all non-abelian groups have (non-trivial) abelian subgroups?

f. How many different generators do the following cyclic groups possess? List all such generators.

(a) $\mathbb{Z}$.

(b) $\mathbb{Z}_{10}$.

(c) $\mathbb{Z}_7$.

(d) $U(\mathbb{Z}_9)$.

g. Explain why the order of g^{-1} is the same as g.

h. Suppose that $g^{-14} = 1$. What can you say about the order of g?

i. Suppose that $g^7 = g^{15}$. What can you say about the order of g?

Exercises

1. Determine the cyclic subgroups of $U(M_2(\mathbb{Z}))$ generated by

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \text{ and } \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

2. Prove that every subgroup of a cyclic group is cyclic.

3. Find an example to show that the converse of Exercise 2 is false: That is, give a non-cyclic group, each of whose proper subgroups is cyclic.

4. Suppose that g is an element of infinite order in a group G. Prove that no two distinct powers of g (with any integer exponent) are equal.

5. If a and b are elements of a group that commute and $\langle a \rangle \cap \langle b \rangle = \{1\}$, what is the order of ab if the order of a is m and the order of b is n? Prove your assertion. Show by example that your assertion is false in general, in the case that a and b do not commute.

6. If a and b are elements of a group whose orders are relatively prime, what can you say about $\langle a \rangle \cap \langle b \rangle$? Prove your assertion.

7. How many generators does an infinite cyclic group have?

8. Prove that if G is a finite cyclic group with more than two elements, then G has more than one generator.

9. (a) Show that in a cyclic group, the equation $x^2 = 1$ has no more than two solutions. (Of course, the identity is always one of the solutions.)

   (b) Give an example of a non-cyclic group where $x^2 = 1$ has more than two solutions.

10. (a) Show that if G is a cyclic group of order m and n divides m, then G has a subgroup of order n. (This subgroup will itself be cyclic.)

   (b) If G is an arbitrary group with order m and n divides m, then G need not have a cyclic subgroup of order n. Find two such examples, one where $n = m$, and one where $n < m$.

11. Generalizing the group given in Exercise 24.7, let $\rho_{i,n}$ be the rotation of a circle counterclockwise through an angle of $2\pi i/n$, for $i = 1, 2, \ldots, n-1$ and all $n \geq 2$. Let e be the 'identity' rotation. Show that this set of rotations forms a group under composition. Show that all elements in this group are torsion (even though the group itself is infinite). Show that each finite subgroup is cyclic.

12. Show that all finite subgroups of the group $\mathbb{S}$, given in Example 24.13, are cyclic.

13. If G is a finite group where every non-identity element is a generator of G, what can you say about the order of G? Prove your assertion.

Section V in a Nutshell

This section defines the abstract notion of group, after examining two important examples: symmetries of regular n-sided polygons (called the *nth dihedral groups*) and symmetries of the regular tetrahedron and the cube in 3-space.

A *group* G is a set of elements with one binary operation ($\circ$) that satisfies three rules:

1. $(g \circ h) \circ k = g \circ (h \circ k)$, for all $g, h, k \in G$,

2. There exists an element $e \in G$ (called the *identity* of G) such that $g \circ e = e \circ g = g$ for all $g \in G$, and

3. For each $g \in G$, there exists an element x (called the *inverse* of g) such that $g \circ x = x \circ g = e$.

In addition to the groups of symmetries mentioned above, other important examples of a group are the additive group of a ring (that is, the elements of a ring with only the addition considered) and the group of units of a ring with unity under multiplication. An *abelian* group is one in which the binary operation $\circ$ is commutative.

Groups enjoy cancellation on both the right and the left, and the solution of equations. Furthermore, the group identity is unique as is the inverse of each element. (All this is Theorem 25.1.)

Paralleling the idea of subrings is the idea of subgroup: If G is a group then a subset H of G is a *subgroup* if it is itself a group under the operation induced from G. To determine if H is a subgroup of a group G, we need only check that $hk^{-1} \in H$ for every $h, k \in H$ (Theorem 25.2).

An important class of subgroup of any group G is the *cyclic subgroup generated by* g, which we denote by $\langle g \rangle$. It consists of the powers of g. The subgroup $\langle g \rangle$ may be infinite or finite, depending on whether the order of g is infinite or finite. If G itself is generated by the powers of a single element, we say that G is a *cyclic group*.

VI

Group Homomorphisms and Permutations

Chapter 27

Group Homomorphisms

At the end of the last chapter we were in need of a way to relate one group to another by a function (to make precise the intuition that all cyclic groups of a given order are 'essentially the same'). We now make the appropriate definition of a *group homomorphism.* This definition is exactly what you should expect, given our earlier experience with *ring homomorphisms.*

27.1 Homomorphisms

In our definition we wish to be careful about where operations are taking place, and so we will depart from our usual practice of denoting the group operation by juxtaposition. Instead, we will denote the group operation by an explicit symbol. So, let G together with operation $\circ$, and H together with operation $*$, be groups. A function $\varphi : G \to H$ such that

$$\varphi(g \circ k) = \varphi(g) * \varphi(k),$$

for all $g, k \in G$ is a **group homomorphism**. Speaking more colloquially, a group homomorphism is a function between groups that preserves the group operation. Note that because g and k are elements of G, we are combining them via the operation $\circ$ in G. But $\varphi(g)$ and $\varphi(k)$ are elements of H, and so we are combining them via the operation $*$ in H.

27.2 Examples

It is time to look at some examples of group homomorphisms. We will begin with a general class of examples.

Namely, let R and S be rings, and $\varphi : R \to S$ a *ring homomorphism.* Now just look at the additive groups of R and S (so the operations on R and S are both just $+$). This function preserves addition, and so is a group homomorphism too. We mention specifically one important example of this type.

Example 27.1

The residue function $\varphi : \mathbb{Z} \to \mathbb{Z}_m$ is a ring homomorphism, and therefore a group homomorphism too.

▷ **Quick Exercise.** Now look at Examples 16.2, 16.3, and 16.4. Because they are ring homomorphisms, they are certainly group homomorphisms too. ◁

Example 27.2

Now consider Example 16.5. It is the function $\rho : \mathbb{Z} \to \mathbb{Z}$ defined by $\rho(n) = 3n$. Here, we are considering $\mathbb{Z}$ as an additive group. This function preserves addition:

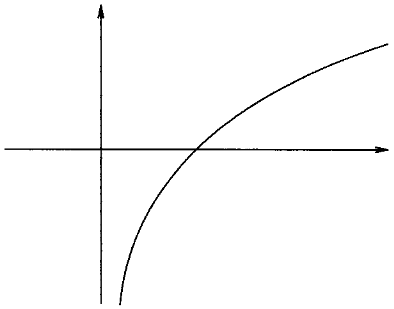
$$\rho(n+m) = 3(n+m) = 3n + 3m = \rho(n) + \rho(m).$$

Thus, ρ is a group homomorphism. In Example 16.5 we saw that this is not a ring homomorphism: It preserves addition but *not* multiplication.

It's time to consider some examples further afield from rings:

Example 27.3

Consider the groups $\mathbb{R}$ under addition, and $\mathbb{R}^+$ of positive real numbers, under multiplication. Recall the function $\log : \mathbb{R}^+ \to \mathbb{R}$, the *natural logarithm* function. (That is, $\log(r)$ is the exponent needed on the irrational number e so that $e^{\log(r)} = r$.)



The log function

Recall first of all that this function is only defined for positive real numbers. The most important and useful property of the logarithm function is this:

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

That is, the logarithm turns multiplication into addition. And this equation is exactly what is required to assert that log is a homomorphism! This example is well worth thinking about. It shows us that the group operations in two groups connected by a homomorphism can be quite different.

Example 27.4

Consider another famous function, this time between the groups $U(M_2(\mathbb{R}))$, the group of units of the 2×2 real-valued matrices, and $\mathbb{R}^*$, the multiplicative group of non-zero reals. Recall that $U(M_2(\mathbb{R}))$ are precisely those matrices in $M_2(\mathbb{R})$ with non-zero determinant. (See Exercise 8.2.) Here, the operation in the first group is matrix multiplication, while in the second it is ordinary real number multiplication. The function is the *determinant* function det:

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc.$$

Let's show that this is a homomorphism. For that purpose, we need to choose two arbitrary matrices,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} r & s \\ t & u \end{pmatrix},$$

where the entries are all real numbers. The product of these two matrices is

$$\begin{pmatrix} ar + bt & as + bu \\ cr + dt & cs + du \end{pmatrix},$$

and the determinant of this matrix is

$$ardu + btcs - bucr - asdt.$$

▷ **Quick Exercise.** Check these computations. ◁

But the product of the determinants is

$$(ad - bc)(ru - ts),$$

which is the same. Thus, the determinant function preserves multiplication. To paraphrase, the determinant of a product is the product of the determinants.

Example 27.5

Consider the group

$$D_3 = \{\iota, \rho, \rho^2, \varphi, \rho\varphi, \varphi\rho\}$$

of symmetries of the equilateral triangle, whose operation is functional composition. Consider also the multiplicative subgroup $\{1, -1\}$ of the integers. Define the function

$$\Phi : D_3 \to \{1, -1\}$$

given as follows: $\Phi(\alpha) = 1$ if α is one of the rotations ι, ρ, ρ^2; $\Phi(\alpha) = -1$ if α is one of the flips $\varphi, \varphi\rho, \rho\varphi$.

To see that this is a homomorphism, it is best to return to the group table we compiled for the symmetry group in Section 22.1. The pattern of F and R we observed there shows us that a rotation times a rotation is a rotation, a flip times a flip is a rotation, and a rotation times a flip (in either order) is a flip. Now replace 'rotation' by 1 and 'flip' by -1 in the previous sentence. This is just the way multiplication in the group $\{1, -1\}$ works!

Example 27.6

Consider the group G of symmetries of the tetrahedron, discussed in Section 23.1. Consider also the additive group $\mathbb{Z}_3$. Define the function

$$\Phi : G \to \mathbb{Z}_3$$

given as follows: $\Phi(\alpha) = 0$ if α is one of the four symmetries $\iota, \varphi_1, \varphi_2$, or φ_3; $\Phi(\alpha) = 1$ if α is one of the four symmetries

$\rho_1, \rho_2, \rho_3, \rho_4$; and $\Phi(\alpha) = 2$ if α is one of the four symmetries $\rho_1^2, \rho_2^2, \rho_3^2, \rho_4^2$.

To see that this is a homomorphism, it is best to return to the group table we compiled for the symmetry group in Section 23.1. The pattern we observed of F, R, and R^2 elements behaves in exactly the same way as addition in $\mathbb{Z}_3$!

▷ **Quick Exercise.** Check that the pattern observed in the group table in Section 23.1 is the same as addition in $\mathbb{Z}_3$. ◁

Thus, we have a group homomorphism from a group where the operation is functional composition, to a group where the operation is addition (modulo 3).

Example 27.7

Consider the function $D : \mathbb{R}[x] \to \mathbb{R}[x]$ defined by $D(f) = f'$ (the derivative of the polynomial f). Of course, we know that the derivative of a polynomial is a polynomial. But we also know that the derivative of a sum is the sum of the derivatives. That is,

$$D(f + g) = D(f) + D(g).$$

This is exactly what is required to show that this function is a group homomorphism. Notice, however, that this function *does not* preserve multiplication (the product rule is not that simple), and so this function is *not* a ring homomorphism. (For more about the derivative function, see Exercise 4.7.)

27.3 Direct Products

Let's recall a construction we made in rings: Given two groups G and H, consider the set

$$G \times H = \{(g, h) : g \in G, h \in H\}$$

of all ordered pairs, with first entry an element of G and second entry an element of H. Equip this set with the component-wise operation

$$(g_1, h_1)(g_2, h_2) = (g_1g_2, h_1h_2).$$

This makes $G \times H$ a group, called the **direct product** of G and H.

▷ **Quick Exercise.** Verify that this is a group. ◁

▷ **Quick Exercise.** Review some examples of direct products of rings. By forgetting the multiplication, they become examples of direct products of groups. ◁

We can now use the direct product to construct two important examples of homomorphisms:

Example 27.8

Let G and H be groups. Consider the function

$$\epsilon : G \to G \times H$$

defined by $\epsilon(g) = (g, 1)$. This is certainly a homomorphism:

$$\epsilon(g_1 g_2) = (g_1 g_2, 1) =$$

$$(g_1, 1)(g_2, 1) = \epsilon(g_1)\epsilon(g_2).$$

Note that this function is one-to-one, but certainly not onto (as long as H has more than one element).

▷ **Quick Exercise.** Check these claims. ◁

Of course, we could define a similar homomorphism on the second component. These homomorphisms are called **embeddings**.

Example 27.9

Let G and H be groups. Consider the function

$$\pi : G \times H \to G$$

defined by $\pi(g, h) = g$. This is called the **projection** onto the first coordinate.

▷ **Quick Exercise.** Check that this is an onto homomorphism. It is not one-to-one (as long as H has more than one element). ◁

A group homomorphism certainly need not be an onto function (see Example 27.2). However, if $\varphi : G \to H$ is not onto, the new function (which we still call φ) obtained by restricting the range to $\varphi(G)$ certainly is onto. Thus, $\varphi : G \to \varphi(G)$ is an onto homomorphism, assuming that $\varphi(G)$ is in fact a subgroup of H. This is the case:

Theorem 27.1 *Let $\varphi : G \to H$ be a homomorphism between groups G and H. Then $\varphi(G)$ is a subgroup of H.*

Proof: This is Exercise 27.6. Model your proof on the corresponding ring theory theorem (Theorem 16.2). □

Chapter Summary

In this chapter we introduced the idea of *group homomorphism*, a function between two groups that preserves the group operation, and we looked at numerous examples. We also introduced the direct product of two groups along with the closely connected homomorphisms *embeddings* and *projections*.

Warm-up Exercises

a. If $\varphi : G \to H$ is a group homomorphism and G is an additive group, need H be an additive group?

b. Is a ring homomorphism necessarily a group homomorphism?

c. Suppose we have a group homomorphism between the additive groups of two rings. Is this necessarily a ring homomorphism too?

d. Check that the determinant function preserves multiplication for two interesting matrices of your choice.

e. Does the determinant function preserve addition? Try it, for two interesting matrices of your choice.

f. We proved that the determinant function preserves multiplication for any pair of 2×2 matrices that are units. Show that this proof works for *any* pair of 2×2 matrices.

g. Does the observation in Exercise f mean that the determinant function defined from $M_2(\mathbb{R})$ onto $\mathbb{R}$ is a *ring* homomorphism? *Hint:* What does Exercise e say?

h. How many polynomials are sent to the polynomial 2 by the derivative homomorphism? Can you describe them all efficiently?

i. Find a one-to-one group homomorphism φ from $\mathbb{Z}_2$ to $\mathbb{Z}_4$. What is the subgroup $\varphi(\mathbb{Z}_2)$?

j. Let G and H be groups with operations $\circ$ and $*$, respectively, and $\varphi : G \to H$ a group homomorphism. Explain why $\varphi(g \circ h \circ k) = \varphi(g) * \varphi(h) * \varphi(k)$.

k. Generalize Exercise j to a product of n terms. Does this mean that $\varphi(g^n) = (\varphi(g))^n$?

Exercises

1. Show that the function $\varphi : \mathbb{Z}[i] \to \mathbb{Z}$ defined by $\varphi(a + bi) = a$ is a group homomorphism (although it is not a ring homomorphism).

2. Can you find a group homomorphism from $\mathbb{Z}_2 \times \mathbb{Z}_2$ onto the multiplicative group $\{1, -1, i, -i\}$?

3. Find three distinct group homomorphisms from $\mathbb{Z}_2 \times \mathbb{Z}_2$ onto $\mathbb{Z}_2$. How many more homomorphisms exist, if we remove the requirement that they be onto?

4. Let G be an abelian group (written additively). Define
$$\psi : G \times G \to G$$
by $\psi(g, h) = g + h$. Make a two-column chart showing what this function does to each element of $G = \mathbb{Z}_3$. Prove that ψ is a homomorphism, for any abelian group G.

5. Show by example that the corresponding homomorphism ψ defined in Exercise 4 need *not* be a homomorphism, when the group is not abelian.

6. Prove Theorem 27.1: That is, let $\varphi : G \to H$ be a homomorphism between the groups G and H. Prove that $\varphi(G)$ is a group.

7. Suppose that G is an abelian group, and φ is a group homomorphism whose domain is G. Prove that $\varphi(G)$ is abelian.

8. Suppose that R and S are rings with unity, and $\varphi : R \to S$ is an onto ring homomorphism. Consider the function $\psi : U(R) \to S$ defined by $\psi(r) = \varphi(r)$. (That is, ψ is just φ restricted to the group of units $U(R)$.) Prove that ψ is a group homomorphism from $U(R)$ into $U(S)$.

9. We know that $\log : \mathbb{R}^+ \to \mathbb{R}$ is a group homomorphism; see Example 27.3. Find a subgroup G of $\mathbb{R}^+$, so that the restricted logarithm function maps G onto the subgroup $\mathbb{Z}$ of $\mathbb{R}$. What about a subgroup of $\mathbb{R}^+$ that maps onto the subgroup $\mathbb{Q}$ of $\mathbb{R}$?

10. Consider the set G of all differentiable real-valued functions. Why is this a group under addition? Is the derivative function D a group homomorphism from G to G? *Warning:* That the function preserves the operation is *not* the issue.

11. Consider the homomorphism in Example 27.5. Define the analogous homomorphism from $D_4 \to \{1, -1\}$, making use of your work in Exercise 22.1.

12. In this problem we consider group homomorphisms from $\mathbb{Z}_n$ to itself.

    (a) Find all homomorphisms from $\mathbb{Z}_5$ into $\mathbb{Z}_5$. Now find all homomorphisms from $\mathbb{Z}_6$ into $\mathbb{Z}_6$.

    (b) Suppose G is a cyclic group of order n. How many homomorphisms are there from G into G? Describe them.

13. In this exercise we generalize the ideas encountered in Exercise 12. A homomorphism from a group into itself is called an **endomorphism** of the group. For an abelian group G, let $\mathrm{End}(G)$ be the set of all its endomorphisms $\varphi : G \to G$.

    (a) Let $\varphi, \psi \in \mathrm{End}(G)$. Define the sum $\varphi + \psi$ of these two homomorphisms by
    $$(\varphi + \psi)(g) = \varphi(g) + \psi(g).$$
    Prove that $\varphi + \psi$ is an endomorphism of G.

    (b) Prove that $\mathrm{End}(G)$, under the addition defined in part a, is an abelian group.

(c) Let $\varphi, \psi \in \mathrm{End}(G)$. Define the product $\varphi\psi$ as the ordinary functional composition:

$$\varphi\psi(g) = \varphi(\psi(g)).$$

Show that $\mathrm{End}(G)$ is a ring with unity when equipped with this multiplication. We call $\mathrm{End}(G)$ the **endomorphism ring** for G.

(d) Let n be a positive integer. *Beware:* in this problem we will consider $\mathbb{Z}_n$ as an abelian group, and as a commutative ring. Prove that the ring $\mathrm{End}(\mathbb{Z}_n)$ is isomorphic as a ring to $\mathbb{Z}_n$.

Chapter 28

Group Isomorphisms

We call a group homomorphism an **isomorphism** if it is a one-to-one and onto function. Intuitively, a group isomorphism preserves all essential group theoretic properties, and so demonstrates that two groups are 'essentially the same'. If $\varphi : G \to H$ is a group isomorphism between the groups G and H, the *inverse function* $\varphi^{-1} : H \to G$ is also a group isomorphism.

▷ **Quick Exercise.** Why is this? *Hint:* Re-read our discussion of this situation in the context of ring isomorphisms, in Chapter 19. ◁

Thus, the relationship of isomorphism is *symmetric*, and so we can speak unambiguously of two groups being **isomorphic**, if there exists an isomorphism between them.

Which of the examples we looked at in the previous chapter are isomorphisms?

▷ **Quick Exercise.** Check that the homomorphisms in Examples 27.1, 27.4, 27.5, 27.6, and 27.7 are not one-to-one and consequently are not isomorphisms. The Quick Exercise following Example 27.9 discusses when that homomorphism is one-to-one. ◁

The homomorphism in Example 27.2 (multiplication by 3, from $\mathbb{Z}$ to $\mathbb{Z}$) is certainly one-to-one, but it is not onto, and consequently not an isomorphism. However, the restricted homomorphism from $\mathbb{Z}$ onto $3\mathbb{Z}$ certainly is an isomorphism. Notice that this says that the group $\mathbb{Z}$ is isomorphic to a proper subgroup of itself.

▷ **Quick Exercise.** To what other proper subgroups is $\mathbb{Z}$ isomorphic? ◁

The homomorphism ϵ from Example 27.8 is certainly one-to-one, but usually not onto. We can assert that the group G is isomorphic to the subgroup

$$\epsilon(G) = \{(g, 1) : g \in G\}$$

of the full direct product.

It remains to consider Example 27.3: the *logarithm* function. Is it one-to-one? Suppose that $\log(r) = \log(s)$. But then

$$r = e^{\log(r)} = e^{\log(s)} = s,$$

which shows that the function is one-to-one. Is it onto? Choose any real number x. Then $\log(e^x) = x$, and so the function is onto as well.

The more elegant way to formulate the proof in the previous paragraph is this: Of course, log is one-to-one and onto, because it possesses an inverse function, namely, the *exponential* function $\exp(x) = e^x$. This function is the corresponding isomorphism

$$\exp : \mathbb{R} \to \mathbb{R}^*.$$

The group-theoretic conclusion to this discussion is perhaps surprising. The additive group of all real numbers is algebraically 'essentially the same' as the multiplicative group of all positive real numbers!

28.1 Structure Preserved by Homomorphisms

It appears that in general the only way to see that two groups are isomorphic is somehow to construct an isomorphism between them. But it is often easy to see that two groups are *not* isomorphic. For one thing, because there exists a one-to-one onto function between them, they must have the same number of elements: A group with 30 elements could not be isomorphic to a group with 24 elements. But there are other group properties preserved by isomorphisms (and, for that matter, homomorphisms). We record some of these properties in the next theorem:

Theorem 28.1 *Let $\varphi : G \to H$ be a homomorphism between the groups G and H, and let 1_G and 1_H be the identities of G and H, respectively.*

a. $\varphi(1_G) = 1_H$.

b. $\varphi(g^{-1}) = (\varphi(g))^{-1}$, for any $g \in G$.

c. Suppose that $g \in G$, and g has finite order. Then the order of $\varphi(g)$ divides the order of g. If φ is an isomorphism, then $o(g) = o(\varphi(g))$.

d. Suppose that G is abelian. Then $\varphi(G)$ is abelian.

Proof: Parts (a) and (b) are proved just as in the ring case (Theorem 16.1, parts a and b); we leave these as Exercise 28.6.

(c): Let $g \in G$, and suppose that $o(g) = n$. Then

$$1 = \varphi(g^n) = (\varphi(g))^n.$$

But then by Theorem 26.2, $o(\varphi(g))$ divides $n = o(g)$, as claimed. If φ is an isomorphism, it possesses an inverse function φ^{-1} which is also an isomorphism. Applying the homomorphism result we just proved to φ^{-1} tells us that $o(g)$ divides $o(\varphi(g))$ too, and because these are both positive integers, $o(g) = o(\varphi(g))$.

(d): This is Exercise 27.7. □

Example 28.1

Consider the groups D_3 and $\mathbb{Z}_6$; they both have the same number of elements. But they cannot be isomorphic, because the second group has an element of order 6, while the first doesn't. But regardless of the order of the elements in each group, the second group is abelian, while the first is not, so they can't be isomorphic.

Example 28.2

The groups $\mathbb{Z}_4$ and $\mathbb{Z}_2 \times \mathbb{Z}_2$ are both abelian groups with four elements. But they are not isomorphic, because the first group has an element of order 4, while the second does not.

28.2 Uniqueness of Cyclic Groups

We are now ready to clear up the unfinished business of Chapter 26 of showing that all cyclic groups of a given order are 'essentially the same':

Theorem 28.2 *All infinite cyclic groups are isomorphic to the additive group $\mathbb{Z}$. All cyclic groups of order n (where n is an integer greater than 1) are isomorphic to the additive group $\mathbb{Z}_n$.*

Proof: Suppose that G is an infinite cyclic group (which we may as well write multiplicatively). Then G has a generator g. That is,

$$G = \{g^m : m \in \mathbb{Z}\},$$

where if $r \neq s$, then $g^r \neq g^s$. Define the function

$$\varphi : G \to \mathbb{Z}$$

by $\varphi(g^m) = m$. (This looks like a symbolic version of the logarithm function because it picks off exponents!) The rule of exponents says that this function takes the operation in G to addition in $\mathbb{Z}$. It is also clearly one-to-one and onto.

$\triangleright$ **Quick Exercise.** Why is φ one-to-one and onto? $\triangleleft$

Suppose now that G is a cyclic group of order n. Then we know that G has a generator g, so that

$$G = \{1, g, g^2, \cdots, g^{n-1}\},$$

and $g^n = 1$. Define the function

$$\varphi : G \to \mathbb{Z}_n$$

by $\varphi(g^m) = [m]$. Let's show that this is a homomorphism. So choose $g^r, g^s \in G$; here, $0 \leq r, s \leq n-1$. Then

$$\varphi(g^r g^s) = \varphi(g^{r+s}).$$

If $r + s \leq n - 1$, then the value of the function is

$$[r+s] = [r] + [s] = \varphi(g^r) + \varphi(g^s),$$

as required. Otherwise, $n \leq r + s \leq 2n - 2$, and so

$$\varphi(g^{r+s}) =$$

$$\varphi(g^n g^{r+s-n}) = \varphi(1 \cdot g^{r+s-n}) = \varphi(g^{r+s-n}) =$$

$$[r+s-n] = [r+s] = [r] + [s],$$

again as required. □

This last theorem is a good example of a recurrent theme in algebra: We have characterized the concrete example $\mathbb{Z}$ as the apparently abstract construct 'infinite cyclic group'. This provides us real insight into the essence of the group structure of $\mathbb{Z}$.

Let's return to the examples of cyclic groups with which we ended Chapter 26.

For example, the cyclic subgroup of the multiplicative group $\mathbb{Z}_7^*$ generated by 2 has three elements. Likewise, so does the cyclic subgroup of the additive group $\mathbb{Z}_{12}$ generated by 4. Consequently, both these groups are isomorphic to $\mathbb{Z}_3$, and hence to each other.

$\triangleright$ **Quick Exercise.** Write down explicit isomorphisms between each pair of these three groups. $\triangleleft$

The cyclic subgroup of $\mathbb{Z}$ generated by 2 is infinite. Likewise, so is the cyclic subgroup of $\mathbb{R}^*$ generated by π. Therefore, both of these groups are isomorphic to $\mathbb{Z}$.

$\triangleright$ **Quick Exercise.** Write down explicit isomorphisms between each pair of these three groups. $\triangleleft$

As a consequence of this isomorphism theorem, when we wish to discuss a cyclic group in the abstract and wish to use multiplicative notation, we will tend to denote it by

$$\langle a \rangle = \{\cdots, a^{-2}, a^{-1}, 1, a, a^2, \cdots\}$$

(if it is infinite), or

$$\langle a \rangle = \{1, a, a^2, \cdots, a^{n-1}\}$$

(if it is finite). We won't worry concretely about what a is, because *any* such cyclic group is essentially the same.

28.3 Symmetry Groups

Another place where the idea of isomorphism can make precise an earlier discussion is in Chapter 22. There we recognized informally that the group D_3 of symmetries of the equilateral triangle could be put in a one-to-one correspondence with the set of matrices

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{pmatrix}, \quad \begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix},$$

$$\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}, \quad \begin{pmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}.$$

We observed in Example 24.14 that this set of matrices is a subgroup of $M_2(\mathbb{R})$. We can now say formally and precisely that this group of matrices is *isomorphic* to the group of symmetries D_3.

Recall that in Chapter 22 we also placed this group of symmetries in one-to-one correspondence with a set of *permutations* of the vertices of the triangle and noted that there seems to be an algebraic connection there also. We will make this precise in the next chapter, where we discuss *permutation groups*, in an abstract setting.

28.4 Characterizing Direct Products

For another illustration of the concept of isomorphism, we return to the notion of direct product. Suppose that a group G is isomorphic to a direct product of the two groups H_1 and H_2. Let's suppose that

$$\varphi : H_1 \times H_2 \to G$$

is the isomorphism. Consider the two subgroups of the direct product

$$H_1 \times \{1\} = \{(a, 1) : a \in H_1\}$$

and

$$\{1\} \times H_2 = \{(1, b) : b \in H_2\}.$$

These subgroups are obviously isomorphic to H_1 and H_2, respectively.

▷ **Quick Exercise.** Give an isomorphism between $H_1 \times \{1\}$ and H_1 and an isomorphism between $\{1\} \times H_2$ and H_2. ◁

Because the structure of the group G is 'essentially the same' as the structure of $H_1 \times H_2$, G must clearly have two corresponding subgroups: They are

$$G_1 = \varphi(H_1 \times \{1\})$$

and

$$G_2 = \varphi(\{1\} \times H_2).$$

Can we characterize abstractly the fact that G is a direct product, in terms of the properties of G_1 and G_2? It turns out we can, and we do this in the next theorem. But first, let's look at an example.

Example 28.3

The group $\mathbb{Z}_6$ is isomorphic to the direct product $\mathbb{Z}_3 \times \mathbb{Z}_2$ (of course, in this example the group operation is *addition*). The isomorphism is

$$\varphi([a]_3, [b]_2) = [4a + 3b]_6.$$

(We have included the subscripts for emphasis, and will henceforth omit them.)

▷ **Quick Exercise.** Check that this is a group isomorphism. ◁

But then

$$G_1 = \varphi(\mathbb{Z}_3 \times \{[0]\}) = \{0, 2, 4\}, \text{ and}$$

$$G_2 = \varphi(\{[0]\} \times \mathbb{Z}_2) = \{0, 3\}.$$

We'll return to this example when we have looked at our theorem.

To state our theorem conveniently, we need a little notation. If G_1 and G_2 are subgroups of a group G, then

$$G_1G_2 = \{g_1g_2 : g_1 \in G_1, g_2 \in G_2\}.$$

That is, G_1G_2 consists of all possible products, where the first factor comes from G_1 and the second factor comes from G_2.

Theorem 28.3 The Internal Characterization Theorem
The group G is isomorphic to $H_1 \times H_2$ if and only if G has two subgroups G_1 and G_2, so that

a. G_1 is isomorphic to H_1 and G_2 is isomorphic to H_2.

b. $G = G_1G_2$,

c. $G_1 \cap G_2 = \{1\}$, and

d. every element of G_1 commutes with every element of G_2.

Before we proceed, note that in Example 28.3 the properties a, b, c, and d hold.

▷ **Quick Exercise.** Check that properties a, b, c, and d hold. (Remember that our operation is addition.) ◁

Proof: Because this thoerem is 'if and only if', we have to prove both directions. Suppose first that G is isomorphic to the direct product of H_1 and H_2, via the isomorphism φ, as above. As before, we can define subgroups G_1 and G_2 and then part (a) is clearly satisfied.

To check part (b), let $g \in G$. Because the isomorphism φ is onto, there exist $h_1 \in H_1$ and $h_2 \in H_2$ so that $\varphi(h_1, h_2) = g$. But then

$$g = \varphi((h_1, 1)(1, h_2)) = \varphi(h_1, 1)\varphi(1, h_2) \in G_1G_2,$$

as required.

To check part (c), suppose that $g \in G_1 \cap G_2$. Then $g = \varphi(h_1, 1)$, and $g = \varphi(1, h_2)$, for some elements $h_1 \in H_1$, and $h_2 \in H_2$. But then

$$1 = \varphi(h_1, 1)(\varphi(1, h_2))^{-1} = \varphi(h_1, h_2^{-1}).$$

Because φ is one-to-one, $h_1 = 1$ and $h_2 = 1$, and so $g = 1$ as required.

To check part (d), suppose that $g_1 \in G_1$ and $g_2 \in G_2$. Then $g_1 = \varphi(h_1, 1)$ and $g_2 = \varphi(1, h_2)$. But such elements clearly commute.

▷ **Quick Exercise.** Why do such elements commute? ◁

For the converse, suppose that G has two subgroups G_1 and G_2 as specified in parts (a) to (d) above. We then define

$$\psi : G_1 \times G_2 \to G$$

by setting $\psi(g_1, g_2) = g_1g_2$. We claim that this is an isomorphism, thus proving the theorem. But ψ is a homomorphism, because of part (d); ψ is onto, because of part (b); and ψ is one-to-one, because of part (c). (You will check these three assertions in Exercise 28.4.) □

▷ **Quick Exercise.** Show that $\mathbb{Z}_{10}$ is isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_5$, by using the two subgroups $\langle 5 \rangle$ and $\langle 2 \rangle$, and the theorem. *Hint:* Examine Example 28.3. ◁

Example 28.4

We work in the additive group of Gaussian integers $\mathbb{Z}[i]$. Consider the infinite cyclic subgroups $G_1 = \langle 1 \rangle$ and $G_2 = \langle i \rangle$. Given an arbitrary Gaussian integer $a + bi$, clearly $a \in G_1$ and $bi \in G_2$, and so part (b) above is satisfied. And because a non-zero number cannot be both real and imaginary, part (c) is satisfied as well. Part (d) holds because the group is abelian. Thus, $\mathbb{Z}[i]$ is isomorphic *as a group* to $G_1 \times G_2$, and hence to $\mathbb{Z} \times \mathbb{Z}$. (Of course, these are by no means isomorphic *as rings*.)

We call this the *Internal Characterization Theorem* because it tells us whether a given group is isomorphic to a direct product, by looking only *inside* the group.

Chapter Summary

In this chapter we introduced the notion of *group isomorphism*, meeting numerous examples along the way. In addition, we characterized cyclic groups as essentially $\mathbb{Z}$ and $\mathbb{Z}_m$. We also used the notion of group isomorphism to characterize direct products of groups.

Warm-up Exercises

a. Give examples of the following group homomorphisms:

(a) A one-to-one homomorphism that is not an isomorphism.

(b) An onto homomorphism that is not an isomorphism.

(c) A homomorphism that is neither one-to-one nor onto.

b. Suppose that φ is a group homomorphism with domain $\mathbb{Z}_8$. What are the possible orders of $\varphi(1)$? What about $\varphi(4)$?

c. Suppose that g has infinite order and φ is a group homomorphism. Need $\varphi(g)$ have infinite order? What if φ is an isomorphism?

d. Explain why none of the following groups are isomorphic (even though all have eight elements):

$$\mathbb{Z}_8, \ \mathbb{Z}_4 \times \mathbb{Z}_2, \ D_4.$$

e. Explain why $\mathbb{Z}[i]$ and $\mathbb{Z}$ are not isomorphic as additive groups. That is, why isn't $\mathbb{Z}[i]$ cyclic?

f. Can a group be isomorphic to a proper subgroup of itself? What if the group is finite?

g. Let R be a finite ring, and consider its additive group and its group of units. Could these two groups be isomorphic?

Exercises

1. Consider the set of complex numbers $\{1, -1, i, -i\}$. Note this is a group under multiplication. Show that this group is isomorphic to $\mathbb{Z}_4$.

2. Use the Internal Characterization Theorem 28.3 to show that $\mathbb{Z}_{341}$ is isomorphic to $\mathbb{Z}_{11} \times \mathbb{Z}_{31}$. Then specify an explicit isomorphism.

3. Suppose that G, H, and K are groups. Prove that the direct products
$$(G \times H) \times K \quad \text{and} \quad G \times (H \times K)$$
are isomorphic. For this reason, we usually omit the parentheses when describing such groups.

4. Complete the proof the Internal Characterization Theorem 28.3; that is, show that the function ψ defined from $G_1 \times G_2$ to G is in fact an isomorphism.

5. If G is a group and g is some fixed element of G, show that the map φ_g defined by $\varphi_g(x) = gxg^{-1}$, for all $x \in G$, is an isomorphism from G onto itself.

6. Here you'll prove parts a and b of Theorem 28.1. Let $\varphi : G \to H$ be a group homomorphism between the groups G and H, where 1_G and 1_H are their respective identities.

   (a) Prove that $\varphi(1_G) = 1_H$.

   (b) For $g \in G$, prove that $(\varphi(g))^{-1} = \varphi(g^{-1})$.

7. Explain why $\mathbb{Q}$ and $\mathbb{Z}$ are not isomorphic as additive groups. That is, why isn't $\mathbb{Q}$ a cyclic group?

8. For group G, consider the map $\varphi : G \to G$ given by $\varphi(g) = g^{-1}$. Show φ is an isomorphism if and only if G is abelian.

9. Let $a, b, c \in \mathbb{R}$, not all zero.

   (a) Show that
   $$P = \{(x, y, z) \in \mathbb{R}^3 : ax + by + cz = 0\}$$
   is a subgroup of $\mathbb{R}^3$.

   (b) Show that
   $$L = \{(ak, bk, ck) \in \mathbb{R}^3 : \ k \in \mathbb{R}\}$$
   is a subgroup of $\mathbb{R}^3$.

   (c) Use the Internal Characterization Theorem to show that $\mathbb{R}^3$ is isomorphic to $P \times L$. What does this mean geometrically?

10. By way of analogy with Example 28.4, show that the additive group $\mathbb{Z}[\sqrt{2}]$ is isomorphic to a direct product of two non-trivial groups.

11. Use the Internal Characterization Theorem to show that $U(\mathbb{Z}_{15})$ is isomorphic to a direct product of two non-trivial groups.

12. Let n be a positive integer, and consider the group G_n described in Exercise 24.15. We will relabel these elements in such a way that we can consider G_n as the list of objects
$$G_n = \{I, R, R^2, \cdots, R^{n-1}, F, FR, FR^2, \cdots FR^{n-1}\},$$
where I is the label we assign to the identity, and the elements $(1, [k])$ are labeled as R^k, and the elements $(-1, [k])$ are labeled as FR^k.

    (a) Show that the elements of G_n under this labeling do satisfy the identities
    $$RF = FR^{n-1}, F^2 = 1, \text{ and } R^n = 1.$$
    In fact, it is possible to prove that our list above of the elements R^k, FR^k, together with these identities, abstractly characterizes the group G_n. We will not rigorously prove this here, however.

(b) Using the abstract characterization of G_n described above, prove that the dihedral group D_n as discussed in Chapter 22, is isomorphic to G_n. (You need only establish a one-to-one correspondence that preserves the identities.) Henceforth, when we refer to the dihedral groups D_n, we will often use the notation of this exercise.

13. In the group D_3, let $H_1 = \{I, F\}$ and $H_2 = \{I, R, R^2\}$. (Note that we are using the notation of the previous exercise.) Determine which of the four criteria of the Internal Characterization Theorem are satisfied by these two subgroups.

14. In this exercise we introduce a more efficient and compact notation for the elements of the group Q_8 of quaternions, discussed as Example 24.15. We will relabel these elements in such a way that we can consider Q_8 as the list of objects

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

where 1 is the label we assign to the identity matrix, i is the label assigned to matrix $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$, j is the label assigned to matrix $\begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}$, and k is the label assigned to matrix $\begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}$; furthermore, we interpret $-1, -i, -j, -k$ as the negatives of the corresponding matrices. Under this labeling, verify that the group elements satisfy the following rules: The element 1 is the identity, and multiplication by -1 (in either order) changes the sign of the element. Furthermore, $i^2 = j^2 = k^2 = -1$, $ij = k, jk = i$, and $ki = j$, while $ji = -k, kj = -i$, and $ik = -j$. We will henceforth use the notation of this exercise when we have occasion to use the group of quaternions.

Chapter 29

Permutations and Cayley's Theorem

In Chapters 22 and 23 we used the idea that a symmetry of a geometric object like a triangle or tetrahedron must take vertices to vertices. For example, specifying where the vertices of a tetrahedron are sent completely determines the symmetry function. We used this reasoning to determine a complete list of symmetries of the tetrahedron. We called a specification of how the vertices are moved a *permutation* of the vertices. We will now consider the notion of permutations in an abstract setting. This leads to an important theorem from group theory, which says that *all* finite groups can be thought of as groups of permutations.

29.1 Permutations

Consider the list $1, 2, 3, 4, \cdots, n$ of the first n positive integers. We wish to rearrange or *permute* this list. To do this, we must tell ourselves which slot the integer 1 should be placed in, which slot the integer 2 should be placed in, and so forth. What this means is that a permutation of this list amounts to a function

$$\alpha : \{1, 2, 3, \cdots, n\} \to \{1, 2, 3, \cdots, n\}$$

that is one-to-one and onto. It is one-to-one, because no two integers can be placed in the same slot. It is onto, because each slot must be filled. Formally then, a **permutation of the set** $\{1, 2, 3, \cdots, n\}$ is a one-to-one and onto function from this set to itself.

$\triangleright$ **Quick Exercise.** It is actually true that if a function from a *finite* set to itself is one-to-one, it is automatically onto. Can you explain why this is true? Give an example to show that this is false in the infinite case. $\triangleleft$

As we did in Chapters 22 and 23, we will denote such functions by a $2 \times n$ array: The top row tells the names of the elements of the set $\{1, 2, \cdots, n\}$, and the bottom row records the slots to which they are sent. Thus,

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

is a permutation of the set $\{1, 2, 3, 4, 5\}$, which does not move 4 or 2, sends 1 to 3, 3 to 5, and 5 to 1. We could use functional notation to express the above facts. For instance, sending 1 to 3 could be written as

$$\begin{pmatrix} 1\,2\,3\,4\,5 \\ 3\,2\,5\,4\,1 \end{pmatrix}(1) = 3.$$

We denote the set of all permutations on the set $\{1, 2, \cdots n\}$ by S_n. This set has $n!$ elements because for 1 there are n slots available, for 2 there are $n - 1$ slots available (because 1 has taken one of them), for 3 there are $n - 2$ slots available, and so on. This gives us

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

possibilities altogether.

29.2 The Symmetric Groups

We claim that S_n is a group, under the operation *functional composition.* We denote that operation by $\circ$, or by juxtaposition. Note first that the composition of two one-to-one, onto functions remains one-to-one and onto.

▷ **Quick Exercise.** Prove this. ◁

This means that the operation is well defined. As we've observed before, functional composition is an associative operation. It thus remains to show that S_n has an identity element, and that every element has an inverse.

But consider the **identity function** ι, where $\iota(k) = k$, for $k = 1, 2, \cdots, n$. We would write this as

$$\iota = \begin{pmatrix} 1\,2\,3 \cdots n \\ 1\,2\,3 \cdots n \end{pmatrix},$$

using our matrix notation. This clearly serves as the identity element; if α is any element of S_n and $1 \le k \le n$, then

$$\iota \circ \alpha(k) = \iota(\alpha(k)) = \alpha(k),$$

and so $\iota \circ \alpha = \alpha$.

▷ **Quick Exercise.** Write down the corresponding proof for $\alpha \circ \iota$. ◁

What about inverses? If $\alpha \in S_n$, because it is a one-to-one onto function, it possesses an **inverse function** α^{-1}:

$$\alpha^{-1}(n) = m \text{ is true exactly if } \alpha(m) = n.$$

This is exactly the function that composes with α to give the identity function ι. Informally, the existence of α^{-1} means that we can 'un-do' any rearrangement of our list.

We call the group S_n the **symmetric group** on n. We will also refer to such a group (or one of its subgroups) as a **group of permutations**. Notice that although for convenience's sake we think of S_n as permutations of the set $\{1, 2, \cdots, n\}$, it really doesn't matter what finite set of objects we're permuting (as long as they are distinguishable).

Let's look at some particular computations in S_n:

Example 29.1

What is the inverse of the permutation $\begin{pmatrix} 1\,2\,3\,4\,5 \\ 3\,2\,5\,4\,1 \end{pmatrix}$ in S_5? It is $\begin{pmatrix} 1\,2\,3\,4\,5 \\ 5\,2\,1\,4\,3 \end{pmatrix}$.

▷ **Quick Exercise.** Verify the result in Example 29.1. ◁

Example 29.2

Consider the permutations

$$\alpha = \begin{pmatrix} 1\,2\,3\,4 \\ 3\,2\,1\,4 \end{pmatrix} \quad \text{and} \quad \beta = \begin{pmatrix} 1\,2\,3\,4 \\ 4\,3\,2\,1 \end{pmatrix}.$$

What is the permutation $\alpha \circ \beta$? Because we read functional composition from right to left, we have

$$\alpha \circ \beta(1) = \alpha(\beta(1)) = \alpha(4) = 4.$$

By doing the three other necessary computations, we obtain that

$$\alpha \circ \beta = \begin{pmatrix} 1\,2\,3\,4 \\ 4\,1\,2\,3 \end{pmatrix}.$$

▷ **Quick Exercise.** Verify the calculation for $\alpha \circ \beta$ given in Example 29.2. Then compute $\beta \circ \alpha, \alpha^2$, and β^{-1}. ◁

One of your results from the previous Quick Exercise drives home a point we had already encountered in Chapters 22 and 23. The symmetric groups are highly non-abelian.

▷ **Quick Exercise.** Choose two elements at random from S_6 and compute their products in both orders. It is highly likely that your answers will not be the same. ◁

Our work in Chapters 22 and 23 (together with our definition of isomorphism in the last chapter) tells us that D_3, the group of symmetries of the equilateral triangle, is isomorphic to the abstract symmetric group S_3. This is the mathematically precise meaning of the correspondence we set up in Chapter 22 between symmetries, and permutations of vertices. Similarly, the group of symmetries of the cube is isomorphic to S_4 because we viewed symmetries of the cube as equivalent to permuting the diagonals. Furthermore, the group of symmetries of the cube is isomorphic to a *subgroup* of S_8. (We got only a subgroup because not all permutations of the eight vertices of the cube can be accomplished by a rigid motion.)

A nice way to view this situation is this: We have *represented* these symmetry groups as abstract groups of permutations. While the symmetries, thought of geometrically, are quite concrete, they are rather difficult to compute with (especially if we don't have a cube or tetrahedron around to handle!). The permutation groups, while more abstract, provide us with a really easy way to compute explicitly. An important theme in group theory has been to represent groups (by various means), in such a way that computation is easier. One way is by means of permutations. Another is by using matrix multiplication. And in fact, we can now view what we did in Chapter 22 as representing D_3 in two different ways: permutations (of the vertices), and matrix multiplication (where the matrices accomplished flips and rotations). We also represented the symmetry groups of the tetrahedron and cube with permutations, while leaving to the ambitious exercise solver the task of using 3×3 matrices for representing these groups. (See Exercise 23.7.)

29.3 Cayley's Theorem

It is a remarkable fact that *all* finite groups can be realized concretely as groups of permutations and as groups of matrices. In the remainder of this chapter we will prove the first of these assertions. We avoid the second because it requires the use of matrices of large size.

We begin with a simple example, the cyclic group $\{1, a, a^2\}$ of order three. (Of course, this is isomorphic to the group $\mathbb{Z}_3$, but we will find it more convenient in this discussion to use multiplicative notation.) We wish to show that this group is a group of permutations of some set, that is, we claim it is a subgroup of S_n for some n. In fact, we will represent the elements of this group as permutations of the set $\{1, a, a^2\}$ itself! This can be rather confusing at times because we will be thinking of the elements of $\{1, a, a^2\}$ as elements of the cyclic group, on the one hand, and as elements of the set to be permuted, on the other. Consider the function

$$\varphi_a : \{1, a, a^2\} \to \{1, a, a^2\}$$

defined by $\varphi_a(x) = ax$. That is, φ_a merely multiplies on the left by a. This means that

$$\varphi_a(1) = a, \quad \varphi_a(a) = a^2, \quad \varphi_a(a^2) = a^3 = 1.$$

We could thus denote this function by

$$\begin{pmatrix} 1 & a & a^2 \\ a & a^2 & 1 \end{pmatrix}.$$

In a similar fashion we can define φ_1 and φ_{a^2}, which can be represented as

$$\begin{pmatrix} 1 & a & a^2 \\ 1 & a & a^2 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & a & a^2 \\ a^2 & 1 & a \end{pmatrix}.$$

Notice that the three functions $\varphi_1, \varphi_a, \varphi_{a^2}$ are all one-to-one and onto functions defined on the three-element set $\{1, a, a^2\}$. Of course, if we re-label this set as $\{1, 2, 3\}$, we can then identify these three functions as permutations of $\{1, 2; 3\}$, that is, as elements of S_3. Explicitly, if we re-label 1 as 1, a as 2, and a^2 as 3, then

$$\varphi_1 = \begin{pmatrix} 1\,2\,3 \\ 1\,2\,3 \end{pmatrix}, \quad \varphi_a = \begin{pmatrix} 1\,2\,3 \\ 2\,3\,1 \end{pmatrix}, \text{ and } \varphi_{a^2} = \begin{pmatrix} 1\,2\,3 \\ 3\,1\,2 \end{pmatrix}.$$

We have demonstrated a means of associating with each element of the group $\{1, a, a^2\}$ an element of S_3. More generally, we wish to associate each element of a group that has n elements with an element of S_n, in such a way that the group operation is preserved.

$\triangleright$ **Quick Exercise.** Multiply two elements of the group

$$\{1, a, a^2\}.$$

Does the result correspond to the composition of the corresponding permutations of the three-element set $\{1, a, a^2\}$? $\triangleleft$

With this example behind us, we are now ready to prove the general theorem:

Theorem 29.1 Cayley's Theorem *Suppose that G is a finite group with n elements. Then G is isomorphic to a subgroup of S_n.*

The idea of the proof that follows is actually very simple, but it is easy to get lost in the forest of details! This proof is merely a more general version of the argument just given for the cyclic group $\{1, a, a^2\}$.

Proof: Our goal is to assign to each element of the group G a permutation belonging to S_n. What we will actually do is assign to each group element a permutation of the set G itself. We will call the group of such permutations S_G. But clearly S_n and S_G are isomorphic groups, and so this will be enough.

Because G has finitely many elements, we label its elements as

$$G = \{g_1, g_2, \cdots, g_n\}.$$

In other words, we have listed the elements of G in some fixed order.

For each integer i, we define a function

$$\varphi_i : G \to G$$

by letting $\varphi_i(g_k) = g_i g_k$. That is, φ_i merely multiplies each element of G on the left by g_i, the ith element of G.

We will show that φ_i is a permutation of the elements of G. That is, we will show that φ_i is a one-to-one and onto function.

φ_i is one-to-one: Suppose that $\varphi_i(g_k) = \varphi_i(g_j)$. This means that $g_i g_k = g_i g_j$. But if we multiply this equation on the left by g_i^{-1}, we see that $g_k = g_j$. Thus, φ_i is one-to-one.

φ_i is onto: Actually, because the set in question (namely, G) is finite, and the function is one-to-one, we could conclude immediately that the function is onto. However, for clarity we shall verify directly that φ_i is onto. For that purpose, choose an arbitrary integer j, where $1 \leq j \leq n$. The equation $g_i x = g_j$ has a unique solution in G. Thus, there must be a k so that $g_i g_k = g_j$. But this evidently means that $\varphi_i(g_k) = g_j$, and so φ_i is onto.

Let's denote by Φ the assignment $g_i \longmapsto \varphi_i$. That is, $\Phi(g_i) = \varphi_i$. Because φ_i is one-to-one and onto (and hence, a permutation of G), we have that Φ is a function from G to S_G. We claim now that this function translates the multiplication in G into the functional composition in S_G; that is, we claim that Φ is a homomorphism.

Φ is a homomorphism: We must show that

$$\Phi(g_i g_k) = \Phi(g_i) \circ \Phi(g_k).$$

Note that both $\Phi(g_i g_k)$ and $\Phi(g_i) \circ \Phi(g_k)$ are functions defined on the set G. To show that two functions are equal, we should check that they have the same value at a generic element of their domain. So pick $g_m \in G$ and compute:

$$\Phi(g_i g_k)(g_m) = (g_i g_k) g_m = g_i(g_k g_m) = g_i(\varphi_k(g_m)) =$$

$$\varphi_i(\varphi_k(g_m)) = (\Phi(g_i) \circ \Phi(g_k))(g_m),$$

which is what we required.

Finally, we require that *Φ is one-to-one*: So suppose that $\Phi(g_i) = \Phi(g_k)$. Because $\Phi(g_i)$ and $\Phi(g_k)$ are functions, they are equal if they do the same thing to every element of their domain G. In particular, then, they must give the same element when applied to $1 \in G$. But then $g_i \cdot 1 = g_k \cdot 1$, or $g_i = g_k$. Thus, Φ is one-to-one.

We have thus proved that G is isomorphic to the subgroup $\Phi(G)$ of the group of permutations S_G. Because S_G is isomorphic to S_n, this completes the proof. □

Cayley's Theorem is usually not very practical for gaining insight into a particular group. For example, if we were to apply it to a group with 8 elements, we would represent it as a subgroup of S_8, which has $8! = 40,320$ elements! Furthermore, if the original group is abelian, we have then represented it as a subgroup of a highly non-abelian group. Nonetheless, the theoretical importance of Cayley's Theorem should not be minimized.

We should perhaps remark that exactly the same proof as we gave above for Cayley's Theorem can be used for an infinite group. This shows that any group whatsoever can be represented as a subgroup of a group of permutations, only this time as permutations of an infinite set! In general, we will avoid groups of permutations of infinite sets in this book.

Historical Remarks

Thus, in principle, to study finite groups we need only study groups of permutations. In the 19th century most group theorists did exactly that. Although the eminent British mathematician Arthur Cayley had enunciated an abstract definition of group in 1853, it was a long time before mathematicians felt comfortable working in an abstract and axiomatic context. The study of finite groups had grown out of work by Lagrange (in the late 18th century) and Galois (in the early 19th century) in studying the roots of polynomial equations. Their insight into such roots was enlarged by thinking about *permuting* them; we will pick up this theme in Chapter 47. Afterward, the French mathematician Cauchy studied permutation groups in their own right, and he introduced our $2 \times n$ matrix notation for permutations. Consequently, even when group theory had grown beyond the particular problems of Lagrange and Galois, the thought that it was still about permutations remained.

Chapter Summary

In this chapter we discussed the *symmetric group* S_n, which consists of all *permutations* of a set with n elements. We proved *Cayley's Theorem*, which asserts that all finite groups are isomorphic to a group of permutations.

Warm-up Exercises

a. Suppose that α is the element of S_7 specified by

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

What is $\alpha(5)$? What is $\alpha(4)$? What is the inverse of α?

b. Suppose in addition that β is the element of S_7 specified by

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

Calculate

$$\beta^2, \quad \beta\alpha, \quad \alpha\beta, \quad \beta\alpha^{-1}, \quad \alpha\beta\alpha.$$

c. Using the elements α and β from the previous exercises, calculate $(\alpha\beta)^{-1}$ twice, once by using your computation of $\alpha\beta$ in Exercise b, and once by computing α^{-1} and β^{-1}, and then composing them (in the proper order!).

d. How many elements are there in S_4? S_5? S_6?

e. When someone says that every finite group is a group of permutations, what does that mean?

f. Explain how the group of symmetries of the cube can be described as isomorphic to a subgroup of S_{12}, S_8, S_6, and S_4, depending on which set of objects associated with the cube you consider permuted.

g. Every finite abelian group is isomorphic to a subgroup of a nonabelian group. Why?

Exercises

1. What are the orders of the permutations $\alpha, \beta, \beta^2, \beta\alpha, \alpha\beta, \beta\alpha^{-1}$, and $\alpha\beta\alpha$ given above in Exercises a and b?

2. Suppose that m and n are positive integers, and $m < n$. Define $I : S_m \to S_n$ as follows: Given $\alpha \in S_m$, we let $I(\alpha)(k) = \alpha(k)$, if $k \leq m$, and $I(\alpha)(k) = k$, if $n \geq k > m$. Show that I is a one-to-one homomorphism, which is not onto. *Note*: You first must check that $I(\alpha) \in S_n$. We can paraphrase the contents of this exercise by asserting that (up to isomorphism) S_m is a subgroup of S_n.

3. Following the proof of Cayley's Theorem 29.1, determine explicitly which permutations of S_4 each of the elements of the group $\{1, -1, i, -i\}$ correspond to.

4. Repeat Exercise 3 for the group of quaternions.

5. Repeat Exercise 3 for S_3. Note that this gives us two representations of S_3 as a group of permutations: the original definition as a permutation of three elements, and the new one as a permutation of six elements!

6. Let n be a positive integer and let k be a fixed integer, $1 \leq k \leq n$. Let
$$G_k = \{\alpha \in S_n : \alpha(k) = k\}.$$
Prove that G_k is a subgroup of S_n. It is called a **stabilizer subgroup** of S_n. How many elements belong to G_k?

7. Generalize Exercise 6: Let K be any subset of $\{1, 2, \cdots, n\}$. Let
$$G_K = \{\alpha \in S_n : \alpha(k) \in K, \text{for all } k \in K\}.$$
Prove that G_K is a subgroup of S_n. How many elements belong to G_K?

Chapter 30

More About Permutations

In Chapter 29 we introduced the symmetric groups S_n. These groups are of theoretical importance because every finite group is isomorphic to a subgroup of such a group, as Cayley's Theorem 29.1 shows. They also provide us with many examples of non-abelian groups. In this chapter we inquire a bit more into permutations and introduce an efficient and illuminating notation for them.

We first make a few notational remarks about permutation groups.

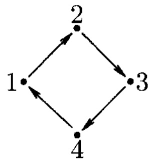
Note that if $m < n$, we can think of S_m as a subgroup of S_n because any permutation of $\{1, 2, \cdots, m\}$ can be thought of as a permutation of the larger set $\{1, 2, \cdots, n\}$, which leaves the elements $m + 1, m + 2, \cdots, n$ fixed. Technically speaking, we should use the language of isomorphism to describe this situation, but we will be content to be a little sloppy here and identify S_m as a subgroup of S_n. (See Exercise 29.2.) Consequently, in what follows we will not bother to be too specific about which permutation group a given permutation belongs to.

Recall that the group operation in S_n is functional composition, since the elements of S_n are actually functions from the set $\{1, 2, \ldots, n\}$ to itself. However, we will for the most part in this (and future chapters) speak less formally as group theorists about this operation; since it is written multiplicatively, we will tend to speak of the *product* of two permutations.

30.1 Cycles

Consider first the permutation
$$\alpha = \begin{pmatrix} 1\,2\,3\,4 \\ 2\,3\,4\,1 \end{pmatrix}.$$

It permutes the elements 1, 2, 3, and 4 *cyclically*, as the following picture suggests:



In general a permutation $\alpha \in S_n$ is a **cycle of length** k, if there exist integers $a_1, a_2, \cdots, a_k$ so that

$$\alpha(a_1) = a_2,\ \alpha(a_2) = a_3,\ \cdots,\ \alpha(a_{k-1}) = a_k,\ \alpha(a_k) = a_1,$$

and α leaves fixed the remaining $n - k$ elements in its domain. We will denote this cycle by the notation

$$(a_1 a_2 a_3 \cdots a_k).$$

Example 30.1

Consider the cycle (125). It is short-hand for the permutation

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

Notice that we could just as well have represented this cycle by (251) or (512). It is clear that (125) represents an element in S_5, at least. But in a particular situation, we might use the same notation to denote the corresponding element

$$\begin{pmatrix} 1\,2\,3\,4\,5\,6\,7 \\ 2\,5\,3\,4\,1\,6\,7 \end{pmatrix}$$

in S_7. Notice that the inverse permutation cycles elements in the opposite direction, which in this case is (152).

▷ **Quick Exercise.** Express the cycle (3251) in our earlier notation. Write this cycle three other ways in cycle notation. What is its inverse? ◁

Of course, not all permutations are cycles: Consider the permutation

$$\beta = \begin{pmatrix} 1\,2\,3\,4\,5 \\ 2\,5\,4\,3\,1 \end{pmatrix}.$$

While it behaves as a cycle on the set $\{1, 2, 5\}$, it also behaves cyclically on the set $\{3, 4\}$. What this really means is that we can factor β as a composition of two cycles:

$$\beta = (125)(34) = (34)(125).$$

▷ **Quick Exercise.** Convince yourself that β really equals the composition of the two-cycle permutations (125) and (34), composed in either order. What is the inverse of β? ◁

30.2 Cycle Factorization of Permutations

This suggests that perhaps we could factor *any* permutation into a product of cycles. To describe this, we need some terminology. The **support** of a permutation α is the set of all integers k so that $\alpha(k) \neq k$. Speaking more colloquially, the support of a permutation is the set of elements in its domain *that it moves*. If $\alpha(k) = k$, then k is not in the support of α; we say that α **fixes** k.

Example 30.2

The permutation (125)(34) has support $\{1, 2, 3, 4, 5\}$. The permutation (3251) has support $\{1, 2, 3, 5\}$; it leaves 4 fixed, and so 4 does not belong to the support of this permutation.

Let's make a technical observation about the support of a permutation, which we will repeatedly find of use in the arguments which follow.

Lemma 30.1 *Suppose that α is a permutation, and k is in the support of α. Then $\alpha(k)$ is in the support of α too.*

Proof: Because k is in the support of α, $\alpha(k) \neq k$. Now apply the function α to these two distinct integers. Because α is a one-to-one function, we must get distinct integers:

$$\alpha(\alpha(k)) \neq \alpha(k).$$

But this means exactly that $\alpha(k)$ is in the support of α. □

Two permutations are **disjoint** if their supports contain no element in common. In the previous Quick Exercise it was the disjointness of (125) and (34) that allowed them to commute. Let's prove this in general:

Theorem 30.2 *Suppose that α and β are disjoint permutations. Then*

$$\alpha\beta = \beta\alpha.$$

Proof: Suppose that α and β are disjoint. We must show that the two functions $\alpha\beta$ and $\beta\alpha$ are the same: So we show they do the same thing to typical elements of their domain.

If we choose an integer k belonging to neither the support of α nor the support of β, then clearly

$$\alpha\beta(k) = \alpha(k) = k = \beta(k) = \beta\alpha(k).$$

Any other integer belongs to the support of exactly one of α and β, because they are disjoint. Without loss of generality, let's suppose that integer k belongs to the support of α (but not that of β). Now by the previous lemma, $\alpha(k)$ also belongs to the support of α, and hence not to that of β. Thus, we have

$$\alpha\beta(k) = \alpha(k)$$

and

$$\beta\alpha(k) = \alpha(k).$$

Hence, the two functions $\alpha\beta$ and $\beta\alpha$ are the same, as claimed. □

We're now ready to state and prove the factorization theorem for permutations. Note that this theorem is very similar in flavor to the Fundamental Theorem of Arithmetic 2.8 and 2.9: It asserts that every permutation can be factored uniquely into simpler pieces (cycles), just as the Fundamental Theorem of Arithmetic asserts that integers can be factored uniquely into simpler pieces (primes).

Theorem 30.3 Cycle Factorization Theorem for Permutations *Every non-identity permutation is either a cycle or can be uniquely factored (up to order) as a product of pairwise disjoint cycles.*

Proof: We first prove the *existence* of the required factorization by using induction on the size of the support of the permutation. Note first that if the size of the support is zero, the permutation moves no element, and is consequently the identity permutation. No permutation can have exactly one element in its support.

▷ **Quick Exercise.** Why can no permutation have exactly one element in its support? ◁

Thus, the base case for the induction is support size two. Suppose that the support of permutation α has two elements. If k is an element of this support, by Lemma 30.1 $\alpha(k)$ must be the other element in the support. But $\alpha(k) \neq \alpha^2(k)$, and so the only choice is that $\alpha^2(k) = k$. That is, α must be a cycle of length two.

Now suppose that α is a permutation with n elements in its support. Our induction hypothesis says that all non-identity permutations with support size less than n are either cycles, or products of disjoint cycles. Pick a_1, an element of the support of α. Then

$$a_1,\ a_2 = \alpha(a_1),\ a_3 = \alpha(a_2),\ \cdots$$

are all elements of the support. Because the support is finite, sooner or later we will have a duplication in this list. Suppose the first duplication in the list occurs at $a_{k+1} = \alpha(a_k)$. Then because α is one-to-one, the only duplication possible is if $\alpha(a_k) = a_1$ (any other duplication would give different elements with the same α value). Now consider the cycle $\beta = (a_1a_2a_3\cdots a_k)$. If $k = n$, then $\alpha = \beta$, which is a cycle, as required. Otherwise, consider the element $\beta^{-1}\alpha$. This function equals α on any integer other than $a_1, a_2, \cdots, a_k$, but does not move the a_i's. Thus, the support of $\beta^{-1}\alpha$ is disjoint from $\{a_1, a_2, \cdots, a_k\}$ and has $n-k$ elements, which is fewer than n. Consequently, by the induction hypothesis, we can factor $\beta^{-1}\alpha$ as a product of disjoint cycles. Then α is the product of these cycles, times β. This completes the proof, by the Principle of Mathematical Induction.

It remains to prove that the factorization we have obtained is unique. This is a similar induction proof, left as Exercise 30.2. □

Henceforth, when we compute with permutations, we will invariably use the disjoint cycle representation guaranteed by this theorem.

Example 30.3

Suppose that

$$\alpha = (154)(23)(689) \quad \text{and} \quad \beta = (14895)(27)$$

are two permutations, written as products of disjoint cycles. Let's compute the product $\alpha\beta$. This is an entirely straightforward matter, if we remember that the operation here is functional composition, and so we operate from right to left. Start with any element of the domain (say, 1):

$$\alpha\beta(1) = \alpha(4) = 1.$$

Thus, the product fixes 1. Let's try 2:

$$\alpha\beta(2) = \alpha(7) = 7;$$
$$\alpha\beta(7) = \alpha(2) = 3;$$
$$\alpha\beta(3) = \alpha(3) = 2.$$

Thus, (273) is one of the cycle factors of the product $\alpha\beta$. We leave it to you to complete the computation to obtain:

$$\alpha\beta = (273)(49)(68).$$

▷ **Quick Exercise.** Complete this computation. Then compute $\beta\alpha$ and $\beta^2\alpha^{-1}$ in the same way. ◁

Once a permutation has been factored as a product of disjoint cycles, we call the set of elements moved by each of its constituent cycle factors its **orbits**. Thus, the permutation

$$(157)(2389)$$

has orbits $\{1,5,7\}$ and $\{2,3,8,9\}$. In addition, it has the **trivial orbits** $\{4\}$ and $\{6\}$. Thus, every element of the domain of the permutation belongs to exactly one orbit, and the other elements of the orbit it belongs to are exactly the locations we can move the element to, by repeatedly applying the permutation.

▷ **Quick Exercise.** Give examples in S_7 of an element with one orbit of seven elements and also an element with 3 orbits, one orbit of 4 elements, one orbit of 2 elements, and a trivial orbit. ◁

30.3 Orders of Permutations

As an application of the Factorization Theorem, let's compute the orders of cycles, and then of arbitrary permutations. For example, the

cycle (1234) has order 4:

$$(1234), \quad (1234)^2 = (13)(24),$$

$$(1234)^3 = (1234)(13)(24) = (1432), \quad (1234)^4 = \iota.$$

In general this is true. A cycle of length k has order k. For if α is such an element, obviously we must apply it to itself exactly k times to bring any element of its support back around the cycle to itself.

Now suppose that α is an arbitrary permutation. By our Factorization Theorem 30.3, we can factor it as a product of k disjoint cycles

$$\alpha = \beta_1\beta_2\cdots\beta_k.$$

Because these cycles are disjoint (and so commute with one another), it is easy to compute the compositions of α with itself:

$$\alpha^m = (\beta_1)^m(\beta_2)^m\cdots(\beta_k)^m.$$

To make such a product the identity, we must guarantee that simultaneously all the terms $(\beta_i)^m$ are the identity (for otherwise, the element α^m would not be the identity on elements in the orbit corresponding to β_i). To accomplish this, we need m to be a common multiple of the orders of all the β_i's. The least common multiple will do the job. In fact, the least common multiple of the orders of all the β_i's is the order of α. (See Exercise 30.7.)

Example 30.4

What is the order of the element (1234)(567)(89)? The least common multiple of 4, 3, and 2 is 12, so this must be the order of this permutation.

Chapter Summary

In this chapter we proved that every permutation is a cycle or can be factored uniquely as a product of disjoint cycles. This makes it much easier notationally and computationally to deal with permutations.

Warm-up Exercises

a. Compute the following products (you should end up with a disjoint cycle representation):

$$(134)(2457), \quad (12)(13)(14), \quad (1567)^3(213)^{-1}.$$

b. Factor the following permutation as a product of disjoint cycles:

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

What is the order of this permutation?

c. Give a permutation with three non-trivial orbits, one with 3 elements and two with 4 elements each. What is the order of the permutation you have constructed?

d. Two non-identity permutations with disjoint supports necessarily commute, but the converse of this statement is false; give an example. *Hint*: *All* groups have non-trivial abelian subgroups!

e. Does the set of 2-cycles in S_4 (together with the identity) form a subgroup? What about 3-cycles, or 4-cycles?

f. In what ways are the Factorization Theorem for Permutations 30.3 and the Fundamental Theorem of Arithmetic 2.8 and 2.9 similar? In what ways are they different? (For the latter, consider whether cycles are *irreducible.*)

Exercises

1. Compute explicitly the cyclic subgroups of S_7 generated by the following permutations:

$$(357), \quad (14)(256), \quad (123)(456).$$

2. Prove the uniqueness part of Factorization Theorem 30.3: The factorization of a permutation into disjoint cycles is unique, up to order.

3. If $\alpha \in S_n$ is a cycle, is α^2 a cycle also? Give an example when this is true and an example when this is false. Now characterize cycles where this is true and cycles where this is false.

4. Give the order of each of the following permutations: $(123)(456)$, $(123)(4567)$, $(123)(45)$.

5. Determine all possible orders of a product of two 3-cycles.

6. Repeat Exercise 5 for the product of two 4-cycles.

7. Suppose the permutation $\alpha = \beta_1\beta_2\cdots\beta_k$, the product of disjoint cycles where the order of cycle β_i is m_i. Show that the order of α is the least common multiple of the m_i.

8. (a) There are four distinct disjoint cycle structures for non-identity elements of S_4. Name them.

   (b) The group of symmetries of the cube is isomorphic to S_4, since it can be viewed as the group of permutations of the diagonals of the cube. Describe what the four disjoint cycle structures of part a mean geometrically.

9. Let n be a positive integer, and k, j fixed integers with $1 \le k, j \le n$. Let G_k and G_j be the stabilizer subgroups of S_n considered in Exercise 29.6. Given $\alpha \in G_k$, define

$$\Phi : G_k \to G_j$$

by setting $\Phi(\alpha) = (kj)\alpha(kj)$ (where here (kj) is a 2-cycle). Prove that Φ is a group isomorphism. (Note that an important verification is to check that $\Phi(\alpha) \in G_j$.)

10. We generalize Exercise 9. Let K and J be subsets of $N = \{1, 2, 3, \cdots, n\}$, with the same number of elements. Let G_K and G_J be the stabilizer subgroups of S_n, discussed in Exercise 29.7. Since K and J have the same number of elements, there exists a one-to-one and onto function $\beta : K \to J$, and similarly there exists a one-to-one and onto function $\gamma : N\backslash K \to N\backslash J$.

    (a) Define a function μ by setting $\mu(m) = \beta(m)$ if $m \in K$, and $\mu(m) = \gamma(m)$ if $m \in N\backslash K$. Argue that $\mu \in S_n$.

    (b) Given $\alpha \in G_K$, define $\Phi : G_K \to G_J$ by setting $\Phi(\alpha) = \mu\alpha\mu^{-1}$. Prove that this is a group isomorphism.

Chapter 31

Cosets and Lagrange's Theorem

In Chapter 27 we introduced the idea of *group homomorphism.* Let's recall the corresponding development that we followed, after introducing the idea of *ring homomorphism.* We obtained the Fundamental Isomorphism Theorem for Rings 19.1, which asserts that knowing about homomorphisms is equivalent to knowing about ideals: Each homomorphism gives rise to an ideal (its *kernel*) and each ideal in turn gives rise to a homomorphism (of which it is the kernel) to a *ring of cosets.* We would like to emulate this powerful and useful theory in the theory of groups, so that we can better understand group homomorphisms. This will be the goal of the next three chapters.

31.1 Cosets

We begin this development by considering the notion of *coset* in the group context. In rings, we started with an ideal and formed its cosets. In groups, we begin with a subgroup and form its cosets.

Let G be a group (written multiplicatively), and suppose that H is a subgroup. For each $g \in G$, we form the set

$$Hg = \{hg : h \in H\}.$$

We call such sets **right cosets** of H in G. We use the term *right* coset, because we are multiplying by g on the right. In a non-abelian group, it might make a difference whether we consider right cosets or left cosets. (We'll come back to this topic in the next chapter.) Note that if G is an additive group, we would write such a coset as

$$H + g = \{h + g : h \in H\}.$$

This is exactly the definition of coset we used, in case G were a ring, and H an ideal.

Let's look at some examples, before we go any further:

Example 31.1

Consider the additive group $\mathbb{Z}$, and its cyclic subgroup $\langle 4\rangle$. Then the sets

$$\langle 4\rangle + 0, \langle 4\rangle + 1, \langle 4\rangle + 2, \langle 4\rangle + 3$$

are the distinct (right) cosets of $\langle 4\rangle$ in $\mathbb{Z}$, just as in the ring context.

Of course, we can generate a long list of examples by referring back to ring theory, by our customary means: Forget the multiplication, and look at the additive group. But let's consider a purely group-theoretic example.

Example 31.2

Consider the subgroup $H = \{\iota, (12)\}$ of the symmetric group S_3. We then obtain three distinct right cosets, as follows:

$$H\iota = H(12) = \{\iota, (12)\} = H,$$

$$H(123) = H(23) = \{(123), (23)\},$$

and

$$H(132) = H(13) = \{(132), (13)\}.$$

The subgroup H is obviously a coset of itself; we obtained this by choosing the two elements in H itself. The other two cosets also have two elements each.

▷ **Quick Exercise.** Verify that the cosets given above do indeed consist of the elements listed. ◁

Example 31.3

Let's compute the right cosets of the subgroup

$$K = \{\iota, (123), (132)\}$$

in the same group S_3. This time we obtain

$$K\iota = K(123) = K(132) = \{\iota, (123), (132)\}$$

and

$$K(12) = K(13) = K(23) = \{(12), (13), (23)\}.$$

Once again K is a coset of itself and there is one other coset, which also has three elements.

▷ **Quick Exercise.** Verify that the cosets listed are correct. ◁

Example 31.4

Consider the group

$$U(\mathbb{Z}_{21}) = \{1, 2, 4, 5, 8, 10, 11, 13, 16, 17, 19, 20\}.$$

Let's compute the right cosets of the subgroup $H = \{1, 4, 16\}$: We obtain

$$H1 = H4 = H16 = \{1, 4, 16\},$$

$$H2 = H8 = H11 = \{2, 8, 11\},$$

$$H10 = H13 = H19 = \{10, 13, 19\},$$

and

$$H5 = H20 = H17 = \{5, 20, 17\}.$$

▷ **Quick Exercise.** Verify that the cosets listed are correct. ◁

31.2 Lagrange's Theorem

The examples from finite group theory that we have looked at are very suggestive. In each case, all the cosets of a given subgroup are the same size and this allows us to decompose the group into finitely many pairwise disjoint pieces of the same size. This observation, when made precise, is a very valuable tool for counting in finite groups; it is called *Lagrange's Theorem.* Before obtaining Lagrange's Theorem, we need to make precise the observations we've made about cosets. We here prove the Coset Theorem for groups, which is directly analogous to the Coset Theorem 18.1 for rings.

Theorem 31.1 The Coset Theorem *Let H be a subgroup of a group G, and $a, b \in G$. Then*

a. If $Ha \subseteq Hb$, then $Ha = Hb$.

b. If $Ha \cap Hb \neq \emptyset$, then $Ha = Hb$.

c. $Ha = Hb$ if and only if $ab^{-1} \in H$.

d. There exists a one-to-one and onto function between any two right cosets Ha and Hb. Thus, if H has finitely many elements, every right coset has that same number of elements.

Proof: (a): Suppose that H is a subgroup of the group G, and a and b are elements of the group for which $Ha \subseteq Hb$. Then

$$a = 1a \in Ha \subseteq Hb,$$

and so there exists $h \in H$ such that $a = hb$. But then $b = h^{-1}a \in Ha$. Now, if $k \in H$, $kb = kh^{-1}a \in Ha$, and so $Hb \subseteq Ha$. That is, $Ha = Hb$.

(b): Suppose that $Ha \cap Hb \neq \emptyset$. Choose c in this intersection. Then $c \in Ha$, and so $Hc \subseteq Ha$. But then by part (a), $Hc = Ha$. But similarly, $Hc = Hb$, and so $Ha = Hb$.

(c): If $Ha = Hb$, then $a = 1a \in Ha = Hb$, and so there exists $h \in H$ such that $a = hb$. But then $ab^{-1} = h \in H$, as required. Conversely, if $ab^{-1} \in H$, then $a = ab^{-1}b \in Hb$. But then $a \in Ha \cap Hb$, and so by part (b) $Ha = Hb$.

(d): Define the function $\varphi : Ha \to Hb$ by $\varphi(x) = xa^{-1}b$. First, note that if $x \in Ha$, then $x = ha$, for $h \in H$. But then $\varphi(x) = \varphi(ha) = (ha)(a^{-1}b) = hb \in Hb$. Thus, our function is well defined. It is one-to-one, because if $\varphi(x) = \varphi(y)$, then $xa^{-1}b = ya^{-1}b$, and multiplying on the right by $b^{-1}a$ gives us that $x = y$. It is onto, because if we choose the arbitrary element $hb \in Hb$, then $\varphi(ha) = ha(a^{-1}b) = hb$.

For finite sets, a one-to-one and onto function establishes that two sets have the same number of elements. And because $H = H1$ is itself a right coset, all right cosets have the same size as H. □

We now introduce some notation, which allows us to state Lagrange's Theorem conveniently. If X is a finite set, let $|X|$ be the number of elements in X. For a group G, we call $|G|$ its **order**. If G is a finite group with subgroup H, we let $[G : H]$ denote the number of distinct cosets of H in G, called the **index of H in G**.

Example 31.5

Let's consider the group $G = U(\mathbb{Z}_{21})$ and the subgroup $H = \{1, 4, 16\}$ as in Example 31.4. Then $|H|$=3. Note that part (d) of Theorem 31.1 tells us that $|Ha| = 3$ for any right coset Ha. Of course, we computed the right cosets explicitly above and saw that this is so. Finally, those computations show us that $[G : H] = 4$. Note that there are 12 elements in G altogether (that is, $|G| = 12$), and we can arrive at this number by counting 4 cosets, each with 3 elements.

▷ **Quick Exercise.** Repeat this reasoning for Examples 31.2 and 31.3 above. ◁

Let's now state the general theorem reflected in these examples:

Theorem 31.2 Lagrange's Theorem *Let G be a finite group with subgroup H. Then*

$$|G| = [G : H]|H|.$$

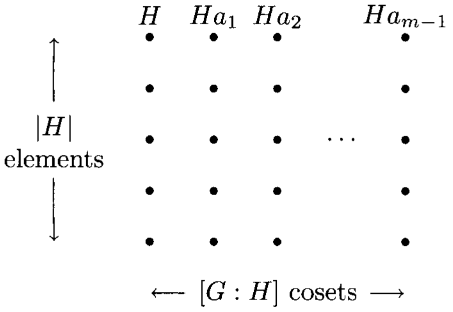
In particular, this means that $|H|$ divides $|G|$.

Proof: Suppose that $[G : H] = m$. Every element of G is in a coset of H, and part (b) of Theorem 31.1 tells us we can decompose G into a union of m pairwise disjoint cosets:

$$G = H \cup Ha_1 \cup Ha_2 \cup \cdots \cup Ha_{m-1}.$$

But each of these cosets has $|H|$ elements. Thus, there must be $[G : H]|H|$ elements in G altogether. □

This is illustrated in the following picture.



Lagrange did his work well before the notion of group had been defined. But he is honored for the theorem named after him because he applied a concrete version of this theorem in his arguments regarding permutations of the roots of a polynomial equation.

31.3 Applications of Lagrange's Theorem

Lagrange's Theorem is a very important one in the theory of finite groups. It places great restrictions on the sort of elements and the sort of subgroups a finite group can have. We make one such inference immediately:

Corollary 31.3 *Let G be a finite group, and $g \in G$. Then the order of g divides the order of G.*

Proof: Consider the cyclic subgroup $\langle g \rangle$ of G generated by g. If the order of g is n, then this subgroup has n elements. By Lagrange's Theorem, n must divide $|G|$. □

Thus, no group of order 15 could possibly have elements of order 2 or 6.

▷ **Quick Exercise.** What are the orders of the elements of the group of symmetries of the cube? Do all these orders divide 24? (Refer to Chapter 23.) ◁

Let's apply this reasoning in the case where the number of elements in a group is a prime p. Now the only positive divisors of the prime p are 1 and p itself. Of course, every group has exactly one element of order 1: namely, the identity. Thus, every other element of the group has order p. But this means that in a group with prime order, every non-identity element has that prime order. And so, the group must be cyclic! (In fact, it is cyclic where *every* non-identity element generates the group.)

Corollary 31.4 *Every group of prime order is cyclic.*

It is important to note that the converse of Lagrange's Theorem 31.2 is false. That is, suppose that G is a finite group with n elements, and m is a divisor of n. We *cannot* infer that G has a subgroup with exactly m elements. (This statement is true for abelian groups; we will meet this as Corollary 35.2.) Here is an example that shows this is true; we prove that the example works, as another application of Lagrange's Theorem 31.2:

Example 31.6

Consider the subgroup of S_4 isomorphic to the symmetric group of the regular tetrahedron. This group has twelve elements, as we discovered in Chapter 23; we will denote this group by A_4. Using the disjoint cycle notation of Chapter 30, the elements of this group are

$$\iota,\ (123),\ (132),\ (124),\ (142),\ (134),\ (143),$$
$$(234),\ (243),\ (12)(34),\ (13)(24),\ (14)(23).$$

If the converse of Lagrange's Theorem were true, then A_4 would have a subgroup H with six elements. By Lagrange's Theorem, this subgroup would have only two distinct cosets. Now pick any element α of A_4 with order 3. Because H has only two distinct cosets, at least two of the cosets $H\iota$, $H\alpha$, and $H\alpha^2$ must be the same. But then, by Theorem 31.1, $\alpha \in H$.

▷ **Quick Exercise.** Why? ◁

So all the elements of A_4 with order 3 belong to the subgroup H. But there are eight such elements, which must all belong to a subgroup with six elements. This contradiction implies that A_4 has no subgroup with six elements. (This elegant argument is due to Joseph Gallian.)

We can however prove a very limited converse to Lagrange's Theorem, which asserts that *prime* divisors of the order of a finite group do lead to subgroups of that order. This theorem is the first of several known as the Sylow theorems. The proof we offer here is a clever one due to James McKay, based on counting a set in two different ways.

Theorem 31.5 *Suppose that p is a prime integer, G is a group, and p divides $|G|$. Then G has an element of order p (and so a subgroup of order p).*

Proof: Consider the set S of all lists $a_1, a_2, \cdots, a_p$, where each $a_i \in G$, and $a_1 a_2 \cdots a_p = 1$. Note that we are allowing for repetitions in the a_i's, and in fact, we are hoping exactly to find an element (other than the identity) of S for which all the entries are the same; this will give us the required element of G with order p.

We are going to count the number of elements in S. For each entry in the list (until the last one), we can choose any element of the group;

but the last entry in the list is required to be $(a_1a_2\cdots a_{p-1})^{-1}$. This means that S has exactly $|G|^{p-1}$ elements.

We will now recount this set. Given a list $a_1, a_2, \cdots, a_p$, consider what happens when we cyclicly permute it. That is, we consider lists of the form $a_k, a_{k+1}, \cdots, a_p, a_1, \cdots, a_{k-1}$. First of all, note that these permuted lists remain in S.

▷ **Quick Exercise.** Why must these permuted lists remain in S? ◁

Second, if at least two of the a_i's are distinct, we obtain a total of p lists from the original one (counting the original when we perform the "trivial" cyclic permutation). However, if all of the a_i's are the same, we do not obtain any new lists. We can thus partition the elements of the set S into the subsets consisting of a given element and the other lists resulting from its cyclic permutations. Some of these subsets have p elements in them, while some have only 1 element in them, depending on whether any of the elements in one of the lists are distinct, or not. Those are the only two possibilities.

We now count the elements in S, by counting the number of elements in each of these subsets, and adding them up. So, let m be the number of these subsets with 1 element in them, and let n be the number of these subsets with p elements. This means that a complete count of all elements in S is given by $m \cdot 1 + n \cdot p$. Thus $|G|^{p-1} = m + np$.

Notice that $m > 0$, because there is at least one element (namely, 1) that when multiplied by itself p times gives 1. We will have completed the proof of the theorem if we can conclude that $m > 1$.

We now finally use the hypothesis of the theorem (that p divides $|G|$) to conclude that p divides m. Since $m > 0$, this means that $m \geq p > 1$. This proves the theorem. □

We shall now use Lagrange's Theorem to determine the number of distinct groups of small size. Notice first that because all cyclic groups of a given order are isomorphic, there is essentially only one group of order 2, 3, 5, etc.: namely, the cyclic groups

$$\mathbb{Z}_2,\ \mathbb{Z}_3,\ \mathbb{Z}_5, \cdots.$$

Let's take the smallest non-prime integer 4: What sort of groups of order 4 are there? Of course, there is $\mathbb{Z}_4$, the cyclic group of order 4. Are there any others?

If a group G has order 4 and is not cyclic, then every non-identity element in the group must be of order 2 (because only 1, 2, and 4 divide 4). Choose two of these elements, and call them a and b. Each is its own inverse (that's what being of order 2 means). So ab has to be the third non-identity element; but so does ba, and so $ab = ba$. (Note that Exercise 24.8 also implies that $ab = ba$.) Thus, we conjecture that the multiplication table for G must look like this:

	1	a	b	ab
1	1	a	b	ab
a	a	1	ab	b
b	b	ab	1	a
ab	ab	b	a	1

To show this is a group, we need to verify that it satisfies all the group axioms. The most tedious of these is associativity. But, more easily, we see that this is just a multiplicative version of the (additive) group $\mathbb{Z}_2 \times \mathbb{Z}_2$. This group is often called the **Klein Four Group**.

▷ **Quick Exercise.** Establish an explicit isomorphism between the multiplicative and the additive representations we have given for the Klein Four Group. ◁

So we have concluded that there are essentially only two groups of order 4: the cyclic group and the Klein Four Group.

There's only one group of order 5 (because 5 is prime). But what about groups of order 6? If such a group had an element of order 6, it would be cyclic. So in a non-cyclic group of order 6, all non-identity elements must be of order 2 or 3. By Theorem 31.5 we know that any such group must have at least one element a with order 3, and at least one element b of order 2. Then $1, a, a^2, b$ are all distinct elements of the group; note that $a^2 \neq b$, because a^2 also has order 3. We now claim that if we add the elements ab and a^2b to this list, we will have a complete list of all elements in the group. To show all of these elements are distinct, let's examine a couple of typical cases. If we suppose that $ab = 1$, then $a^2 = b$, which we have already concluded is impossible. And if $ab = a^2$, then $b = a$, which is clearly impossible.

▷ **Quick Exercise.** Finish checking that each of these elements is distinct from each of the others. ◁

We thus have that our group consists of the set $\{1, a, a^2, b, ab, a^2b\}$. But obviously ba must be an element too. It is easy to check that the only possibilities for this element are ab and a^2b. If $ba = ab$, then the group is abelian, and the element ab is an element of order 6.

▷ **Quick Exercise.** Check that ab is a generator in this case, either by direct calculation or else by using Exercise 26.5. ◁

So the only remaining case is if $ba = a^2b$. It follows from this that $ba^2 = a^2ba = a^4b = ab$. Is there such a group? Of course! This is exactly S_3 (where (123) could play the role of a, and (12) could play the role of b).

We have thus concluded that there are essentially only two groups of order 6: $\mathbb{Z}_6$ and S_3. Note that this makes S_3 the non-abelian group with smallest order.

You will inquire in Exercises 31.8 and 31.9 into the groups of order 8. It turns out that there are five of them.

Chapter Summary

In this chapter we defined the notion of *right coset* for a subgroup of a group. We determined that the set of right cosets of a subgroup divides the group into pairwise disjoint pieces, each of the same size. In the context of finite groups, this leads to the counting theorem known as *Lagrange's Theorem.* Lagrange's Theorem allows us to prove that the order of an element divides the order of the group, and that all groups of prime order are cyclic.

Warm-up Exercises

a. Determine the set of right cosets of the subgroup $\langle 6\rangle$ in $\mathbb{Z}_{20}$. Check that Lagrange's Theorem 31.2 holds in this case.

b. Determine the set of right cosets of the subgroup $\langle 7\rangle$ in $U(\mathbb{Z}_{20})$. Check that Lagrange's Theorem holds in this case.

c. Determine the set of right cosets of the subgroup $\{1, -1\}$ in the group of quaternions Q_8. Check that Lagrange's Theorem holds in this case. (Note that we are here using the shorthand notation of Exercise 28.14 for the quaternions.)

d. If H is a subgroup of G and two cosets of H share an element, then what can you say?

e. Is a coset ever a subgroup?

f. If $a \in Hb$, how are the two cosets Ha and Hb related?

g. Suppose that G is an infinite group, and H is a subgroup of G with finitely many elements. How many distinct cosets does H have?

h. Suppose that G is an infinite group, and H is a subgroup of G with finitely many cosets. How many elements does H have?

Exercises

1. Determine explicitly the right cosets of the subgroup $\langle(124)\rangle$ in the group A_4 (see Example 31.6).

2. Determine explicitly the right cosets of the subgroup $\langle(124)\rangle$ in the whole group S_4.

3. Find the *left* cosets of the subgroup $H = \{\iota, (12)\}$ of S_3. How do these cosets compare with the right cosets of H found in Example 31.2?

4. How do the left cosets of K compare with the right cosets found in Example 31.3?

5. If G is a group of order 8 and G is not cyclic, why must $x^4 = 1$ for all x in G? What similar statement can you make about a group of order 27? Generalize this further.

6. Let G be a group of order p^2, where p is prime. Show that every proper subgroup of G is cyclic.

7. Replicate the argument about the number of groups of order 6 for groups of order 10. You should conclude that there are essentially only two such groups; namely, $\mathbb{Z}_{10}$ and the dihedral group D_5.

8. (a) What are the possible orders for a subgroup of a group of order 8?

   (b) Prove that any group of order 8 must have a subgroup of order 2. (This is easy, and does not require that you use either Theorem 31.5 or the results of Exercise 9.)

9. In this problem we analyze how many groups there are of order 8. Consider the following cases (with hints):

   Case 1: There is an element of order 8.

   Case 2: If there is no element of order 8, but there is an element a of order 4, choose an element b outside $\langle a \rangle$. Now, either you can find such a b of order 2 or else all such b have order 4. If you can find such a b of order 2, then, depending on whether a and b commute, argue that we must obtain either $\mathbb{Z}_2 \times \mathbb{Z}_4$ or D_4. Now assume that all such b have order 4. Argue that you must obtain the quaternions Q_8.

   Case 3: Now suppose that there are no elements of order 4. So, all non-identity elements have order 2. First, argue that the group must be abelian. Then, pick three elements of order 2 and argue that you must obtain $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$.

10. Show that every non-identity element of a group generates the group if and only if the group is cyclic of prime order.

11. Let n be a positive integer and $1 \leq k \leq n$. Consider the stabilizer subgroup G_k of S_n. (See Exercise 29.6.) What does Lagrange's Theorem say about $[S_n : G_k]$? Demonstrate this explicitly, by placing the set of right cosets of G_k into a natural one-to-one correspondence with $\{1, 2, \cdots, n\}$.

12. (a) Suppose G is a finite group and $a \in G$. Show that $a^{|G|} = 1$.

    (b) Now prove Fermat's Little Theorem 8.7: *If p is prime and $0 < x < p$, then $x^{p-1} = 1 \pmod{p}$.*

    (c) For any positive integer n, $\phi(n)$ is defined to be the number of positive integers less than n that are also relatively prime to n. (This is called **Euler's phi function**.) So, $\phi(6) = 2$ because only 1 and 5 are less than 6 and also relatively prime to 6. Note that if p is prime, $\phi(p) = p - 1$. Show that if a is relatively prime to n, then

    $$a^{\phi(n)} = 1 \pmod{n}.$$

    This generalization of Fermat's Little Theorem 8.7 is called **Euler's Theorem**.

13. Refer to the definition in Exercise 12 of Euler's phi function.

    (a) Compute $\phi(7)$, $\phi(20)$, and $\phi(100)$.

    (b) Use Euler's Theorem to efficiently compute

    $$2^{38} \pmod{7}, \qquad 17^{18} \pmod{20}, \qquad 7^{82} \pmod{100}.$$

14. Prove that the set of (right) cosets of the subgroup $\mathbb{Z}$ in $\mathbb{R}$ can be placed in a one-to-one correspondence with the set of real numbers x, with $0 \leq x < 1$.

15. Consider the group $\mathbb{S}$, the unit circle, a multiplicative subgroup of $\mathbb{C}^*$. Provide a simple condition on complex numbers α, β characterizing when the cosets $\mathbb{S}\alpha$ and $\mathbb{S}\beta$ are equal, and prove that it works.

Chapter 32

Groups of Cosets

Our goal now is to place a group structure on the set of right cosets of a subgroup of a given group, just as we did for the cosets of an ideal in a given ring. Given a group G with a subgroup H, our experience with ring theory would suggest that the group operation on the set of cosets should be defined like this:

$$(Ha)(Hb) = Hab.$$

That is, to multiply two cosets, pick a representative of each, multiply them together, and then form its coset. It is clear that we must prove that this operation is *well defined.* (That is, it should be independent of the coset representative chosen.)

Before proceeding, let's look at an example of this; we return to Example 31.2. Let's try multiplying the cosets

$$H(123) = H(23) = \{(123), (23)\},$$

and

$$H(132) = H(13) = \{(132), (13)\},$$

using our provisional definition. If our definition is to work, and we compute the product of *any* element from the first coset, times *any* element of the second coset, we should always land in the same coset. In this case there are four possible products:

$$(123)(132) = \iota, \ (23)(132) = (12),$$

$$(123)(13) = (23), \ \text{and} \ (23)(13) = (123).$$

These elements *do not* all belong to the same coset! (Because, among other reasons, there are four distinct elements obtained and a coset of H must have only two elements.) Consequently, it does not seem that here we can make a group out of the set of right cosets of H in G.

32.1 Left Cosets

Let's beat a strategic retreat and return to a notion that we left behind in the previous chapter: there we defined *right* cosets. But it makes perfectly good sense to also speak of *left* cosets. Given a subgroup H of a group G, a **left coset** of H in G is a set of the form

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

Let's compute the left cosets for the subgroups H and K of S_3; we computed their right cosets in Examples 31.2 and 31.3.

Example 32.1

(See Example 31.2.) We still obtain exactly 3 distinct cosets; this time they are

$$\iota H = (12)H = \{\iota, (12)\},$$

$$(123)H = (13)H = \{(123), (13)\},$$

and

$$(132)H = (23)H = \{(132), (23)\}.$$

Example 32.2

(See Example 31.3.) We still obtain exactly 2 distinct cosets; this time they are

$$\iota K = (123)K = (132)K = \{\iota, (123), (132)\}$$

and

$$(12)K = (13)K = (23)K = \{(12), (13), (23)\}.$$

A comparison of these two examples reveals a real difference! For K, each of its left cosets is equal to the corresponding right coset, while this is false for H. Of course, if the group had been abelian, then the left and right cosets would obviously be equal, because multiplying on the left, or on the right, amounts to the same thing in an abelian group. And we know that coset addition works for rings. This suggests that this property might be of importance. In fact, it turns out to be *exactly* what we need to build a group of cosets.

32.2 Normal Subgroups

Let H be a subgroup of the group G. We say that H is a **normal subgroup** if $gH = Hg$, for all $g \in G$. If H is a normal subgroup, we can safely talk about its *cosets* (without an adjective). We denote the set of all these cosets by G/H.

Example 32.3

The subgroup $\langle(123)\rangle$ is a normal subgroup of S_3, while the subgroup $\langle(12)\rangle$ isn't.

Example 32.4

Every subgroup of an abelian group is normal.

In the next two chapters we will eventually obtain a host of examples of normal subgroups, but it's now time for the theorem we have been building up to. This is the group analogue of Theorem 18.2, where the idea of normal subgroups plays the role analogous to ideal in a ring.

Theorem 32.1 *Let H be a normal subgroup of G. Then the set G/H of cosets of H in G is a group, under the operation*

$$(Ha)(Hb) \quad = \quad Hab.$$

Proof: We must first check that the operation specified in the statement of the theorem is *well defined*; that is, it should be independent of coset representatives. This verification is the crucial part of the proof, and will depend in an essential way on the fact that H is normal.

Suppose then that $Ha = Hc$ and $Hb = Hd$. We claim that $Hab = Hcd$. By Theorem 31.1c, this amounts to checking that $ab(cd)^{-1} \in H$. Because $Hb = Hd$, we know that $bd^{-1} \in H$. Now H is a normal subgroup, and so $aH = Ha$. This means that

$$a(bd^{-1}) = ha$$

for some $h \in H$. Thus,

$$ab(cd)^{-1} = abd^{-1}c^{-1} = hac^{-1}.$$

But $Ha = Hc$ means exactly that $ac^{-1} \in H$, and so therefore $hac^{-1} \in H$ too. This completes the proof that the operation is well defined.

The operation is clearly associative:

$$(HaHb)(Hc) = HabHc = H(ab)c =$$

$$Ha(bc) = HaHbc = Ha(HbHc).$$

The element $H1$ serves as the identity for G/H:

$$H1Ha = H1a = Ha = Ha1 = HaH1.$$

And the element Ha has an inverse; namely, Ha^{-1}:

$$HaHa^{-1} = Haa^{-1} = H1 = Ha^{-1}a = Ha^{-1}Ha.$$

Thus, G/H is a group, as we claim. □

The proof of this theorem is more complicated than that for the corresponding ring theorem, because the group need not be abelian. The crucial point in the proof is showing that the operation on G/H is well defined. Loosely speaking, the normality of H says that elements of G commute with H, and this is just enough to make the argument work.

We call G/H the **group of cosets of G modulo H**; we also call G/H a **quotient group**.

We have encountered many examples of this construction already: Merely take a ring of cosets as we discussed in Chapter 18 and thereafter, and forget the multiplicative structure. However, these examples do not call sufficient attention to the importance of the normality of the subgroup, because, as we've pointed out before, *all* subgroups of an abelian group are normal. Consequently, we must examine some purely group-theoretic examples, involving some non-abelian groups.

Before we do this, however, we will provide a characterization of normality that is slightly easier to work with. We will inquire further into this version of normality in the exercises, and in future chapters. For the moment, we view the following proposition only as a computational aid to checking normality.

Theorem 32.2 *Let H be a subgroup of the group G. Then H is normal in G if and only if $g^{-1}Hg = H$, for all $g \in G$. In fact, this is true if and only if $g^{-1}Hg \subseteq H$, for all $g \in G$.*

Proof: Suppose that H is normal in G and $g \in G$; then $Hg = gH$. So for any $h \in H$, there exists $h_1 \in H$ such that $hg = gh_1$. But then $g^{-1}hg = h_1 \in H$. Thus, $g^{-1}Hg \subseteq H$. A very similar argument, which we save for Exercise 32.1, shows that the reverse inclusion also holds.

Conversely, suppose that $g^{-1}Hg = H$ for all $g \in G$. Given any element $h \in H$, this means that $h = g^{-1}h_1g$, for some $h_1 \in H$. But then $gh = h_1g$; that is, $gH \subseteq Hg$. You will prove the reverse inclusion in Exercise 32.2.

If $g^{-1}Hg \subseteq H$ for *all* g, this applies to g^{-1} as well: that is, $gHg^{-1} \subseteq H$. But this latter statement means that $H \subseteq g^{-1}Hg$, and so these two sets are actually equal. □

32.3 Examples of Groups of Cosets

Example 32.5

Consider the group A_4 in Example 31.6:

$$A_4 = \{\iota, (234), (243), (143), (134), (124), (142),$$

$$(132), (123), (12)(34), (13)(24), (14)(23)\}.$$

Consider the subgroup

$$H = \{\iota, (12)(34), (13)(24), (14)(23)\}.$$

Notice that H is a group with four elements that is not cyclic; that is, H is isomorphic to the Klein Four Group (see the discussion in Section 31.3). We claim that H is normal. Let's perform a few of the computations necessary to check this:

$$(132)^{-1}(12)(34)(132) = (123)(12)(34)(132) = (14)(23),$$

$$(243)^{-1}(12)(34)(243) = (234)(12)(34)(243) = (13)(24),$$

$$(124)^{-1}(13)(24)(124) = (142)(13)(24)(124) = (12)(34).$$

In each case, we see that an element of the form $\beta^{-1}\alpha\beta$, where $\beta \in A_4$ and $\alpha \in H$, remains in H. It would be tedious indeed to check all the other cases!

▷ **Quick Exercise.** Perform two more of the required calculations. ◁

In Exercise 32.16 you will provide a complete proof that H is normal in A_4, cleverer than just verifying all the cases. We will also return to considering the normality of H in Example 33.5. For the moment, we shall accept that this is true.

We can thus construct the group A_4/H. Let's use Lagrange's Theorem 31.2 to determine the size of this group. The order of A_4 is 12, while the order of H is 4. Thus, the order of A_4/H must be $12/4 = 3$. But there is essentially only one group of order 3, the cyclic group. Let's write out the elements of this group of cosets, and see why it should be isomorphic to $\mathbb{Z}_3$:

$$A_4/H = \{\iota, H(123), H(132)\}.$$

▷ **Quick Exercise.** Check that these three cosets are distinct, and so this list is really a complete list of the elements of A_4/H. ◁

But now it is clear that A_4/H is a cyclic group generated by $H(123)$.

Example 32.6

Consider now the group $\mathcal{G} = U(M_2(\mathbb{R}))$ of 2×2 invertible matrices. Recall that this consists of those matrices with non-zero determinant. (See Exercise 8.2.) Now consider the set

$$\mathcal{H} = \{A \in \mathcal{G} : \det(A) = 1\}.$$

First note that this is a subgroup. (You actually showed this in Exercise 25.18.) Given $A, B \in \mathcal{H}$, we need to check that $AB^{-1} \in \mathcal{H}$. But

$$\det(AB^{-1}) = \det(A)(\det(B))^{-1} = 1 \cdot 1 = 1.$$

This shows that $\mathcal{H}$ is a subgroup.

Furthermore, we claim that $\mathcal{H}$ is a normal subgroup. Because, given $A \in \mathcal{H}$ and $C \in \mathcal{G}$, we claim that $C^{-1}AC \in \mathcal{H}$. But

$$\det(C^{-1}AC) = (\det(C))^{-1}\det(A)\det(C) = \det(A) = 1.$$

This shows that $\mathcal{H}$ is normal.

We thus know that $\mathcal{G}/\mathcal{H}$ is a group. Let's see if we can determine the nature of this group. Let's look at a particular coset $\mathcal{H}A$. A typical element of this coset looks like BA, where $\det(B) = 1$. But

$$\det(BA) = \det(B)\det(A) = \det(A).$$

Thus, all elements of $\mathcal{H}A$ have the same determinant. In fact, we claim that $\mathcal{H}A$ consists exactly of all those matrices with determinant equal to the non-zero real number $\det A$. To see this, suppose that $C \in \mathcal{G}$, and $\det(C) = \det(A)$; we claim that $C \in \mathcal{H}A$. But $C = CA^{-1}A$, and $\det(CA^{-1}) = 1$, and so $C \in \mathcal{H}A$.

We thus have established a one-to-one correspondence between the elements of the group $\mathcal{G}/\mathcal{H}$ and the non-zero real numbers. This suggests the possibility that $\mathcal{G}/\mathcal{H}$ is isomorphic to $\mathbb{R}^*$, the multiplicative group of non-zero real numbers. We could prove this now but will instead wait until the next chapter, when we can prove this carefully, and easily, using the Fundamental Isomorphism Theorem for Groups 33.4; see Example 33.8.

We can much enlarge our fund of examples, by showing that in the special case when a subgroup has index 2 in the group, it is automatically normal:

Theorem 32.3 Index 2 Theorem *Suppose that G is a group with subgroup H, and $[G : H] = 2$. Then H is a normal subgroup of G, and G/H is isomorphic to $\mathbb{Z}_2$.*

Proof: Suppose that H is a subgroup of G and $[G : H] = 2$. This means that there are only two right cosets of H in G: We can write these two cosets as $H1$ and Hg, where g is any element of G that does not belong to H. But by Lagrange's Theorem 31.2, there can also be only two left cosets of H, and they must be of the form $1H$ and gH. But because $H1 = H = 1H$, this means that $Hg = gH$, for all $g \notin H$. In other words, H is normal in G.

Because $[G : H] = 2$, we know that G/H is a group with two elements and there is only one such group (up to isomorphism), namely, $\mathbb{Z}_2$. □

Example 32.7

Consider the subgroup A_4 of S_4 (see Example 31.6). This subgroup has 12 elements, and so its index in S_4 is 2. Theorem 32.3 then implies that G is a normal subgroup of S_4. We will generalize this example in Chapter 34.

Example 32.8

Consider the dihedral group D_n, where we will here use the notation introduced in Exercise 28.12. The cyclic subgroup $\langle R\rangle$ is clearly the subgroup of rotations in D_n and has n elements. But because D_n has $2n$ elements, $\langle R\rangle$ is a normal subgroup of D_n.

Chapter Summary

In this chapter we defined a *normal subgroup* and proved that if a subgroup is normal, then a group structure can be given to the set of cosets, which we then call a *quotient group.*

Warm-up Exercises

a. Given a subgroup H of a group G, at least one left coset of H always equals its corresponding right coset. Why?

b. Why is every subgroup normal, if the group is abelian? (The converse is false; see Exercise 10 below.)

c. If H is a normal subgroup of the group G, and $[G : H] = 7$, to what group is G/H necessarily isomorphic?

d. The ring $\mathbb{Z}$ is a subring of $\mathbb{R}$ but is not an ideal. (Why?) We thus were unable to form the ring of cosets $\mathbb{R}/\mathbb{Z}$. But $\mathbb{Z}$ is clearly a normal subgroup, and so we are able to form the group of cosets $\mathbb{R}/\mathbb{Z}$. Explain this. (For more information about the group $\mathbb{R}/\mathbb{Z}$, see Exercise 33.1.)

e. Show that $\{1\}$ is a normal subgroup of any multiplicative group G.

f. Consider the set

$$\mathcal{K} = \{A \in U(M_2(\mathbb{R})) : \det(A) = 3\}.$$

By using an argument just like that in Example 32.6, it is easy to see that if $A \in \mathcal{K}$, then $C^{-1}AC \in \mathcal{K}$, for all matrices $C \in U(M_2(\mathbb{R}))$. Does this make $\mathcal{K}$ a normal subgroup of $U(M_2(\mathbb{R}))$? (Be careful!)

Exercises

1. Suppose that H is a normal subgroup of the group G. Prove that $H \subseteq g^{-1}Hg$. *Note:* This is an omitted verification in the proof of Theorem 32.2.

2. Suppose that H is a subgroup of the group G, and $g^{-1}Hg = H$, for all $g \in G$. Prove that $Hg \subseteq gH$. *Note:* This is another omitted verification in the proof of Theorem 32.2.

3. Are $\langle(124)\rangle$ and $\langle(12)\rangle$ a normal subgroups of S_4? (Consider Exercise 31.2.) Is $\langle(124)\rangle$ a normal subgroup of A_4? (Consider Exercise 31.1.)

4. Is $\{\iota, (123), (132)\}$ a normal subgroup of S_3? (Consider Example 31.3.)

5. Suppose that $n > 2$ is a positive integer and $1 \leq k \leq n$. Show that the stabilizer subgroup G_k is not normal in S_n. (See Exercise 29.6.)

6. Let H be a (not necessarily normal) subgroup of the group G, and $g \in G$. Prove that $g^{-1}Hg$ is always a subgroup of G. What group is it, if H is normal? The subgroup $g^{-1}Hg$ is called a **conjugate** of the subgroup H. Argue that if H_1 is a conjugate of the subgroup H_2, then H_2 is a conjugate of H_1 (that is, conjugacy is a *symmetric* relation).

7. Suppose that n is a positive integer, and d is a positive integer that (properly) divides n. Prove that

$$\mathbb{Z}_n/\langle d\rangle$$

is isomorphic to $\mathbb{Z}_d$. *Note:* This exercise foreshadows developments in the next chapter and will be easy to do with the main theorem of that chapter; your experience with rings should suggest the appropriate approach. See Exercise 33.4.

8. Suppose that H and K are normal subgroups of the group G. Prove that $H \cap K$ is a normal subgroup. *Note:* In Exercise 25.5 you have already shown that $H \cap K$ is a subgroup.

9. Consider the direct product $G \times H$ of the groups G and H. Prove that the subgroups $G \times \{1\}$ and $\{1\} \times H$ are normal subgroups of $G \times H$.

10. Show that every subgroup of the group Q_8 of quaternions is normal, even though the group is not abelian. (See Warm-up Exercise b.)

11. Consider the subgroup $H = \{1, -1\}$ of Q_8. In the previous exercise, you showed that H is a normal subgroup of Q_8. Give the group table for Q_8/H. To what group is this group of cosets isomorphic?

12. Consider the group of cosets $\mathbb{R}^*/\mathbb{Q}^*$. Exhibit an element of this group with order 2. What about an element of order n, for any positive integer n? Do you believe that there are elements of infinite order in this group? (You should be able to conjecture an element that works, but you will not be able to prove this rigorously.)

13. Suppose that H and K are subgroups of the group G, and K is a normal subgroup. Let $HK = \{hk : \ h \in H, \ k \in K\}$.

    (a) Prove that HK is a subgroup of G.

    (b) Suppose in addition that H is normal. Prove that HK is a normal subgroup of G.

14. Suppose that G is a group with subgroups H, K. Suppose further that H is a normal subgroup and $H \subseteq K \subseteq G$. Prove that H is a normal subgroup of the group K.

15. Suppose G is a group, and that $g, h \in G$. Furthermore, assume that the order of h is n. Prove that the order of $g^{-1}hg$ is n. The element $g^{-1}hg$ is called a **group conjugate** of the element h; an element of group shares many properties with its conjugates.

16. Consider the groups A_4 and H described in Example 32.5. Use Exercise 15 to provide a complete proof that H is normal in A_4.

17. Let G be an abelian group, and let

$$t(G) = \{g \in G : o(g) \text{ is finite}\}.$$

We call $t(G)$ the **torsion subgroup** of G.

(a) Prove that $t(G)$ is a subgroup of G.

(b) Prove that $G/t(G)$ is torsion-free. (Recall that this means that G has no non-identity elements of finite order.)

18. Suppose that H is a normal subgroup of the group G. Prove that G/H is an abelian group if and only if $g^{-1}h^{-1}gh \in H$, for all $g, h \in G$. *Note:* Elements of the form $g^{-1}h^{-1}gh$ are called **commutators** in the group G.

19. We generalize Exercise 18 slightly. Given a group G, define H to be the smallest subgroup of G that contains all the commutators. The subgroup H is called the **commutator subgroup**. Argue that H consists of all finite products of commutators. For a normal subgroup K of G, prove that G/K is abelian if and only if $K \supseteq H$.

Chapter 33

The Isomorphism Theorem for Groups

In this chapter we prove the *Isomorphism Theorem for Groups*, the important theorem analogous to the ring theory result we obtained in Chapter 19. The ring theory theorem asserts that knowing about ideals is essentially the same as knowing about ring homomorphisms. We will establish a similar connection between normal subgroups and group homomorphisms.

33.1 The Kernel

Let's begin with a homomorphism $\varphi : G \to H$ between the groups G and H. By way of analogy with ring theory, we define the **kernel** of φ to be

$$\ker(\varphi) = \{g \in G : \varphi(g) = 1\}.$$

We can express this colloquially: The kernel of φ is the set of elements in G that are sent to the identity in H by φ. We can use a slightly different notation to emphasize the definition of the kernel. It is the **pre-image** of the identity; that is,

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

The following theorem should not be a surprise:

Theorem 33.1 *Let $\varphi : G \to H$ be a homomorphism between the groups G and H. Then* $\ker(\varphi)$ *is a normal subgroup of G.*

Proof: To show that $\ker(\varphi)$ is a subgroup, choose two elements $a, b \in \ker(\varphi)$; we must show that ab^{-1} is in the kernel (by the Subgroup

Theorem 25.2). To show this, we compute:

$$\varphi\left(ab^{-1}\right) = \varphi(a)(\varphi(b))^{-1} = 1.$$

That is, $ab^{-1} \in \ker(\varphi)$.

To show that $\ker(\varphi)$ is normal, we must choose an arbitrary element $g \in G$, and check that

$$g^{-1}\ker(\varphi)g \subseteq \ker(\varphi)$$

(by Theorem 32.2). For that purpose, choose $a \in \ker(\varphi)$, and compute again:

$$\varphi\left(g^{-1}ag\right) = (\varphi(g))^{-1}\varphi(a)\varphi(g) = (\varphi(g))^{-1}1\varphi(g) = 1.$$

That is, $g^{-1}ag \in \ker(\varphi)$, as required. □

Let's look at some examples of kernels of homomorphisms.

▷ **Quick Exercise.** Review several of the examples of kernels of ring homomorphisms, discussed in Chapter 17. ◁

Next, let's compute the kernels of a number of the homomorphisms in the examples in Chapter 27.

Example 33.1

(Example 27.2) Consider the function $\rho : \mathbb{Z} \to \mathbb{Z}$ that multiplies elements by 3. The equation $\rho(n) = 3n = 0$ has only one solution, and so the kernel is $\{0\}$, which is evidently a normal subgroup.

Example 33.2

(Example 27.3) Consider the logarithm function $\log : \mathbb{R}^+ \to \mathbb{R}$. The equation $\log(r) = 0$ has a unique solution, and so again the kernel here is the trivial subgroup, this time written multiplicatively: $\{1\}$.

Example 33.3

(Example 27.4) Consider the determinant function

$$\det : U(M_2(\mathbb{R})) \to \mathbb{R}^*.$$

Here, the kernel is the subgroup

$$\mathcal{H} = \{A \in U(M_2(\mathbb{R})) : \det(A) = 1\}.$$

It is difficult to describe more succinctly the subgroup $\mathcal{H}$. But note that it has infinitely many elements.

▷ **Quick Exercise.** Provide 4 distinct matrices belonging to $\mathcal{H}$. ◁

By Theorem 33.1, $\mathcal{H}$ is normal; in Example 32.6 we showed directly that this is the case. We will return to this example yet again, later in this chapter.

Example 33.4

(Example 27.5) Consider the homomorphism $\varphi : D_3 \to \{1, -1\}$ that takes the rotations $\{1, \rho, \rho^2\}$ to 1, and the remaining elements (the flips) to -1. Obviously, the kernel is the subgroup $\langle\rho\rangle$. According to Theorem 33.1, this is a normal subgroup.

Example 33.5

(Example 27.6) Consider the homomorphism $\Phi : G \to \mathbb{Z}_3$, where G is the group of symmetries of the tetrahedron. (We've seen in Examples 31.6 and 32.5 that G is (isomorphic to) the group A_4.) If you check the discussion in Example 27.6, it is obvious that the kernel is

$$\{\iota, \varphi_1, \varphi_2, \varphi_3\}.$$

In Example 32.5 we started to verify that this subgroup (or rather, the corresponding subgroup of the corresponding group of permutations) is normal, by brute force. It follows easily here, now that we note that it is the kernel of a homomorphism. (If you did Exercise 32.16, you obtained another proof that this subgroup is normal.)

Example 33.6

(Example 27.7) Consider the differentiation function $D : \mathbb{R}[x] \to \mathbb{R}[x]$. The kernel consists of exactly those polynomials whose derivative is the zero polynomial. That is, the kernel consists of the constant polynomials:

$$\ker(D) = \mathbb{R} \subseteq \mathbb{R}[x].$$

Example 33.7

(Examples 27.8 and 27.9) The kernel of the embedding homomorphism $\epsilon : G \to G \times H$ defined by $\epsilon(g) = (g, 1)$ is the trivial subgroup of G. The kernel of the projection homomorphism $\pi : G \times H \to G$ defined by $\pi(g, h) = g$ is the subgroup

$$\{1\} \times H = \{(1, h) : h \in H\}.$$

▷ **Quick Exercise.** Verify that the kernels of Example 33.7 are what is claimed. Check directly that these are normal subgroups (or see Exercise 32.9). ◁

33.2 Cosets of the Kernel

The kernel contains indirectly more information than just the pre-image of the identity. Just as for rings, we have:

Theorem 33.2 *Let $\varphi : G \to H$ be a homomorphism between the groups G and H. Then $\varphi^{-1}(h)$ equals the coset $\ker(\varphi)g$, where g is any given element of $\varphi^{-1}(h)$.*

Proof: This is left for you to prove, using the proof of the corresponding ring result (Theorem 17.2) as a model; see Exercise 33.6. □

An important special consequence of this theorem follows in the case when the kernel is the trivial subgroup. In that case, *every* inverse image of the homomorphism consists of a single element. In other words, the homomorphism in question is one-to-one. This is directly analogous to the results for rings (Corrollary 17.4).

Corollary 33.3 *A group homomorphism is one-to-one if and only if its kernel is the trivial subgroup.*

▷ **Quick Exercise.** Check this corollary for each of the examples above. ◁

We now know that each group homomorphism leads to a normal subgroup, namely, its kernel. But each normal subgroup likewise leads to a homomorphism (of which it is the kernel). This requires another definition, again analogous to the ring case:

Let H be a normal subgroup of the group G. Then form the group of cosets G/H. Consider the function

$$\nu : G \to G/H$$

defined by $\nu(g) = Hg$. It is quite evident (from the definition of the group operation on G/H) that this is a homomorphism, that its kernel is exactly H, and that it is onto. (See Exercise 33.3.)

We call ν the **natural homomorphism from G onto G/H.**

33.3 The Fundamental Theorem

It remains to show that any onto group homomorphism is 'essentially the same' as the natural homomorphism from the domain group onto the group of cosets formed from the kernel. This is the next theorem.

Theorem 33.4 Fundamental Isomorphism Theorem for Groups *Let $\varphi : G \to H$ be an onto homomorphism between groups, and let $\nu : G \to G/\ker(\varphi)$ be the usual natural homomorphism. Then there exists an isomorphism $\mu : G/\ker(\varphi) \to H$ such that $\mu \circ \nu = \varphi$.*

We exhibit this situation in the following diagram:

$$\begin{tikzcd}
G \arrow[rr, "\varphi"] \arrow[dr, "\nu"'] & & H \\
& G/\ker(\varphi) \arrow[ur, "\mu"'] &
\end{tikzcd}$$

Proof: Suppose that G and H are groups, $\varphi : G \to H$ is an onto group homomorphism, and $\nu : G \to G/\ker(\varphi)$ is the natural homomorphism. Clearly, what we need to do is define a function μ, and prove that it has the desired properties. In this proof (unlike the ring case), you will be doing the required verifications.

Choose an arbitrary element of $G/\ker(\varphi)$. Such an element is a coset of the form $\ker(\varphi)g$, where $g \in G$. What element of H should it correspond to? If the composition of functions required in the theorem is to work as we wish, we must have that

$$\mu(\ker(\varphi)g) = \varphi(g).$$

And so, this is how we define the map $\mu : G/\ker(\varphi) \to H$.

There is an immediate problem we must solve: This definition apparently *depends on the particular representative* g. That is, our function does not appear to be unambiguously defined. But it is well defined; you will check this in Exercise 33.5.

After showing that μ is well defined, we must next show that μ is a group isomorphism. This requires showing that:

μ preserves the group operation,

μ is onto,

and that

μ is one-to-one.

You check these things in Exercise 33.5. □

We now look at a couple of examples of this theorem.

▷ **Quick Exercise.** Review the ring theory examples in Chapter 19. ◁

Example 33.8

Consider again the determinant homomorphism

$$\det : U(M_2(\mathbb{R})) \to \mathbb{R}^*.$$

This homomorphism is onto.

▷ **Quick Exercise.** Why? ◁

We have calculated its kernel in Example 33.3: It is $\mathcal{H}$, the set of all matrices with determinant 1. The Fundamental Isomorphism Theorem for Groups 33.4 now allows us to conclude that the groups $U(M_2(\mathbb{R}))/\mathcal{H}$ and $\mathbb{R}^*$ are isomorphic, as we conjectured in Example 32.6. In fact, from the proof of the Fundamental Isomorphism Theorem we can extract the function that establishes the isomorphism; namely, $\mu(\mathcal{H}A) = \det(A)$.

The inverse function, which is also necessarily an isomorphism, should assign to each non-zero number r the coset of a matrix whose determinant is r. In the previous chapter we laboriously determined that each coset of $\mathcal{H}$ consists of those matrices with a certain determinant. We proved this again in this chapter, in a more general context, because we proved that the cosets of $\mathcal{H} = \ker(\varphi)$ are exactly the inverse images of real numbers under the function det. Because of this, to define the inverse function of μ, *any* choice of such a matrix with the appropriate determinant will work as a representative for the coset. Here is a particularly understandable version:

$$\mu^{-1}(r) = \mathcal{H}\begin{pmatrix} r & 0 \\ 0 & 1 \end{pmatrix}.$$

Example 33.9

Let's return to the derivative function of Example 33.6. It is evidently onto $\mathbb{R}[x]$.

▷ **Quick Exercise.** Prove that the derivative function is onto. (Use a little calculus!) ◁

Thus, the Fundamental Isomorphism Theorem 33.4 says that the groups $\mathbb{R}[x]/\mathbb{R}$ and $\mathbb{R}[x]$ are isomorphic. (Isn't this strange?)

Chapter Summary

In this chapter we introduced the notion of the *kernel* of a group homomorphism. It is a normal subgroup, and every normal subgroup is the kernel of some homomorphism. Furthermore, we proved the *Fundamental Isomorphism Theorem for Groups*, which asserts that every onto homomorphism can be viewed as a natural homomorphism onto the group modulo its kernel.

Warm-up Exercises

a. Explain why every normal subgroup is the kernel of some homomorphism.

b. Explain why knowing the kernel of a homomorphism tells us essentially everything about the homomorphism.

c. Can you tell if a homomorphism is onto, by just looking at its kernel?

d. Can you tell if a homomorphism is one-to-one, by just looking at its kernel?

Exercises

1. Consider the function $\varphi : \mathbb{R} \to \mathbb{S}$ defined by

$$\varphi(r) = e^{2\pi r i} = \cos(2\pi r) + i\sin(2\pi r),$$

where $\mathbb{S}$ is the unit circle of Example 24.13. Show that φ is an onto homomorphism. What is the kernel of φ? What two groups are isomorphic, according to the Fundamental Isomorphism Theorem 33.4?

2. Consider the subgroup G of $U(M_2(\mathbb{R}))$ given by

$$G = \left\{ \begin{pmatrix} a & 0 \\ b & 1 \end{pmatrix} : a \neq 0 \right\}.$$

(See Exercise 25.19.) Define $\varphi : G \to \mathbb{R}^*$ by letting

$$\varphi \begin{pmatrix} a & 0 \\ b & 1 \end{pmatrix} = a.$$

(a) Prove that this is an onto group homomorphism.

(b) What is the kernel of this homomorphism?

(c) What two groups does the Fundamental Isomorphism Theorem assert are isomorphic?

(d) What is the relationship between this exercise and Example 33.8?

3. Suppose H is a normal subgroup of G. Show that the natural homomorphism $\nu : G \to G/H$ is indeed a homomorphism, is onto, and has kernel equal to H, as claimed in Section 33.2.

4. Let's redo Exercise 32.7, in light of the Fundamental Isomorphism Theorem. Suppose that n is a positive integer and d is a positive integer which (properly) divides n. Prove that

$$\mathbb{Z}_n/\langle d\rangle$$

is isomorphic to $\mathbb{Z}_d$.

5. Complete the proof of the Fundamental Isomorphism Theorem 33.4, relying if necessary on the corresponding proofs for the Fundamental Isomorphism Theorem for Commutative Rings 19.1:

(a) Show that the function μ is well defined (that is, it does not depend on the coset representative chosen).

(b) Show that the function μ is a group isomorphism (that is, that it preserves the group operation and is one-to-one and onto).

6. Prove Theorem 33.2, using Theorem 17.2 as a model, if necessary: Suppose that $\varphi : G \to H$ is an onto group homomorphism, and $h \in H$. Prove that $\varphi^{-1}(h)$ is a coset of $\ker(\varphi)$.

7. Let G be a group and fix some $g \in G$. Consider the map $\varphi : \mathbb{Z} \to G$ defined by $\varphi(n) = g^n$. Show that φ is a group homomorphism. What are the possible kernels for φ? Describe the image of φ in G.

8. Consider the function $\varphi : \mathbb{C}^* \to \mathbb{R}^*$ defined by $\varphi(\alpha) = |\alpha|$. Prove that φ is an onto homomorphism. What two groups are isomorphic, according to the Fundamental Isomorphism Theorem?

9. Suppose that G is a group with normal subgroups H, K. Consider the normal subgroups $H \cap K$ and HK. (See Exercises 32.8, 32.13, 32.14.) Prove that the groups $H/(H \cap K)$ and HK/K are isomorphic. (You should compare this to the corresponding result in ring theory, in Exercise 19.23.)

10. Consider the group $\mathbb{R}^3$, and suppose that $a, b, c \in \mathbb{R}$, not all zero. Define $\varphi : \mathbb{R}^3 \to \mathbb{R}$ by $\varphi(x, y, z) = ax + by + cz$. Show that φ is a group homomorphism. What is the kernel of φ (considered *geometrically*)? What does the Fundamental Isomorphism Theorem say in this context? What is the geometric meaning of this assertion? (Compare this exercise to Exercise 28.9.)

11. Consider the dihedral group D_4. Every element of D_4 can be written in the form F^iR^j, where $i = 0, 1$, and $j = 0, 1, 2, 3$, where F is the 'flip' about the vertical axis (12)(34), and R is the counterclockwise rotation (1432). (See Exercise 28.12.). Define

$$\psi : D_4 \to \mathbb{Z}_2 \times \mathbb{Z}_2$$

by setting $\psi(F^iR^j) = ([i]_2, [j]_2)$.

   (a) Prove that this is an onto homomorphism.

   (b) What is the kernel of this homomorphism?

   (c) The subgroup you obtained in part b is necessarily normal. (Why?) Prove this directly.

   (d) What two groups does the Fundamental Isomorphism Theorem assert are isomorphic?

12. Suppose that G is a group with normal subgroup H.

   (a) Define a correspondence between the subgroups of G containing H, and the subgroups of G/H, by assigning $K \to K/H$. Why is Exercise 32.14 relevant to this definition?

   (b) Show that the correspondence from part a is one-to-one and onto.

   (c) Furthermore, show that this correspondence also establishes a one-to-one correspondence between *normal* subgroups of G containing H and *normal* subgroups of G/H.

13. Suppose that H and K are normal subgroups of the group G, and that $K \subseteq H$. Then the group G/H is isomorphic to

$$(G/K)/(H/K).$$

(You should compare this to the corresponding result in ring theory, in Exercise 19.21.)

Chapter 34

The Alternating Groups

In this chapter we inquire further into the symmetry groups S_n. In the process we gain further insight into the importance and significance of normal subgroups.

34.1 Transpositions

For obvious reasons, we call a cycle of length 2 a **transposition**.

We now show that any permutation can be factored as a product of transpositions.

Theorem 34.1 *Any permutation can be factored as a product of transpositions.*

Proof: Because every permutation is a product of cycles (Theorem 30.3), it clearly suffices to show that each cycle can be factored as a product of transpositions. But this is easy: Consider the cycle $(a_1a_2 \cdots a_n)$ of length n. Clearly,

$$(a_1a_2 \cdots a_n) = (a_1a_n)(a_1a_{n-1}) \cdots (a_1a_3)(a_1a_2).$$

□

Notice that the method of factorization suggested by the proof of the theorem provides $n - 1$ transpositions for each cycle of length n. So the 5-cycle (12345) can be factored into a product of 4 transpositions:

$$(12345) = (15)(14)(13)(12).$$

The product of cycles can then also be factored into a product of transpositions:

$$(1345)(267) = (15)(14)(13)(27)(26).$$

However, this method of factoring into transpositions does *not* give us a factorization which is unique (unlike the factorization into disjoint cycles in Theorem 30.3).

Example 34.1

Here's a different factorization of the last permutation into transpositions:

$$(1345)(267) = (25)(67)(57)(25)(45)(35)(15).$$

(Don't worry about where this factorization came from.)

▷ **Quick Exercise.** Check that this factorization works. Can you come up with another essentially different factorization of this permutation into transpositions? ◁

Example 34.2

A surprising place where permutations and transpositions occur is in the theory of *bell ringing.* A bell ringer rings n bells. *Ringing the changes* means ringing each of the $n!$ orders of the n bells exactly once. The n bells are typically arranged in a row on a table and the ringer starts on the left and rings each bell in sequence. According to bell ringing practice, after a certain permutation is rung, the next permutation to be rung must be the same, except for one transposition. This is true because the bell ringer has only two hands with which to interchange two bells! Our theorem in essence asserts that we can get from any one order of the bells to any other, by repeated transpositions.

▷ **Quick Exercise.** Suppose that a bell ringer has six bells, in order $1, 2, 3, 4, 5, 6$. Give a list of transpositions, which when done one after another, results in the bells in order $6, 1, 4, 3, 5, 2$. ◁

34.2 The Parity of a Permutation

Our earlier example revealed that two distinct factorizations of a given permutation into transpositions need not even have the same number of factors. For example, our first factorization of (1345)(267) had five factors, while our second had seven factors. In fact, any given permutation can be expressed as a product of transpositions in an infinite number of ways. To see this, we need only notice that $\iota = (12)(12)$ and that $(ab) = (ca)(cb)(ca)$.

▷ **Quick Exercise.** Why does this mean that any permutation can be expressed as a product of transpositions in infinitely many ways? ◁

However, it does turn out that *any* factorization for the permutation (1345)(267) will have an odd number of factors. In fact, for any permutation, all the factorizations of that permutation into transpositions involve an even number of transpositions, or they all involve an odd number of transpositions. To see that the number of factors is always even or always odd, consider the polynomial

$$g_n = (x_1 - x_2)(x_1 - x_3)\cdots(x_1 - x_n)(x_2 - x_3)\cdots(x_{n-1} - x_n)$$
$$= \prod_{i<j}(x_i - x_j).$$

Notice that this polynomial consists of the product of all terms of the form $x_i - x_j$, where $i < j$, and so each pair x_i, x_j appears together in a factor of this product exactly once. Now consider a permutation of the n terms $x_1, \ldots, x_n$. Any such permutation will either map g_n to g_n or g_n to $-g_n$. Now clearly a single transposition maps g_n to $-g_n$. So, the product of an even number of transpositions will map g_n to g_n and the product of an odd number of transpositions will map g_n to $-g_n$. Thus, if a particular permutation of n elements leaves g_n unchanged, then clearly *any* representation of this permutation as a product of transpositions must do the same, and so such a product must have an even number of tranpositions. Similarly, if the permutation changes g_n to $-g_n$, any representation of this permutation as a product of transpositions must contain an odd number of transpositions.

So, we can now call a permutation **even** if it can be expressed as a product of an even number of transpositions, and **odd** if it can be expressed as a product of an odd number of tranpositions. Saying whether a permutation is even or odd is to specify its **parity.**

▷ **Quick Exercise.** Determine whether the following permutations are even or odd: ι, $(123)(4789)$, $(123)(342)$. ◁

34.3 The Alternating Groups

Let A_n be the set of even permutations. Take two such permutations, and multiply them together. Any representation of this product as a product of transpositions will have an even number of transpositions. Thus, the product of two even permutations is even. That is, A_n is closed under multiplication. But what about inverses? Given a permutation represented as a product of transpositions, its inverse can be computed as the product of the same transpositions, in the opposite order!

▷ **Quick Exercise.** Why is the inverse of a product of transpositions the same product in the opposite order? ◁

Thus, the inverse of an even transposition is still even. This all means that A_n is a subgroup of S_n. We call A_n the **alternating group.**

Note that the argument we have just given does *not* mean that the set of odd permutations forms a subgroup: The sum of two odd integers is even, and so the product of two odd permutations is an even permutation. And there's an even simpler reason why the odd permutations cannot form a subgroup: Any subgroup must contain the identity, and the identity is an even permutation (because 0 is even).

Note that cycles of *odd* length are the *even* permutations!

▷ **Quick Exercise.** Why are cycles of odd length even permutations? ◁

Example 34.3

We have looked at the group A_4 before (see Example 32.5). It is isomorphic to the group of symmetries of the tetrahedron and consists of the twelve permutations

$$\iota,\ (123),\ (132),\ (124),\ (142),\ (234),\ (243),$$

$$(134),\ (143),\ (12)(34),\ (13)(24),\ (14)(23).$$

▷ **Quick Exercise.** List the elements in the group A_3. ◁

In the examples we've looked at, it certainly appears that exactly half the permutations are even, which would mean that the group A_n has exactly $n!/2$ elements (where $n > 1$, to avoid the trivial case A_1). This is quite easy to prove. Consider the function

$$\Lambda : A_n \to S_n \backslash A_n$$

defined by $\Lambda(\alpha) = (12)\alpha$. That is, Λ merely multiplies each element by the transposition (12). If α is even, quite evidently $(12)\alpha$ is odd, and so this function is well defined. But Λ is also one-to-one and onto. (See Exercise 34.2.) Thus, there are as many even permutations as odd.

34.4 The Alternating Subgroup is Normal

It is now easy to see that the alternating group A_n is a normal subgroup of S_n; the argument we've just done shows that $[S_n : A_n] = 2$, and so the result follows from the Index 2 Theorem 32.3. We will however provide here another proof, which will give us added insight into the arithmetic of permutations.

By Theorem 32.2, to show that A_n is normal, we must check that for any permutation α,

$$\alpha^{-1} A_n \alpha \subseteq A_n.$$

In other words, we must show that if β is even, then $\alpha^{-1}\beta\alpha$ is too.

We introduce some general terminology and notation to deal with this. Let G be a group and $g, h \in G$. We call the element $g^{-1}hg$ the **conjugate** of h by g. Our goal then can be rephrased as this: We must show that *conjugation preserves the parity of permutations.* (Note that in Exercise 32.15 you proved that conjugates have the same order.) But in fact, much more is true:

Theorem 34.2 Conjugation Theorem *Conjugation in S_n preserves disjoint cycle structure. That is, if we express an element α as a product of m disjoint cycles of lengths $k_1, k_2, \cdots, k_m$, then any conjugate of α can be expressed as a product of m disjoint cycles, of lengths $k_1, k_2, \cdots, k_m$.*

Before proceeding to the proof of this theorem, let's look at a couple of examples.

Example 34.4

Consider the element $\alpha = (12)(3456)(789)$, the product of a 2-cycle, a 3-cycle and a 4-cycle. Consider some other permutation, like $\beta = (14567)(239)$. Then $\beta^{-1} = (76541)(932)$, and

$$\beta^{-1}\alpha\beta = \beta^{-1}(153724689) = (1452)(368)(79).$$

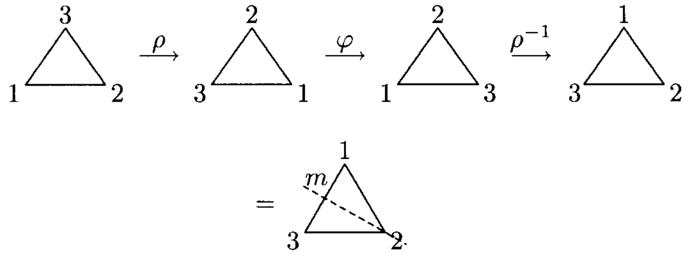
This new element has exactly the same disjoint cycle structure as α.

▷ **Quick Exercise.** Now pick some other permutation β at random, and use it to conjugate α. Is the disjoint cycle structure preserved? ◁

Example 34.5

In Chapter 22 we looked at S_3 as the set of symmetries of the equilateral triangle. Let's consider what conjugation might mean geometrically in this context, and thus understand why we should expect it to preserve disjoint cycle structure.

In particular, consider the symmetry φ, which is represented by the permutation (12). Recall that φ is the reflection of the triangle through the vertical line ℓ. Let's see what the conjugate $\rho^{-1}\varphi\rho$ of φ by ρ means, where ρ rotates the triangle 120° counterclockwise. The rotation ρ re-orients the triangle, so that the flip φ will permute the vertices at original positions 1 and 2. We then rotate back via ρ^{-1}. We have in effect accomplished the flip $\rho\varphi$ through the line m. (See the diagram below.) In other words, $\rho^{-1}\varphi\rho = (13)$.



A metaphor for the interpretation of conjugation in this example might be illustrative. Suppose you, an English-speaking mathematics student, are asked to solve a mathematical problem written in French! Your thought process would probably be this: Translate the problem into English (operation β), solve the problem (operation α), and then translate back into French (β^{-1}). The general message is this: if a group element can be thought of an operation or process, then any conjugate of it will be an operation or process of a similar type: the conjugate of a 3-cycle is a 3-cycle; the conjugate of a flip is a flip. (See Exercise 34.g.)

This discussion suggests the general method of proof for our theorem:

Proof of the Conjugation Theorem: We prove the theorem first in the case where α is a cycle of length k. So suppose that $\alpha = (a_1 a_2 \cdots a_k)$, and β is an arbitrary permutation. Then we claim that $\beta^{-1}\alpha\beta$ is the k-cycle

$$\chi = (\beta^{-1}a_1 \; \beta^{-1}a_2 \; \cdots \; \beta^{-1}a_k).$$

By direct computation it is easy to see that the conjugate $\beta^{-1}\alpha\beta$ and the k-cycle χ behave in the same way on the set

$$\{\beta^{-1}a_1, \beta^{-1}a_2, \cdots, \beta^{-1}a_k\}.$$

▷ **Quick Exercise.** Check this. ◁

Now suppose m is some integer not belonging to this set; that is, $\chi(m) = m$. Then βm is not in the set $\{a_1, \cdots, a_k\}$. That is, βm is not in the support of α. Thus,

$$\beta^{-1}\alpha\beta(m) = \beta^{-1}\beta(m) = m = \chi(m).$$

We have just shown that $\chi = \beta^{-1}\alpha\beta$, and so $\beta^{-1}\alpha\beta$ is also a cycle of length k.

Now suppose that α has been factored as a product

$$\chi_1\chi_2\cdots\chi_n$$

of disjoint cycles. Note that

$$\beta^{-1}\alpha\beta =$$

$$\beta^{-1}\chi_1\beta\beta^{-1}\chi_2\beta\cdots\beta^{-1}\chi_n\beta,$$

and so the conjugate of α is evidently a product of cycles of the same length as the corresponding cycles making up α. It remains only to show that these cycles are disjoint. But if $\beta^{-1}\chi_i\beta$ and $\beta^{-1}\chi_j\beta$ both move r, then χ_i and χ_j would both move $\beta^{-1}(r)$, which is impossible, since χ_i and χ_j are disjoint. □

▷ **Quick Exercise.** Check this theorem for three distinct conjugates of the permutation $(13)(54789)$. ◁

Now it is easy to see (again) that A_n is a normal subgroup of S_n:

Corollary 34.3 *The alternating group is a normal subgroup of the symmetric group.*

Proof: A conjugate of an even permutation has the same disjoint cycle structure. But this means that the conjugate is even, too. □

Example 34.6

> As another example of the conjugation theorem, consider again the subgroup
>
> $$\{\iota, (12)(34), (13)(24), (14)(23)\}$$
>
> of the alternating group A_4. (See Example 34.3 for the elements in A_4.) Any conjugate of a non-identity element of this group has the same disjoint cycle structure (a product of two disjoint transpositions). But the subgroup consists of the complete collection of elements of this form (together with ι). We first looked at this in Example 32.5 where we found it difficult to prove completely the normality of this subgroup. In Example 33.5 you showed that it is normal by exhibiting a homomorphism, of which it is the kernel.

34.5 Simple Groups

As we've seen already, A_n is always normal in S_n, for all positive integers n. Does A_n have any normal subgroups? The answer for $n = 4$ is certainly yes (as we've seen Example 34.6). But it is possible to prove that for $n \geq 5$, A_n itself has no proper normal subgroups (except the trivial subgroup). This proof is quite a bit more technical, but similar in flavor, to the proof of the Conjugation Theorem 34.2. Note that these groups A_n have many non-trivial proper subgroups (just consider any cyclic subgroup). However, no such subgroup is normal.

This situation merits a definition: We call a group with no proper normal subgroups (except the trivial subgroup) a **simple** group. The following theorem gives a large collection of simple groups.

Theorem 34.4 *The groups A_n are simple, for $n \neq 4$.*

Proof: For $n < 4$, showing that A_n is simple is left as Exercise 34.4. We will now prove that A_n is simple, for $n \geq 5$. For that purpose, suppose that N is a non-trivial normal subgroup of A_5. Choose a non-identity element $\alpha \in N$ whose support is of minimal size. We wish to show that α is a 3-cycle.

Note first that in its disjoint cycle representation, all factors of α must be cycles of the same size. For if $\alpha = (a_1 \cdots a_k)(b_1 \cdots b_m)\cdots$, where $k < m$, then $\alpha^k \in N$, but α^k leaves the integers a_i fixed, and so has a smaller support than α.

Suppose next that α is a product of disjoint 2-cycles (that is, transpositions). So, we may suppose (without loss of generality) that $\alpha = (12)(34)\cdots$, where the remainder of α consists of further disjoint transpositions. Because N is normal, we then have that $\beta = (123)\alpha(123)^{-1}\alpha^{-1} \in N$. But $\beta = (13)(24)$, and so because α has a support on minimum size in N, we must have that $\alpha = (12)(34)$. Now $n > 4$, and so $(125) \in A_n$. But then $(125)(12)(34)(125)^{-1}((12)(34))^{-1} \in N$. But this latter element is (152), and this contradicts the minimality of the support of α.

We now may suppose that α is a product of disjoint k-cycles, where $k \geq 3$. Without loss of generality we then have that

$$\alpha = (123\cdots)(456\cdots)\cdots.$$

But then $(12534) = (124)\alpha(124)^{-1}\alpha^{-1} \in N$, which again contradicts the minimality of the support of α.

At this point we have concluded that α must be a single odd k cycle, with $k \geq 3$. Now if $k > 3$, we may suppose that $\alpha = (12345\cdots)$. But then $(124) = (123)\alpha(123)^{-1}\alpha^{-1} \in N$. This yet again contradicts the minimality of the support of α, and so we conclude that α itself must be a 3-cycle.

We may as well suppose that $\alpha = (123)$. But for any other 3-cycle (abc), We can construct a permutation τ by first setting $\tau(a) = 1, \tau(b) = 2, \tau(c) = 3$, and then filling in the other values of τ in such a way that we have a permutation. It is then clear that $\tau^{-1}(123)\tau = (abc)$. By normality, this would say that $(abc) \in N$, if $\tau \in A_n$. However, as we have constructed it τ may not be an even permutation. However, if the permutation τ we have first constructed is odd, the permutation $\tau(45)$ will certainly belong to A_n, and still take (123) to (abc) by conjugation. Of course, this last step is permissible only because $n \geq 5$.

We have thus shown that if N is a non-trivial normal subgroup of A_n, it contains all 3-cycles. But by Exercise 34.12, the set of 3-cycles generates A_n, and so $N = A_n$, as we claimed. □

It turns out that this theorem will be of considerable importance to us in Chapter 49, when we prove that polynomial equations with degree greater than 4 cannot always be solved by ordinary arithmetic and root extractions.

There are other simple groups that are very familiar to us:

Example 34.7

> Consider the cyclic group $\mathbb{Z}_5$. This group has no non-trivial proper subgroups. If H is any non-trivial subgroup, choose a non-identity element $h \in H$. By Lagrange's Theorem 31.2, this non-identity element has order 5, and so $\langle h \rangle = \mathbb{Z}_5$. Thus, $\mathbb{Z}_5$ is simple.

The argument of this example can clearly be extended, and so we have:

Theorem 34.5 *The groups $\mathbb{Z}_p$ are simple, for all primes p.*

▷ **Quick Exercise.** Prove this by repeating the argument above for $\mathbb{Z}_p$, p prime. ◁

Historical Remarks

In the theory of finite groups that has been developed in the 20th century, simple groups can be viewed as the building blocks out of which more complicated groups can be constructed. Thus, to understand all finite groups, we need to understand simple groups, and how they can be put together to form more complicated groups. Both questions are very difficult.

One of the triumphs of 20th-century mathematics has been the complete classification of all finite simple groups. Dozens of mathematicians have contributed to the solution of this problem, and a full proof would require many hundreds of pages of technical mathematics. We have met two families of finite simple groups—namely, the cyclic groups of prime order, and the alternating groups—and there are other such families. But part of the difficulty in the classification theorem lies in the fact that there are some finite simple groups that do not belong to such naturally defined families. These simple groups are called *sporadic*, and the last of these 26 groups was finally constructed in the 1980s.

The problem of putting together simple groups (called the *extension problem*) is far from solved. The easiest way to put together smaller groups to build more complicated ones is the idea of direct product, which we understand well. But there are much more complicated ways. We will turn our attention to how abelian groups can be put together in certain ways to form a much larger class called *solvable* groups, in Chapter 36.

The general program for understanding finite groups is typical in much of algebra. We wish to prove a *structure theorem*: All objects of a given type can be built in a specified way, from well-understood pieces. In the next chapter we will discuss such a program for finite abelian groups. In this case, the well-understood pieces turn out to be cyclic groups, and the method of putting them together is the idea of direct product. In Chapter 46, we will provide a structure theorem for finite fields.

Chapter Summary

In this chapter we proved that the set of even permutations forms a subgroup of the symmetric group, called the *alternating group*. We proved the *Conjugation Theorem*, which says that conjugation preserves disjoint cycle structure. It follows that A_n is normal in S_n.

Warm-up Exercises

a. Compute explicitly three conjugates of the permutation

$$(145)(96)(237),$$

checking that its disjoint cycle structure is preserved.

b. How many elements are in A_5? A_6?

c. Because conjugation preserves parity, the conjugate of an odd permutation is odd. Why doesn't that make the set of odd permutations a normal subgroup?

d. Why can't a simple group have a non-trivial subgroup of index 2?

e. Give an example of a non-normal subgroup, whose index is 3. (Thus, Theorem 32.3 is false if 2 is replaced by 3.)

f. Is the identity permutation even or odd?

g. Suppose the operation β means "paint the north wall of your room", and the operation α means "move to your sister's room". What is the meaning of the conjugate $\alpha^{-1}\beta\alpha$? Remember that functional composition is read from right to left.

Exercises

1. Is the product of an even permutation and an odd permutation always an odd permutation? Prove this, or give a counterexample.

2. Show that the map $\Lambda : A_n \to S_n \backslash A_n$ defined by $\Lambda(\alpha) = (12)\alpha$ is one-to-one and onto, completing the argument of Section 34.3 that there are as many even as odd permutations. (The notation $S_n \backslash A_n$ means the elements of S_n that are not in A_n.)

3. (a) List the elements of A_5. (See Exercise b.)

   (b) List all cyclic subgroups of A_5.

   (c) Let $\alpha \in A_5$. Show $\langle \alpha \rangle$ is not a normal subgroup of A_5.

4. Show that A_n is simple for $n < 4$, proving part of Theorem 34.4.

5. By following the proof of the Conjugation Theorem 34.2, demonstrate explicitly that

$$(1456)(29) \quad \text{and} \quad (7895)(13)$$

   are conjugates in S_9.

6. Let $n \geq 5$. Prove that A_n is the only non-trivial normal subgroup of S_n.

7. Prove that any permutation can be expressed as a product of the transpositions $(12), (13), (14), \cdots, (1n)$.

8. Prove that any permutation can be expressed as a product of the transpositions $(12), (23), (34), \cdots, (n-1\ n)$.

9. Prove that S_n is generated by the two elements $(1234\cdots n)$ and (12).

10. In Example 31.6 we showed that A_4 has no subgroup with 6 elements. Provide a different proof for this fact: Assume that H is such a subgroup; then H must be normal. (Why?) Now count the three cycles. Show that at least one of the three cycles must be in H. Then argue that H must contain all the three cycles, which is absurd.

11. Let $n \geq 2$ and let H be a subgroup of S_n. Prove either that all elements of H are even, or else exactly half of the elements of H are even.

12. Let n be a positive integer, $n \geq 3$. Prove that A_n is generated by its 3-cycles. That is, prove that every element of A_n can be expressed as a product of 3-cycles.

13. Suppose that H is a subgroup of the permutation group S_n. We say that H is a **transitive subgroup** if for any integers k, j with $1 \leq k, j \leq n$, we can find $\alpha \in H$ so that $\alpha(k) = j$.

    (a) Determine all transitive subgroups of S_4.

    (b) Argue that A_n is a transitive subgroup of S_n, for any $n > 2$.

(c) Suppose that H is a transitive subgroup of S_n, and $\alpha \in S_n$. Prove that the conjugate subgroup $\alpha^{-1}H\alpha$ is also a transitive subgroup of S_n. (See Exercise 32.6 for more about conjugate subgroups.)

Chapter 35

Fundamental Theorem for Finite Abelian Groups

In this chapter we state the *Fundamental Theorem for Finite Abelian Groups*, which completely describes the structure of such groups. This powerful theorem provides an easy-to-understand recipe by which all such groups can be constructed, using the two familiar notions of *cyclic group* and *direct product*. The theorem is relatively difficult to prove, and so we will not prove it here. The interested reader can refer to any introductory text in group theory. However, we will look more closely at a very special type of finite group called a p-group and prove a couple of important facts regarding them.

35.1 The Fundamental Theorem

Theorem 35.1 The Fundamental Theorem for Finite Abelian Groups *Every finite abelian group is isomorphic to a direct product of cyclic groups; each cyclic group in this decomposition is of order p^n, where p is prime. That is, each finite abelian group is isomorphic to a group of the form*

$$\mathbb{Z}_{p_1^{k_1}} \times \mathbb{Z}_{p_2^{k_2}} \times \cdots \times \mathbb{Z}_{p_n^{k_n}}$$

where the p_i's are primes (not necessarily distinct), and the k_i's are positive integers (not necessarily distinct).

Note that the group

$$\mathbb{Z}_{p_1^{k_1}} \times \mathbb{Z}_{p_2^{k_2}} \times \cdots \times \mathbb{Z}_{p_n^{k_n}}$$

has order $p_1^{k_1} p_2^{k_2} \cdots p_n^{k_n}$.

Example 35.1

What finite abelian groups of order 8 are possible? Clearly, the cyclic groups used to build such groups can only have order 2, 4, or 8, and the product of the orders of the cyclic groups must be 8. So the only possible abelian groups of order 8 are

$$\mathbb{Z}_8, \quad \mathbb{Z}_4 \times \mathbb{Z}_2, \quad \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2.$$

▷ **Quick Exercise.** We did not include $\mathbb{Z}_2 \times \mathbb{Z}_4$, because it is isomorphic to $\mathbb{Z}_4 \times \mathbb{Z}_2$. Give an isomorphism. ◁

Are these three groups really non-isomorphic? We can answer this affirmatively by looking at the orders of elements in these groups. Of them, only $\mathbb{Z}_8$ has any elements of order 8; $\mathbb{Z}_4 \times \mathbb{Z}_2$ has elements of order 4, while all non-identity elements of $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$ are of order 2.

▷ **Quick Exercise.** Verify our assertions in the last sentence, thus checking that these groups are not isomorphic. ◁

Note that in Exercise 31.9 you actually determined a complete list of non-isomorphic groups of order 8, whether abelian or not.

Example 35.2

What finite abelian groups of order 10 are possible? By the Fundamental Theorem for Finite Abelian Groups 35.1, since $10 = 2 \cdot 5$ (a product of primes), there is only one possibility: $\mathbb{Z}_2 \times \mathbb{Z}_5$. However, you might notice that $\mathbb{Z}_{10}$ is also abelian and of order 10. But these two groups are isomorphic. This is easy to see, because the element $(1,1) \in \mathbb{Z}_2 \times \mathbb{Z}_5$ has order 10, and so this group is cyclic of order 10.

More generally, these two groups are necessarily isomorphic, on account of Theorem 21.2. Of course, that theorem is phrased as a theorem about rings, but any cyclic group of the form $\mathbb{Z}_n$ can be equipped with a ring structure, and so we can use it here.

▷ **Quick Exercise.** Review Theorem 21.2. ◁

Thus, there is (up to isomorphism) only one abelian group of order 10.

▷ **Quick Exercise.** How many distinct abelian groups of order 9 are there? How about order 12? ◁

By Theorem 21.2, if p and q are two *distinct* primes, then $\mathbb{Z}_{p^k} \times \mathbb{Z}_{q^j}$ is isomorphic to $\mathbb{Z}_{p^k q^j}$. (See Exercise 35.1.)

As a bonus, the Fundamental Theorem for Finite Abelian Groups shows that a finite abelian group has many subgroups. In fact, it has as many subgroups as it possibly could have:

Corollary 35.2 *If G is a finite abelian group of order r and m divides r, then G has a subgroup of order m.*

Lagrange's Theorem 31.2 asserts that the order of a subgroup of a finite group divides the order of the group; this corollary is the converse of this, for the abelian case. However, this converse is not true in general, as we saw in Example 31.6.

Proof: By the Fundamental Theorem,

$$G = \mathbb{Z}_{p_1^{k_1}} \times \mathbb{Z}_{p_2^{k_2}} \times \cdots \times \mathbb{Z}_{p_n^{k_n}},$$

where $p_1^{k_1} p_2^{k_2} \cdots p_n^{k_n} = r$. Because m divides r, $m = p_1^{j_1} p_2^{j_2} \cdots p_n^{j_n}$, where $j_i \leq k_i$, for all i. We need only show that each $\mathbb{Z}_{p_i^{k_i}}$ has a cyclic subgroup of order $p_i^{j_i}$ and then

$$H = \mathbb{Z}_{p_1^{j_1}} \times \mathbb{Z}_{p_2^{j_2}} \times \cdots \times \mathbb{Z}_{p_n^{j_n}},$$

is the desired subgroup. But this follows from Exercise 26.10. □

Example 35.3

Consider the group $G = \mathbb{Z}_8 \times \mathbb{Z}_{16} \times \mathbb{Z}_9$, which is of order 1152. To find a subgroup of order 24, we must combine cyclic subgroups of the three groups $\mathbb{Z}_8, \mathbb{Z}_{16}$, and $\mathbb{Z}_9$ whose orders multiply together to give 24. For example, we could find a subgroup isomorphic to $\mathbb{Z}_8 \times \{0\} \times \mathbb{Z}_3$, or else $\mathbb{Z}_4 \times \mathbb{Z}_2 \times \mathbb{Z}_3$, together with some other possibilities. Note that the two subgroups just given are not isomorphic.

▷ **Quick Exercise.** Why are the two subgroups above not isomorphic? ◁

Example 35.4

The method suggested still works, even if the direct product of cyclic groups is not that given by the Fundamental Theorem. For example, consider the group $G = \mathbb{Z}_{12} \times \mathbb{Z}_{10}$, which is of order 120. If we wish to find a subgroup of G of order 8, we can still simply piece together subgroups of the cyclic groups. Clearly, $\mathbb{Z}_4 \times \mathbb{Z}_2$ is such a subgroup.

35.2 p-groups

Now let's look more closely at a particular type of finite group called a p-group. A **p-group** is a group, each of whose elements has order some power of the prime number p.

Example 35.5

Consider the group $\mathbb{Z}_{27}$. By Lagrange's Theorem 31.2, the order of every element divides 27, and so must be a power of 3. Thus, $\mathbb{Z}_{27}$ is a 3-group.

Example 35.6

Consider the group $\mathbb{Z}_2 \times \mathbb{Z}_8$. This is a 2-group.

The concept of p-group makes sense for non-abelian groups. Although we will have no more use for such examples in the rest of this chapter, we do provide one such example here:

Example 35.7

The group of symmetries of the square is a 2-group.

Theorem 31.5 tells us that if p is a positive prime integer where p divides the order of G, then G has an element of order p. (This is true whether G is abelian or not.)

It is important to note that we cannot omit the hypothesis that p is prime. For if G is any non-cyclic group with n elements, then n divides

$|G|$, while G has no element of order n. For example, $\mathbb{Z}_4 \times \mathbb{Z}_2$ has no element of order 8.

We now use this theorem to prove the following easy corollary about p-groups:

Corollary 35.3 *Let G be a finite abelian p-group. Then $|G| = p^n$, for some positive integer n.*

Proof: Suppose instead that $|G|$ is divisible by some prime q other than p. But by Theorem 31.5, G must have an element of order q, which is impossible, if G is to be a p-group. □

Chapter Summary

In this chapter we stated the *Fundamental Theorem for Finite Abelian Groups*. This theorem completely describes the structure of such groups, as direct products of cyclic p-groups. As a corollary, we showed that a finite abelian group has a subgroup of order m for all m that divide the order of the group. We also showed that the order of every p-group is a power of p.

Warm-up Exercises

a. How many distinct abelian groups are there of the following orders: 14, 18, 25, 29?

b. Describe the following finite abelian groups as direct products as specified by the Fundamental Theorem for Finite Abelian Groups 35.1:

$$\mathbb{Z}_4 \times \mathbb{Z}_6, \quad \mathbb{Z}_{150}, \quad U(\mathbb{Z}_{20}), \quad U(\mathbb{Z}_{19}), \quad U(\mathbb{Z}_{21}).$$

c. Determine whether the following groups are p-groups; if so, for which p?

$$Q_8, \quad A_4, \quad \mathbb{Z}_{27}, \quad \mathbb{R}, \quad A_3, \quad D_8.$$

d. A finite abelian group has 50 elements. Why do we know it is *not* a p-group?

e. Give examples where the converse of Lagrange's Theorem 31.2 fails. That is, give a group G and an integer n, where n divides $|G|$, but G does not possess an element of order n. Find examples other than those mentioned in the text.

f. Up to isomorphism, how many abelian groups are there of order pq where p and q are distinct primes?

g. Up to isomorphism, how many abelian groups are there of order p^2, for prime p?

Exercises

1. Show that if p and q are distinct primes, that $\mathbb{Z}_{p^k} \times \mathbb{Z}_{q^j}$ is isomorphic to $\mathbb{Z}_{p^k q^j}$ by finding an element of $\mathbb{Z}_{p^k} \times \mathbb{Z}_{q^j}$ of order $p^k q^j$. (Of course, this result follows from Theorem 21.2, but we want you to actually find a generator.)

2. Show that $\mathbb{Z}_{12} \times \mathbb{Z}_{10}$ has no element of order 8 and, hence, no subgroup isomorphic to $\mathbb{Z}_8$.

3. Suppose G is a finite abelian group of order m, where m is square-free. (That is, if p divides m, then p^2 does not.) Show that G is cyclic.

4. If there are k abelian groups of order m and j abelian groups of order n, how many abelian groups are there of order mn if m and n are relatively prime?

5. Suppose that G, H, and K are finite abelian groups. Suppose that the direct products $G \times K$ and $H \times K$ are isomorphic as groups. Prove that G and H are isomorphic.

6. The conclusion of Exercise 5 seems so plausible that you might conjecture that it is true for all groups (or at least for all abelian groups). Construct a counterexample to this, using abelian groups.

Chapter 36

Solvable Groups

In this chapter we will introduce solvable groups, as a natural generalization of abelian groups. Solvable groups can be viewed as groups built out of finitely many abelian pieces. This gives us a chance to see a bit of how the extension problem in groups is approached; we discussed these ideas briefly in the historical note at the end of Chapter 34. We will use the notion of solvability in Chapter 49, when we discuss whether a polynomial equation can be solved using ordinary arithmetic and the extraction of roots.

36.1 Solvability

The symmetry group S_3 is not abelian, but we can think of it as consisting of two abelian pieces, put together in the appropriate way. Namely, S_3 has an abelian subroup A_3, and the homomorphic image S_3/A_3 is also abelian. We generalize this idea to finitely many steps in our definition. We say that a group G is **solvable** if it has a finite collection of subgroups $G_0, G_1, \cdots G_n$, so that

$$G_n = \{1\} \subseteq G_{n-1} \subseteq G_{n-2} \subseteq \cdots \subseteq G_1 \subseteq G_0 = G,$$

and furthermore each G_{i+1} is normal in G_i, and the group of cosets G_i/G_{i+1} is abelian. We shall call such a finite sequence of subgroups satisfying the definition of solvability a **subnormal series with abelian quotients** for the group G.

The terminology we have chosen for the series required for a solvable group is cumbersome, but it is consistent with the mathematical literature. In general, a subnormal series requires only that each G_{i+1} is normal in G_i, without any requirement on the group G_i/G_{i+1}. The word 'subnormal' places emphasis on the fact that we require only that

G_{i+1} is normal in the next larger subgroup G_i; it is not necessarily normal in the entire group G.

Example 36.1

We note first that any abelian group G is of course solvable because $\{1\} \subseteq G$ is a subnormal series with abelian quotients.

Example 36.2

The group S_3 is solvable because $\{\iota\} \subseteq A_3 \subseteq S_3$, and S_3/A_3 and $A_3/\{\iota\} = A_3$ are abelian. Now we can also think of S_3 as the symmetries of an equilateral triangle, that is, as the dihedral group D_3. We can now generalize this example to see that any of the dihedral groups D_n are solvable, because the subgroup of rotations is abelian, and of index two in D_n.

▷ **Quick Exercise.** Check this. ◁

Example 36.3

The group S_4 is solvable, although it will require more than two subgroups to build a subnormal series with abelian quotients. We will use the notation from Example 32.5, where we denote by H a subgroup of S_4 isomorphic to the Klein Four Group:

$$\{\iota\} \subseteq H \subseteq A_4 \subseteq S_4.$$

Each of these subgroups is normal in the next bigger subgroup. The quotient groups are abelian because they are isomorphic to $H, \mathbb{Z}_3$ and $\mathbb{Z}_2$, respectively.

▷ **Quick Exercise.** Check these assertions. ◁

Example 36.4

The group S_5 is not solvable. By Exercise 34.6 we know that A_5 is the only non-trivial normal subgroup of S_5. But by Theorem 34.4, A_5 is a non-abelian group with no non-trivial normal subgroups, and so S_5 cannot have a solvable series. The fact that S_5 is not solvable will be important for us in Chapter 49.

36.2 New Solvable Groups from Old

In this section we will present some natural ways to get from solvable groups from a given solvable group.

Theorem 36.1 *Every homomorphic image of a solvable group is solvable.*

Proof: Suppose that G is a solvable group with normal subgroup H. We shall assume that G has the following subnormal series with abelian quotients:

$$G_n = \{1\} \subseteq G_{n-1} \subseteq G_{n-2} \subseteq \cdots \subseteq G_1 \subseteq G_0 = G.$$

Consider each of the sets $HG_i = \{hg : h \in H, g \in G_i\}$. By Exercise 32.13a we know that HG_i is a subgroup of HG_{i-1}. We claim that HG_i is normal in HG_{i-1}. To prove this, choose $h \in H$ and $a \in G_{i-1}$. The right coset $HG_iha = HG_ia$, because $H \subseteq HG_i$. But then

$$HG_iha = HG_ia = H(G_ia) = H(aG_i) = (Ha)G_i = (aH)G_i = aHG_i,$$

because G_i is normal in G_{i-1} and H is normal in G. But $aHG_i = aa^{-1}haHG_i = haHG_i$, because $a^{-1}ha \in H$. This equality of right and left cosets means that HG_i is normal in HG_{i-1}.

By Exercise 33.12 we have for each i that HG_i/H is a normal subgroup of HG_{i-1}/H, and so we have a subnormal series for G/H:

$$H/H \subseteq HG_{n-1}/H \subseteq HG_{n-2}/H \subseteq \cdots \subseteq HG_1/H \subseteq G/H.$$

Now by Exercise 33.13

$$(HG_{i-1}/H)/(HG_i/H)$$

is isomorphic to HG_{i-1}/HG_i.

It remains only to show that the latter group is abelian to conclude that G/H is solvable. For that purpose, suppose that $h, k \in H$ and $a, b \in G_{i-1}$. Then we can use normality and the fact that G_{i-1}/G_i is abelian to conclude the following:

$$\begin{aligned} HG_i(ha)(kb) &= HG_iakb = HG_i(aka^{-1})ab = HG_iab \\ &= H(G_iab) = H(G_iba) = HG_iba \\ &= HG_i(bhb^{-1})ba = HG_ibha = HG_i(kb)(ha). \end{aligned}$$

□

A similar proof shows the following:

Theorem 36.2 *Every subgroup of a solvable group is solvable.*

Proof: We leave this as Exercise 36.2. □

We can also build larger solvable groups, by putting together a normal subgroup and the corresponding homomorphic image:

Theorem 36.3 *Suppose that G is a group with normal subgroup H. Then G is solvable if and only if G/H is solvable and H is solvable.*

Proof: If G is solvable, the conclusions about H and G/H are just the previous two theorems.

Now suppose that H and G/H are solvable. Then we have subnormal series with abelian quotients for both these groups. By Exercise 33.12, we can write the subgroups of G/H in the form G_i/H:

$$H_m = \{1\} \subseteq H_{m-1} \subseteq \cdots \subseteq H_1 \subseteq H_0 = H$$

and

$$G_n/H = H/H \subseteq G_{n-1}/H \subseteq \cdots \subseteq G_1/H \subseteq G_0/H = G/H.$$

But then we can put together these subgroups to obtain a subnormal series with abelian quotients for G:

$$\begin{aligned} H_m = \{1\} &\subseteq H_{m-1} \subseteq \cdots \subseteq H_1 \subseteq H_0 = H \\ &\subseteq G_{n-1} \subseteq \cdots \subseteq G_1 \subseteq G_0 = G. \end{aligned}$$

▷ **Quick Exercise.** Check that the above series is subnormal with abelian quotients. Why is Exercise 33.12 relevant? ◁ □

We can informally paraphrase this theorem by asserting that a solvable extension of a solvable group is solvable. It is easy to extend this inductively to obtain the following:

Theorem 36.4 *A group G is solvable if and only if it has a subnormal series*

$$G_n = \{1\} \subseteq G_{n-1} \subseteq G_{n-2} \subseteq \cdots \subseteq G_1 \subseteq G_0 = G,$$

where each quotient group G_i/G_{i+1} is solvable.

Proof: If G is solvable, it has a subnormal series with abelian quotients, which are of course solvable.

For the converse, suppose that G has such a subnormal series. We will proceed by induction on n. If $n = 1$, then $G = G_0$ is obviously solvable. We now suppose that if G has such a subnormal series with length $n-1$, it is in fact solvable. Given the series above of length n, we can now conclude by induction that the group G_1 is in fact solvable. But then G_1 is a solvable normal subgroup of G, and by assumption G/G_1 is solvable. By Theorem 36.3 G is then solvable. □

Historical Remarks

We have barely scratched the surface here regarding the study of group extensions. The general project of understanding groups in terms of extensions built of simpler pieces is a large one that we cannot develop fully here. In the general theory, the 'simpler pieces' to use are precisely the simple groups, whose classification was such a major part of 20th century group theory — for more information you should consult the Historical Remarks following Chapter 34. This idea leads to subnormal series whose quotients are simple; these are called **composition series**. (In Exercise 36.3 you look at this in the specific context of finite solvable groups.) Any finite group has such a composition series (see Exercise 36.7). That the factor groups in such a series are unique (up to order) is an important theorem proved by Camille Jordan in the context of permutation groups, and in the general case by Otto Hölder.

One of the most important steps in the classification problem for finite simple groups was the difficult theorem of John Thompson and Walter Feit that all finite groups of odd order are in fact solvable! Consequently, we need look only to groups with even order in our search for finite simple groups. The Feit-Thompson theorem took up an entire issue of the Pacific Journal of Mathematics, when the theorem was first published in 1963.

Chapter Summary

In this chapter we explored the notion of *solvable group*. Such groups can be built as finitely many abelian extensions of an abelian group, and as such are a natural generalization of the abelian groups. We then saw how the class of solvable groups is closed under taking subgroups, homomorphic images, and building group extensions.

Warm-up Exercises

a. Is every abelian group solvable? If not, give an example of a non-solvable abelian group.

b. Is every solvable group abelian? If not, give an example of a non-abelian solvable group.

c. Show that the group of quaternions Q_8 is solvable. That is, give an appropriate subnormal series for this group.

Exercises

1. Suppose that G and H are solvable groups. Prove that $G \times H$ is solvable.
2. Prove Theorem 36.2.
3. Suppose the G is a finite abelian group. Prove that G has a composition series; that is, show that G has a collection of subgroups G_i so that $G_n = \{1\} \subseteq G_{n-1} \subseteq \cdots \subseteq G_1 \subseteq G_0 = G$, and G_{i-1}/G_i is a simple group.
4. Suppose that G is a finite solvable group. Prove that G has a composition series.
5. Prove that $\mathbb{Z}$ does not have a composition series.
6. Let G be a group and H a proper normal subgroup. Prove that G/H is a simple group if and only if H is a maximal normal subgroup.
7. Generalize Exercise 36.4. That is, prove that every finite group has a composition series.

Section VI in a Nutshell

This section starts by considering group homomorphisms, by way of analogy with ring homomorphisms: A function between groups $\varphi : G \to S$ is a *group homomorphism* if $\varphi(gh) = \varphi(g)\varphi(h)$ for every $g, h \in G$. A group homomorphism preserves the group identity and inverses, the image of G is a subgroup of S, and if G is abelian then so is $\varphi(G)$ (Theorem 28.1). If φ is one-to-one and onto, then we say φ is an *isomorphism*, in which case $\ker(\varphi) = 1_G$.

A subgroup H of G is *normal* if $gH = Hg$ for all $g \in G$; that is, if the left and right cosets of H are the same for each element of G. This is equivalent to saying that $g^{-1}Hg \subseteq H$ for all $g \in G$. If H is normal in G, then the collection of cosets of H in G, denoted G/H, forms a group under the operation $(Ha)(Hb) = Hab$ (Theorem 32.1). G/H is then called the *group of cosets of G mod H* or the *quotient group of G mod H.* Normal subgroups are to groups what ideals are to rings. The kernel $\ker(\varphi)$ of any group homomorphism is a normal subgroup of G.

Paralleling rings, there is the *Fundamental Isomorphism Theorem for Groups* (Theorem 33.4) which says that if $\varphi : G \to S$ is an onto homomorphism, then $G/\ker(\varphi)$ is isomorphic to S.

Whether the subgroup H is normal in G or not, the number of left cosets of H is the same as the number of right cosets. Assuming G is finite, this number is called the *index of H in G* and denoted by $[G : H]$. *Lagrange's Theorem* (Theorem 31.2) says that $|G| = [G : H]|H|$. Thus the order of a subgroup must divide the order of the group and so the order of any element must divide the order of the group. Thus, any group of prime order is cyclic.

An important group is the group of permutations on the set $\{1, 2, \ldots, n\}$, also called the nth *symmetric group* and denoted by S_n. Clearly $|S_n| = n!$. *Cayley's theorem* (Theorem 29.1) says that every group of order n is isomorphic to a subgroup of S_n.

A *transposition* is a cycle of length two. Any permutation can be factored as a product of transpositions (Theorem 34.1), and, while this factorization is not unique, all factorizations of a given permutation have the same parity. Thus we can classify a permutation as either even or odd. The set A_n of even permutations is a group called the nth *alternating group.* The alternating group A_n is a normal subgroup of the symmetric group S_n and $[S_n : A_n] = 2$. Furthermore, A_n is simple for $n \neq 4$.

All finite abelian groups can be completely described by the *Fundamental Theorem for Finite Abelian Groups* (Theorem 35.1): Every finite abelian group is isomorphic to a direct product of cyclic groups; each cyclic group in this decomposition is of order p^n, where p is prime. That is, each finite abelian group is isomorphic to a group of the form

$$\mathbb{Z}_{p_1^{k_1}} \times \mathbb{Z}_{p_2^{k_2}} \times \cdots \times \mathbb{Z}_{p_n^{k_n}}$$

where the p_i's are primes (not necessarily distinct), and the k_i's are positive integers (not necessarily distinct).

It follows from this theorem that if G is a finite abelian group of order r and m divides r, then G has a subgroup of order m (Corollary 35.2). That is, G has subgroups of every possible order.

A *p-group* is a group where each of its elements has order a power of the prime p. If G is a finite abelian p-group, then $|G| = p^n$, for some n (Corollary 35.3).

Finally, the idea of abelian group is generalized by the idea of *solvable* groups. A group is solvable if it has a finite collection of subgroups $G_0, G_1, \cdots G_n$, so that

$$G_n = \{1\} \subseteq G_{n-1} \subseteq G_{n-2} \subseteq \cdots \subseteq G_1 \subseteq G_0 = G,$$

and furthermore each G_{i+1} is normal in G_i, and the group of cosets G_i/G_{i+1} is abelian. Such a sequence of subgroups is called a *subnormal series with abelian quotients* for G. An important group that is *not* solvable is S_5 (Example 36.4). This important fact we will use in Section IX.

Every homomorphic image of a solvable group is also solvable (Theorem 36.1) as is every subgroup of a solvable group (Theorem 36.2). In fact, if G has a normal subgroup H, then G is solvable if and only if both H and G/H are solvable (Theorem 36.3).

VII

Constructibility Problems

Chapter 37

Constructions with Compass and Straightedge

You probably recall doing various constructions with a compass and a straightedge in high school geometry. We will imagine *idealized* tools. Thus, the straightedge is an unmarked ruler, which in principle can be as long as necessary. With it we can draw line segments of arbitrary length, perhaps passing through a particular point or connecting two given points. Likewise, the compass is as large as necessary; with it we can draw arcs and circles and duplicate distances. For instance, if A and B are marked on one line and point C is marked on another, the compass allows us to mark a point D on the second line so that the distance between C and D is the same as the distance between A and B. Of course, in actually carrying out these constructions with real rulers and compasses, there is always error involved. However, we are concerned with *idealized* constructions—perfect constructions with no error.

37.1 Construction Problems

From around the fifth century B.C. Greek mathematicians wondered what constructions could be carried out using only a compass and a straightedge. In the axioms that begin Euclid's *Elements* we discover the postulation of an idealized compass and straightedge. The theorems about plane geometry that follow show how successful the Greeks were at doing plane geometry with compass and straightedge. You should recall some of the constructions possible:

- find the midpoint of a line segment;
- construct a line perpendicular to a given line through a given

point on the line;

- construct a line perpendicular to a given line through a point off the line;
- construct a line parallel to a given line through a point off the given line;
- given an angle, bisect it.

▷ **Quick Exercise.** How are these constructions done? Better yet, get out your compass and straightedge and do them. ◁

There were, however, three famous constructions that ancient Greek mathematicians could *not* accomplish with compass and straightedge:

- **Doubling the cube:** Given a line segment, which represents the edge of a cube, construct another line segment representing the edge of another cube, whose volume is twice that of the original cube.
- **Trisecting an angle:** Given an arbitrary angle, divide it into three equal parts.
- **Squaring the circle:** Construct a square that has the same area as a given circle.

The question of whether such constructions are possible bedeviled mathematicians for over 2000 years.

It is important to note that in each case we desire a general method that works for *all* instances of the given problem. For example, the angle trisection problem is to find a method of trisection that works for all angles. *Certain* angles can be easily trisected, but this does not mean that the general problem has been solved. For example, a 90° angle can be trisected, because this is equivalent to constructing a 30° angle, which is easy to do. In fact, there are a number of ways to do this. For instance, you could construct an equilateral triangle and then bisect one of the angles. Or, you could directly construct a right triangle with angles of 60° and 30°.

▷ **Quick Exercise.** Construct an equilateral triangle. Also, directly construct a 30°–60°–90° triangle, by starting with a shorter leg of length 1. ◁

In the 19th century all three of the famous constructions above were shown to be impossible; the proofs were largely algebraic, rather than geometric. Think for a moment about what a proof of impossibility means. It is not at all clear how it is possible to show that a construction is *impossible*—we certainly can't try all possibilities! Proving the impossibility of these three classical construction problems was one of the great triumphs of mathematics in general and algebra in particular.

37.2 Constructible Lengths and Numbers

We start the attack on these problems by first deciding which lengths can be constructed. Specifically, we want to answer the following question: *Given a line segment in the plane, which we say is of length 1, for what values α can we construct a line segment of length α?*

Note that if we talk about lengths, we must start with some unit of measure; hence, we designate some particular line segment as the *unit line segment.* Our goal is to give a complete algebraic description of which numbers we can construct, when starting with a unit line segment. Our modest beginning is the observation that all the natural numbers can be constructed:

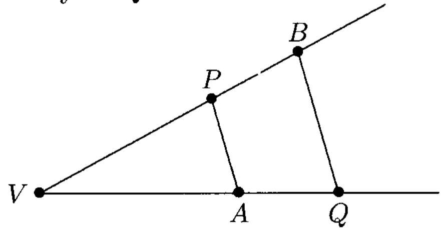
Lemma 37.1 *Given a line segment of length 1 and a natural number n, it is possible to construct a line segment of length n.*

Proof: Merely use the compass to lay off n copies of the unit line segment, next to one another on a line (which can be made as long as necessary, using the straightedge). □

Hence, we say that the natural numbers are constructible. In general, we say that the real number α is **constructible** if, given a line segment of length 1, it is possible to construct a line segment of length $|\alpha|$. Thus, the integers are constructible because the natural numbers are.

Theorem 37.2 *Given line segments of length 1, a, and b, it is possible to construct segments of lengths $a+b$, $a-b$ (if $a \geq b$), ab, and a/b.*

Proof: The constructibility of $a+b$ and $a-b$ are obvious. To construct ab, consider the figure below. Start by constructing two rays with vertex V. On one ray mark A so that the length of the line segment from V to A is a. (We will write $|\overline{VA}|$ for the length of the line segment from V to A.) Then on the other ray mark points P and B so that $|\overline{VP}| = 1$ and $|\overline{VB}| = b$. Now draw the line segment from P to A. Finally, construct a line parallel to this line segment and through the point B, as shown in the figure. Label the point where this line intersects the other ray as Q.



Now, $\triangle VAP$ is similar to $\triangle VQB$; thus,

$$\begin{aligned} \frac{|\overline{VA}|}{|\overline{VP}|} &= \frac{|\overline{VQ}|}{|\overline{VB}|}, \quad \text{or} \\ \frac{a}{1} &= \frac{|\overline{VQ}|}{b}, \quad \text{and so} \\ ab &= |\overline{VQ}|. \end{aligned}$$

A similar construction can be made for a/b. This is left as Exercise 37.1. □

The previous theorem asserts that the sum, difference, product, and quotient of two constructible numbers is constructible. This makes the next corollary obvious:

Corollary 37.3 *Given a line segment of length 1, the set of all constructible numbers is a field.*

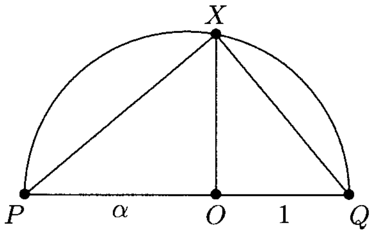
We will therefore refer to this set as the **field of constructible numbers** and denote it by $\mathbb{K}$ (the German word for 'construct' is 'konstruieren', and for 'field' is 'Körper'). Our lemma tells us that $\mathbb{Z}$ is a subring of $\mathbb{K}$. But $\mathbb{K}$ is closed under division, and so $\mathbb{K}$ contains all quotients of integers; that is, the field $\mathbb{Q}$ of rational numbers is a subfield of $\mathbb{K}$. Furthermore, by definition all constructible numbers must be real numbers. That is, $\mathbb{K}$ is a subfield of $\mathbb{R}$.

Are there constructible numbers other than the rationals? To answer this, construct a square with side one. Then draw the diagonal; this means that $\sqrt{2}$ is a constructible number! And we've seen before (Exercise 2.14) that $\sqrt{2}$ is an irrational number, and so not all constructible numbers are rational. We shall see, however, that not all real numbers are constructible. Indeed, the highlight of the next chapter is to exactly describe the field of constructible numbers. There, we shall see that constructing the square root is of utmost importance.

Accordingly, we generalize the fact that $\sqrt{2}$ is constructible, in the next theorem:

Theorem 37.4 *If α is constructible, then so is $\sqrt{|\alpha|}$.*

Proof: We assume that α is positive. Here is one way of constructing $\sqrt{\alpha}$. Refer to the figure below. First mark points P, O, and Q on a line so that $|\overline{PO}| = \alpha$ and $|\overline{OQ}| = 1$. Now find the midpoint of the line segment from P to Q and using that as the center, draw a semicircle of radius $(\alpha + 1)/2$, as shown. Draw a perpendicular to the line through the point O. Label the point where this perpendicular intersects the semicircle as X.



Note that $\triangle XOQ$ is similar to $\triangle POX$. Hence,

$$\begin{aligned} \frac{|\overline{PO}|}{|\overline{OX}|} &= \frac{|\overline{OX}|}{|\overline{OQ}|}, \quad \text{or} \\ \frac{\alpha}{|\overline{OX}|} &= \frac{|\overline{OX}|}{1}, \quad \text{and so} \\ \alpha &= |\overline{OX}|^2. \end{aligned}$$

That is, $|\overline{OX}| = \sqrt{\alpha}$. □

By repeatedly applying the last theorem, we can construct $\sqrt[k]{\alpha}$ if α is a positive constructible number and k is a power of two. So, for instance, $\sqrt{5}$, $\sqrt[4]{5}$, and $\sqrt[8]{5}$ are all constructible. Thus, there are infinitely many constructible numbers in addition to the rationals. Our goal in the next chapter is to characterize algebraically all elements of the field $\mathbb{K}$.

Although the constructions we have discussed in this chapter would have been familiar and understandable to an ancient Greek geometer, our emphasis on constructing *numbers*, rather than *line segments*, would have seemed strange. But we must change our point of view to bring to bear our modern algebraic tools.

Chapter Summary

We introduced the three famous construction problems of the ancient Greeks: Is it possible, using only a compass and a straightedge, to

- double the cube,
- trisect an angle, or
- square the circle?

We defined a *constructible number* and showed that the rationals are constructible. Indeed, we showed that the set $\mathbb{K}$ of constructible numbers is a field. We also showed that $\sqrt{|\alpha|}$ is constructible if α is.

Warm-up Exercises

a. Can you trisect the angle 45°?

b. Suppose you could square a particular circle. Could you then square any circle?

c. Did we discuss a real number in this chapter that is *not* constructible?

d. Is

$$\sqrt{2-\sqrt{5}}$$

constructible? (Be careful!)

e. Suppose that $\sqrt{\pi}$ were a constructible number. Why would this mean that we could square a circle of radius 1? (We'll see in Chapter 39 that $\sqrt{\pi}$ is not constructible.)

f. Suppose that $\sqrt[3]{2}$ were a constructible number. Why would this mean that we could double a cube with edge length 1? (We'll see in Chapter 39 that $\sqrt[3]{2}$ is not constructible.)

g. Euclid's first postulate states: "Let the following be postulated: to draw a straight line from any point to any point." His second postulate says: "To produce a finite straight line continuously in a straight line." With what tool do these postulates equip us? Why is the tool 'idealized'?

h. Euclid's third postulate states: "To describe a circle with any center and any distance." With what tool does this postulate equip us? Why is the tool 'idealized'?

Exercises

1. Give the construction for a/b required in the Theorem 37.2.

2. (a) Construct $\sqrt{5}$, using the method described in the proof of the last theorem.

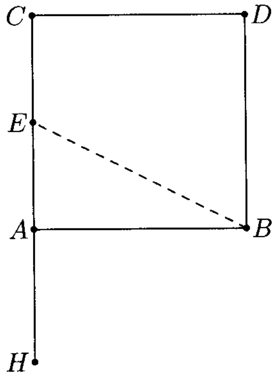
   (b) Now provide an easier construction of $\sqrt{5}$, by constructing the diagonal of an appropriately chosen rectangle.

3. Explain the steps necessary to construct $\sqrt[4]{3}+1$ and $\sqrt{\frac{5}{2}+\sqrt{7}}$, when starting with a line segment of length 1.

4. Explain the steps necessary to construct $\sqrt{5+2\sqrt{6}}$ and $\sqrt{2}+\sqrt{3}$. Then show that these two numbers are the same.

5. Show that $\sqrt{3+2\sqrt{2}}$ is of the form $a+b\sqrt{2}$, for some integers a and b. This exercise reinforces the lesson of Exercise 4: A given element of $\mathbb{K}$ may be constructible by quite a distinct list of steps!

6. (a) Perform the following construction with compass and straightedge, justifying each step: Take a line segment $\overline{AB}$, and construct on it a square $ABDC$. Then find the midpoint E of $\overline{AC}$. Extend line segment $\overline{AC}$ so that A is between E and H, and $\overline{BE}=\overline{EH}$. (See diagram below.)



(b) Now apply the Pythagorean Theorem to the right triangle EAB to show that

$$\overline{AB}(\overline{AB}-\overline{AH}) = \overline{AH}^2.$$

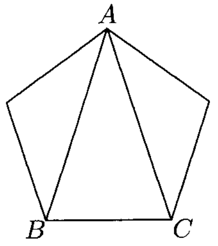
(c) If $|\overline{AB}| = 1$, what quadratic equation in $\rho = |\overline{AH}|$ do we obtain from part b?

(d) Show that $1/\rho = \rho + 1$.

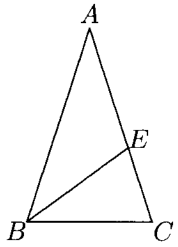
(e) What is the value of the constructible number ρ? This number (or its reciprocal) is called the **Golden Section**, and the construction we have given for it appears as Proposition 11 in Book 2 of Euclid's *Elements*.

7. In this exercise you will show that the regular pentagon is constructible.

(a) In the diagram below, show that $\angle BAC = 36°$ and $\angle ABC = \angle ACB = 72°$.



(b) Let E be the point where the bisector of $\angle ABC$ intersects the opposite side.



Note that $\triangle ABC$ is similar to $\triangle BCE$. If $\overline{BC} = 1$ and $\overline{EC} = x$, show that $x^2 + x - 1 = 0$.

(c) From Exercise 6c, you should have noted that the positive solution to this quadratic is the golden section, $\rho = \frac{1+\sqrt{5}}{2}$. Now argue that the regular pentagon is constructible.

8. Technically speaking, the compass given by Euclid's third postulate (see Exercise h) is *collapsible*: it cannot be used directly to transfer distances from one line segment to another. Prove that distances can be transferred, with a collapsible compass and a straightedge, and so the modern compass is mathematically equivalent to the Greek collapsible compass.

Chapter 38

Constructibility and Quadratic Field Extensions

In the previous chapter we saw that constructible numbers can be obtained from the unit segment by addition, subtraction, multiplication, division (the field operations), and by taking square roots of positive numbers. We may, of course, perform any number of these operations in any order we please. Note that being closed under taking square roots of positive elements is a property that $\mathbb{K}$ does not share with all fields. For example, $2 \in \mathbb{Q}$ (the smallest field of constructible numbers, as we have seen) but $\sqrt{2} \notin \mathbb{Q}$.

38.1 Quadratic Field Extensions

Let's build a bigger field than $\mathbb{Q}$ that does contain $\sqrt{2}$. Consider the set

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

It is straightforward to show that $\mathbb{Q}(\sqrt{2})$ satisfies all the field axioms.

▷ **Quick Exercise.** Check that $\mathbb{Q}(\sqrt{2})$ is a subring of $\mathbb{R}$, and then verify that it is in fact a field. ◁

In fact, $\mathbb{Q}(\sqrt{2})$ is the *smallest* field containing both $\mathbb{Q}$ and $\sqrt{2}$; that is, $\mathbb{Q}(\sqrt{2})$ is contained in every subfield of $\mathbb{C}$ containing both $\mathbb{Q}$ and $\sqrt{2}$. For if $\sqrt{2}$ is an element of a field $F \subseteq \mathbb{C}$ because $\mathbb{Q}$ necessarily is contained in F, then so is $a + b\sqrt{2}$ for all a in $\mathbb{Q}$. Hence, $\mathbb{Q}(\sqrt{2}) \subseteq F$. Because F was arbitrary, $\mathbb{Q}(\sqrt{2})$ is contained in all subfields of $\mathbb{C}$ containing both $\sqrt{2}$ and $\mathbb{Q}$.

We say that $\mathbb{Q}(\sqrt{2})$ is a **field extension** of $\mathbb{Q}$, because $\mathbb{Q}$ is a subfield of $\mathbb{Q}(\sqrt{2})$; since $\mathbb{Q}(\sqrt{2})$ is a strictly larger field, we call this a **proper** field extension. More specifically, $\mathbb{Q}(\sqrt{2})$ is a **quadratic field exten-**

sion of $\mathbb{Q}$. In general, a field $H \subseteq \mathbb{C}$ is a quadratic field extension of a field F if

$$H = \{a + b\sqrt{k} : a, b, \in F\}$$

for some $k \in F$ such that $\sqrt{k} \notin F$. (If $\sqrt{k} \in F$, then $H = F$, and so H would not be a proper extension.) We will return to this definition of quadratic field extension in Chapter 42, when we will put it in a more general context. Note also that when talking about field extensions, we will often omit the word 'field' if the context is clear.

▷ **Quick Exercise.** Show that if $F \subseteq \mathbb{C}$ is a field and $k \in F$, then

$$H = \{a + b\sqrt{k} : a, b \in F\}$$

is a field. If $\sqrt{k} \notin F$, then H properly contains F. That is, H is a quadratic field extension of F. (This is simply a generalization of $\mathbb{Q}(\sqrt{2})$ being a quadratic field extension of $\mathbb{Q}$.) ◁

Example 38.1

Consider the field $\mathbb{Q}(i)$; it is a quadratic extension of $\mathbb{Q}$ because $i = \sqrt{-1}$. This is the field of all complex numbers whose real and imaginary parts are rational. Of course, $\mathbb{Q}(i)$ does not consist entirely of constructible numbers, because $\mathbb{K} \subseteq \mathbb{R}$.

Example 38.2

Let's build a quadratic extension of the field $\mathbb{Q}(\sqrt{2})$. To do so, we need an element belonging to this field whose square root does not. Does $\sqrt{3} \in \mathbb{Q}(\sqrt{2})$? If so, then for some a and b in $\mathbb{Q}$,

$$\sqrt{3} = a + b\sqrt{2},$$

and thus

$$3 = (a + b\sqrt{2})^2 = (a^2 + 2b^2) + (2ab)\sqrt{2}.$$

But then $2ab = 0$, and so at least one of a and b must be zero. Hence, there is no solution to the equation $3 = a^2 + 2b^2$. Thus, we can consider the quadratic extension field $\mathbb{Q}(\sqrt{2})(\sqrt{3})$, which we usually write as $\mathbb{Q}(\sqrt{2}, \sqrt{3})$.

But what are the elements of this field? A typical element should look like

$$(a + b\sqrt{2}) + (c + d\sqrt{2})\sqrt{3} = a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6}.$$

▷ **Quick Exercise.** Check that the quadratic extension $\mathbb{Q}(\sqrt{3})(\sqrt{2})$ consists of the same elements as $\mathbb{Q}(\sqrt{2})(\sqrt{3})$. Thus, the notation $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ is not ambiguous about the order in which the elements $\sqrt{2}$ and $\sqrt{3}$ are adjoined to $\mathbb{Q}$. ◁

In this second example (unlike the first) we obtain a field of constructible numbers. This is more generally true: If F is a field of constructible numbers—that is, F is a subfield of $\mathbb{K}$—and k is an element of F, then $F(\sqrt{|k|})$ is also a field of constructible numbers. We state this in the following lemma.

Lemma 38.1 *Suppose that F is a field of constructible numbers and $k \in F$ with $k > 0$. Then $F(\sqrt{k}) \subseteq \mathbb{K}$.*

Proof: If $\sqrt{k} \in F$, then $F(\sqrt{k}) = F$, and we are done. So assume that $\sqrt{k} \notin F$. By Theorem 37.4, $\sqrt{k}$ is constructible. If $\alpha \in F(\sqrt{k})$, then $\alpha = a + b\sqrt{k}$ for some a and b in F, and so α is also constructible, because $\mathbb{K}$ is a field. □

38.2 Sequences of Quadratic Field Extensions

We can thus start with the rational numbers and repeatedly take quadratic extensions by square roots of positive elements, in the process building larger and larger fields, always contained within the field of constructible numbers. We formally state this in the following theorem.

Theorem 38.2 *Suppose that*

$$\mathbb{Q} = F_0 \subset F_1 \subset \cdots \subset F_n$$

is a sequence of fields such that $F_{i+1} = F_i(\sqrt{k_i})$ for some $k_i \in F_i$, with $k_i > 0$ for $i = 0, 1, \ldots, n-1$. Then $F_n \subseteq \mathbb{K}$.

Proof: Use induction and the previous lemma. □

Are there any constructible numbers that cannot be obtained by this process of repeatedly extending our field by taking square roots of positive elements? The surprising answer is 'No'! Proving this is our next task.

Before we do, let's consider an example of a constructible number, say,

$$\sqrt{6+\frac{4}{3}\sqrt{2+2\sqrt{7}}}.$$

This number could be constructed by successively constructing the following sequence of numbers:

$$\begin{aligned}
X_1 &= 7,\\
X_2 &= \sqrt{7},\\
X_3 &= \sqrt{7}+\sqrt{7} = 2\sqrt{7},\\
X_4 &= 2,\\
X_5 &= 2+2\sqrt{7},\\
X_6 &= \sqrt{2+2\sqrt{7}},\\
X_7 &= \frac{4}{3},\\
X_8 &= \frac{4}{3}\sqrt{2+2\sqrt{7}},\\
X_9 &= 6,\\
X_{10} &= 6+\frac{4}{3}\sqrt{2+2\sqrt{7}},\\
X_{11} &= \sqrt{6+\frac{4}{3}\sqrt{2+2\sqrt{7}}}.
\end{aligned}$$

The fields corresponding to the above sequence of numbers would be $F_1 = \mathbb{Q}$, $F_2 = F_1(\sqrt{7})$, $F_5 = F_4 = F_3 = F_2$, $F_6 = F_5\left(\sqrt{2+2\sqrt{7}}\right)$, $F_{10} = F_9 = F_8 = F_7 = F_6$, and

$$F_{11} = F_{10}\left(\sqrt{6+\frac{4}{3}\sqrt{2+2\sqrt{7}}}\right).$$

So, compressing this sequence, we see that the sequence of quadratic field extensions given in the above theorem would be

$$\mathbb{Q} \subset F_1 \subset F_2 \subset F_3,$$

$$\text{where}\quad F_1 = \mathbb{Q}(\sqrt{7}),\ F_2 = F_1(\sqrt{2+2\sqrt{7}}),$$

$$\text{and}\quad F_3 = F_2(\sqrt{6+\frac{4}{3}\sqrt{2+2\sqrt{7}}}).$$

(We've abbreviated the original sequence of points somewhat. For instance, to construct 7 we might have constructed 2 ($= 1+1$), 3 ($= 2+1$), 4 ($= 2+2$), and finally 7 ($= 3+4$). Likewise, 6 and $\frac{4}{3}$ would take some intermediate steps. All those 'missing' numbers are in $\mathbb{Q}$, however.)

Note that there may be different paths to reach a given number and hence a different sequence of fields reflecting the order in which the numbers are constructed. (For examples illustrating this, look at Exercises 37.4 and 37.5.) The important point is that each field extension is a quadratic extension and only a finite sequence of extensions is needed.

38.3 The Rational Plane

We now return to using the compass and straightedge. This discussion will clarify what is actually meant by constructing with these tools. In constructing numbers with a compass and a straightedge, we start with a given unit segment somewhere in the plane and start constructing lengths. We can impose a Cartesian coordinate system on the plane so that the left-hand endpoint of the given unit segment is at the origin, and the right-hand endpoint is on the x-axis at location $(1,0)$. All rational numbers on the x-axis can be located by applying only the constructions necessary to carry out field operations (addition, subtraction, multiplication, division). Note that a rational number being constructible means that a line segment of the appropriate length can be constructed somewhere in the plane. But we can easily transfer this length to the x-axis so that one end of the line segment is at the origin. Thus, on the x-axis we can locate $\pm q$ for any $q \in \mathbb{Q}$. We can easily transfer these points to the y-axis with the compass, and so we can locate any point (p,q) in the plane where p and q are in $\mathbb{Q}$.

▷ **Quick Exercise.** How do you locate (p,q) in the plane once p has been located on the x-axis and q has been located on the y-axis? ◁

We have thus located all points in the plane that we can construct by means of only the field operations: We call this the **rational plane**, or the **plane of** $\mathbb{Q}$.

38.4 Planes of Constructible Numbers

Suppose now that somewhere in the plane we construct the length $\sqrt{k}$, where k is some positive rational number, and $\sqrt{k} \notin \mathbb{Q}$. By applying only the field operations to $\sqrt{k}$ and elements from $\mathbb{Q}$, we can locate points on the x and y axes that are in the quadratic extension $\mathbb{Q}(\sqrt{k})$. Thus, we can locate all points in the plane with coordinates (p, q), where p and q are in $\mathbb{Q}(\sqrt{k})$; we call this set of points the **plane of** $\mathbb{Q}(\sqrt{k})$.

More generally, suppose that F is any subfield of $\mathbb{K}$—that is, a field of constructible numbers. Then the **plane of** F consists of the set of points (p, q) for which p and q are in F.

We now wish to consider what further points can be reached using compass and straightedge alone, when working in the plane of F, where F is some field of constructible numbers. Let's consider the compass first. To use the compass we need to know two points at which to place our compass: one for the center, and one for some point on the circumference. In the case we're describing, these two points must belong to the plane of F. We will call such a circle **a circle in the plane of** F. Likewise, with a straightedge we can draw a line that passes through two points in the plane of F. We will call such a line **a line in the plane of** F. Other constructions are a combination of these two simple constructions.

At this point you might object and say that a constructed circle could have center at *any* point in the entire plane with *any* radius—simply close your eyes and put the compass down. Or, a line drawn with the straightedge need not pass through two points in the plane of F—again, lay the straightedge down arbitrarily. But steps like these are not permitted by the axioms of Greek geometry. Instead, we require step-by-step procedures that can be *replicated*. For example, a successful solution to the angle trisection problem should be an unambiguous list of constructions that when carried out by anyone leads to the same solution. This means that when we draw lines or circles with the straightedge or compass, we must have unambiguous information: A circle is determined by its center and a point on its circumference, a line by two points.

Let's be explicit algebraically about what equations for lines and circles in the plane of a field F look like.

The equation for a line in the plane of F is of the form

$$ax + by + c = 0$$

where a, b, and c are in F, and a and b are not both zero. For if the line passes through the points (x_1, y_1) and (x_2, y_2), then any point (x, y) on this line must satisfy

$$\frac{y - y_1}{x - x_1} = \frac{y_2 - y_1}{x_2 - x_1},$$

provided $x_1 \neq x_2$. If (x_1, y_1) and (x_2, y_2) are both in the plane of F, then putting the above equation into the form $ax + by + c = 0$ will give us a, b, and c in the field F. The case where $x_1 = x_2$ is an easily handled special case.

▷ **Quick Exercise.** Show that we can obtain an equation of the form $ax + by + c = 0$ with $a, b, c \in F$ for the line through two points in the plane of F with equal x-coordinates. ◁

▷ **Quick Exercise.** Determine the equation of the line passing through $(2 + \sqrt{5}, -\sqrt{5})$ and $(4 + 3\sqrt{5}, 2 + 7\sqrt{5})$, and check that the coefficients a, b, and c do belong to $\mathbb{Q}(\sqrt{5})$. ◁

The equation for a circle in the plane of F is of the form

$$x^2 + y^2 + dx + ey + f = 0,$$

where d, e, and f are in F. For if the circle has center at (x_1, y_1) and a point on the circumference is (x_2, y_2), then the radius is

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

and so if (x, y) is any point on the circle, it must satisfy

$$(x - x_1)^2 + (y - y_1)^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2.$$

This can be put into the desired form where d, e, and f are in the field F if (x_1, y_1) and (x_2, y_2) are in the plane of F.

▷ **Quick Exercise.** Put this equation into the form

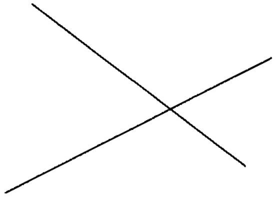
$$x^2 + y^2 + dx + ey + f = 0$$

and note that d, e, and f are elements of F. ◁

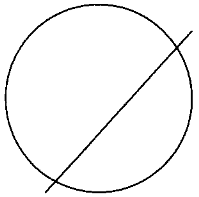
▷ **Quick Exercise.** Determine the equation of the circle with center $(2+\sqrt{5}, -\sqrt{5})$ and $(4+3\sqrt{5}, 2+7\sqrt{5})$ on the circumference, and check that the coefficients d, e, and f are in $\mathbb{Q}(\sqrt{5})$. ◁

So, given the circle and line constructions we can make in the plane of F, what new points can we locate? New points can be located in one of three ways:

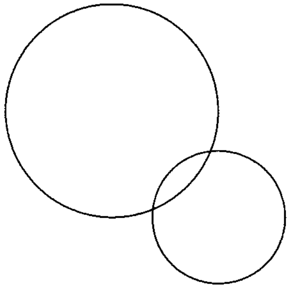
1. The intersection of two lines in the plane of F.



2. The intersection of a circle in the plane of F with a line in the plane of F.



3. The intersection of two circles in the plane of F.



We can easily locate points found by method (1) by solving the system of two linear equations. Notice that the solution, if the lines are not parallel, is a point (p_1, p_2) with p_1 and p_2 in the field F, because our method of solving two equations in two unknowns involves *only field operations.* In other words, method (1) can locate no points outside of the plane of F.

▷ **Quick Exercise.** Find the simultaneous solution to the equations $a_1x + b_1y + c_1 = 0$ and $a_2x + b_2y + c_2 = 0$. Under what conditions will these two equations have no simultaneous solution and thus represent parallel lines? ◁

Method (2) involves the simultaneous solution of an equation for a circle and an equation for a line. Method (3) involves the simultaneous solution of two equations for circles. Notice that method (3) reduces to method (2), for if

$$x^2 + y^2 + d_1x + e_1y + f_1 = 0$$

is subtracted from

$$x^2 + y^2 + d_2x + e_2y + f_2 = 0,$$

we get

$$(d_2 - d_1)x + (e_2 - e_1)y + (f_2 - f_1) = 0,$$

and so the simultaneous solution of this linear equation (which gives the equation for the common chord; see Exercise 38.8) with either of the two circle equations is the desired solution. Thus, what remains is to find which points can be obtained from method (2).

So, suppose we wish to solve simultaneously

$$\begin{gathered} x^2 + y^2 + dx + ey + f = 0, \text{ and} \\ ax + by + c = 0, \end{gathered}$$

where all the coefficients belong to the field F. Because a and b can't both be zero we will assume that $b \neq 0$ (the case where $a \neq 0$ is similar). We solve for y in terms of x:

$$y = -\frac{a}{b}x - \frac{c}{b}.$$

Substituting this into the circle equation yields:

$$x^2 + \left(-\frac{a}{b}x - \frac{c}{b}\right)^2 + dx + e\left(-\frac{a}{b}x - \frac{c}{b}\right) + f = 0$$

which is a quadratic equation in x. We could use the quadratic formula to solve it; we will not carry out the calculations explicitly.

▷ **Quick Exercise.** Apply the quadratic formula to this quadratic equation, and obtain the solutions in terms of the coefficients. ◁

The solution you just obtained has the form $A \pm B\sqrt{k}$, with $A, B, k \in F$. If $k < 0$, then there are no *real* solutions, and this means geometrically that there is no intersection. If $k = 0$, then there is a unique solution; this means geometrically that the line is *tangent* to the circle. Finally, if $k > 0$, we have two distinct intersections.

In case we have solutions, we can determine y by substitution: This yields an expression of the form $A' \pm B'\sqrt{k}$, with $A', B' \in F$.

▷ **Quick Exercise.** Show that if there are solutions for x of the form $A \pm B\sqrt{k}$, with $A, B, k \in F$, then y has the form $A' \pm B'\sqrt{k}$, with $A', B' \in F$. ◁

So, notice that both x and y are in the field $F(\sqrt{k})$; if $\sqrt{k} \in F$, then of course $F(\sqrt{k}) = F$, and so x and y are in F. That is, the point (x, y) is in the plane of $F(\sqrt{k})$.

38.5 The Constructible Number Theorem

We are now ready to prove the main result of this chapter.

Theorem 38.3 Constructible Number Theorem *The following two statements are equivalent:*

a. The number α is constructible.

b. There exists a finite sequence of fields

$$\mathbb{Q} = F_0 \subset F_1 \subset \cdots \subset F_N$$

with $\alpha \in F_N$ and $F_{i+1} = F_i(\sqrt{k_i})$ for some $k_i \in F_i$, with $k_i > 0$ for $i = 0, \ldots, N-1$.

Proof: That (b) implies (a) is the previous theorem.

To show that (a) implies (b), suppose α is constructible. Then we can construct the point $(\alpha, 0)$, which we will label P, starting only with the segment of length 1 along the positive x-axis. To construct P, we must construct a finite sequence of points, $P_0, P_1, \ldots, P_M = P$. Because the points $(0,0)$ and $(1,0)$ are the first two points constructed, we set $P_0 = (0,0)$ and $P_1 = (1,0)$. Of course, P_0 and P_1 are both elements of the plane of $\mathbb{Q}$. Now P_2 is constructed using only P_0, P_1 and one of the three constructions for locating new points in the plane. Hence, by the above discussion, $P_2 \in \mathbb{Q}(\sqrt{k})$ for some positive $k \in \mathbb{Q}$. (This field may be equal to $\mathbb{Q}$ if $\sqrt{k} \in \mathbb{Q}$.) Then $F_0 = F_1 = \mathbb{Q}$ and $F_2 = \mathbb{Q}(\sqrt{k})$.

We now proceed inductively. Let F_i be the smallest field containing the points $P_0, \ldots, P_i$. By the induction hypothesis, for each $i = 3, 4, \ldots, M$, P_i was constructed using only the points constructed before it and one of the three constructions for locating points in the plane. Hence, $P_i \in F_{i-1}(\sqrt{k_{i-1}})$ for some positive $k_{i-1} \in F_{i-1}$.

Noting only those times the field F_i is a proper extension of F_{i-1}, we thus have that α is in the field F_N where

$$\mathbb{Q} = F_0 \subset F_1 \subset \cdots \subset F_N$$

and $F_{i+1} = F_i(\sqrt{k_i})$ where $k_i \in F_i$ with $k_i \geq 0$, for $i = 0, 1, \ldots, N-1$. (Notice that N, the number of different fields needed in the construction of α, is probably smaller than M, the number of points actually constructed.) □

Thus, for any constructible number, there is a *finite* sequence of quadratic extension fields, the last of which will contain the given number. Of course, this was exactly what we discovered for the particular constructible number

$$\sqrt{6 + \frac{4}{3}\sqrt{2 + 2\sqrt{7}}}.$$

In Chapters 42–44 we will provide a more general theory of field extensions (including degrees other than 2); but this will require some new concepts, which we introduce in Chapters 40 and 41. See Exercise 38.5 for an example of a non-quadratic extension.

Chapter Summary

We defined *quadratic field extension.* We examined what new points could be constructed from points previously constructed and found that a new point was always in a quadratic field extension of the field determined by the old points.

Finally, we proved the main result of this section, the *Constructible Number Theorem*, which completely describes in *algebraic* terms which numbers are constructible.

Warm-up Exercises

a. Explain why the Constructible Number Theorem 38.3 guarantees that the following numbers are constructible:

$$\sqrt{6}, \quad \sqrt[8]{6}, \quad \sqrt{2+\sqrt[4]{5}}.$$

b. What is the quadratic field extension $\mathbb{R}(i)$ usually known as?

c. Does $\mathbb{C}$ admit any proper quadratic field extensions?

d. Does $\mathbb{K}$ admit any proper quadratic field extensions? Does it admit any proper quadratic field extensions that are subfields of $\mathbb{R}$?

e. Did we discuss a real number in this chapter that is *not* constructible?

f. Suppose that ℓ_1 and ℓ_2 are lines in the plane of a field F of constructible numbers. What can you say about the intersection of ℓ_1 and ℓ_2?

g. Suppose that a, b, and c are constructible numbers, and $ax^2 + bx + c$ has real roots. Are these roots constructible numbers?

Exercises

1. Show that $\sqrt{5} \notin \mathbb{Q}(\sqrt{3})$ by showing that 5 cannot be the square of a number of the form $a + b\sqrt{3}$ where a and b are in $\mathbb{Q}$.

2. If $F = \mathbb{Q}(\sqrt{3})$, describe the set of elements of the field $F(\sqrt{5})$. Show that this field is the same as the quadratic extension

$$\mathbb{Q}(\sqrt{5})(\sqrt{3}).$$

3. Generalize Exercise 2: Suppose that p and q are distinct positive prime integers; prove that

$$\mathbb{Q}(\sqrt{p})(\sqrt{q}) = \mathbb{Q}(\sqrt{q})(\sqrt{p}),$$

and describe the elements of the field.

4. Give a sequence of numbers necessary to construct the number

$$\sqrt[4]{2+4\sqrt{3}}.$$

Give the corresponding sequence of fields.

5. We've been able to give a nice description of the smallest field containing both $\sqrt{2}$ and $\mathbb{Q}$; namely, $\{a + b\sqrt{2} : a, b \in \mathbb{Q}\}$. In this problem we try to give a description of the smallest field containing $\sqrt[3]{2}$ and $\mathbb{Q}$. Why is $\{a + b\sqrt[3]{2} : a, b \in \mathbb{Q}\}$ not the answer? Now consider $\{a + b\sqrt[3]{2} + c\sqrt[3]{4} : a, b, c \in \mathbb{Q}\}$. When proving this is a field, the difficult part is showing that a typical element $a + b\sqrt[3]{2} + c\sqrt[3]{4}$, with $a, b, c \in \mathbb{Q}$ and not all zero, has a multiplicative inverse. You might not be successful in this part, and we will give an explicit hint below. But do at least verify that this set is indeed a commutative ring with unity.

Here is a hint for how to show that a typical element indeed has a multiplicative inverse. In fact, the multiplicative inverse of $a + b\sqrt[3]{2} + c\sqrt[3]{4}$ is $d + e\sqrt[3]{2} + f\sqrt[3]{4}$ where

$$d = \frac{-2bc + a^2}{2b^3 + 4c^3 + a^3 - 6bca}$$
$$e = \frac{2c^2 - ba}{2b^3 + 4c^3 + a^3 - 6bca}, \quad \text{and}$$
$$f = \frac{ca - b^2}{2b^3 + 4c^3 + a^3 - 6bca}.$$

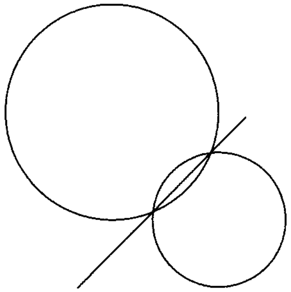
You can verify this by computing $(a+b\sqrt[3]{2}+c\sqrt[3]{4})(d+e\sqrt[3]{2}+f\sqrt[3]{4})$ and seeing that it simplifies to 1. (This is obviously a non-trivial calculation!)

We will show (in a less grubby manner) that this set is a field in Theorem 43.3. Note also that we have by no means claimed that $\sqrt[3]{2}$ is a constructible number (see Section 39.1).

6. Describe the elements of the quadratic extension field $F(\sqrt[4]{2})$, where $F = \mathbb{Q}(\sqrt{2})$.

7. Suppose that a, b, and c are constructible numbers. Consider the polynomial $ax^4 + bx^2 + c$: Are its real roots constructible? (See Warm-up Exercise g above.)

8. Argue that the linear equation that results when two circles are intersected algebraically gives the equation of the common chord of the two circles. (See the figure below.)



9. Find the equation of tangent line to the circle

$$(x-1)^2 + y^2 = 5$$

at the point $(3, 1)$ by algebra: This is the only line that intersects this circle in exactly this point. *Note:* This was Rene Descartes' method for determining a tangent line. Doing this problem should make you appreciate the *geometric* approach to finding the tangent (it is perpendicular to a radial line) and the *calculus* approach (calculate dy/dx).

Chapter 39

The Impossibility of Certain Constructions

The Constructible Number Theorem 38.3 is the most important piece of machinery necessary to show that the three construction problems of the Greeks are indeed impossible with a compass and a straightedge.

39.1 Doubling the Cube

We first tackle the problem of doubling the cube. Recall the problem:

> *Given a line segment representing the edge of a cube, construct another line segment representing the edge of a cube with twice the volume of the original cube.*

We start with a line segment of length 1, and consider the cube with this segment as one edge. A cube of twice the volume would have edges of length $\sqrt[3]{2}$. So, doubling the cube then amounts to constructing the number $\sqrt[3]{2}$. We will show that $\sqrt[3]{2}$ is not constructible by using the Constructible Number Theorem. We must prove that $\sqrt[3]{2}$ cannot be an element of a field at the end of a finite sequence of quadratic field extensions that starts with the rational numbers. The next lemma is the key to showing this.

Lemma 39.1 *Let $F(\sqrt{k})$ be a real quadratic field extension of a field F. If $\sqrt[3]{2} \in F(\sqrt{k})$, then $\sqrt[3]{2} \in F$.*

Proof: Suppose that $\sqrt[3]{2} \in F(\sqrt{k})$, a proper quadratic extension of the field F. Then $\sqrt[3]{2} = a + b\sqrt{k}$, with $a, b \in F$ and $\sqrt{k} \notin F$. We want to show that $b = 0$. But,

$$2 = (a + b\sqrt{k})^3$$

$$= a^3 + 3a^2b\sqrt{k} + 3ab^2k + b^3k\sqrt{k}$$
$$= (a^3 + 3ab^2k) + (3a^2b + b^3k)\sqrt{k}.$$

If $3a^2b+b^3k \neq 0$, then we could solve the above equation for $\sqrt{k}$; the resulting equation would show that $\sqrt{k} \in F$. Hence, $3a^2b + b^3k = 0$. But then,

$$\begin{aligned}(a - b\sqrt{k})^3 &= (a^3 + 3ab^2k) - (3a^2b + b^3k)\sqrt{k}\\ &= a^3 + 3ab^2k\\ &= 2,\end{aligned}$$

and so $a - b\sqrt{k}$ is also a cube root of 2. Thus, $a + b\sqrt{k}$ and $a - b\sqrt{k}$ are both real roots of $x^3 - 2$. But

$$x^3 - 2 = (x - \sqrt[3]{2})(x^2 + \sqrt[3]{2}x + \sqrt[3]{4}),$$

and the quadratic factor is irreducible in $\mathbb{R}[x]$.

▷ **Quick Exercise.** Use the quadratic formula to check that this quadratic factor is irreducible in $\mathbb{R}[x]$. ◁

Hence, $a + b\sqrt{k} = a - b\sqrt{k}$ which implies that $b = 0$, as we wished.□

It now follows from Lemma 39.1 that:

Theorem 39.2 *It is impossible to double the cube.*

Proof: As noted, doubling the cube is equivalent to constructing $\sqrt[3]{2}$. But if $\sqrt[3]{2}$ were constructible, by the Constructible Number Theorem 38.3, there would exist a finite sequence of quadratic field extensions, $\mathbb{Q} = F_0 \subset F_1 \subset \cdots \subset F_N$, with $\sqrt[3]{2} \in F_N$.

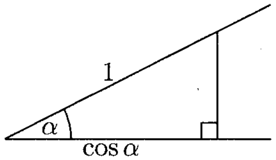
But Lemma 39.1 says that if $\sqrt[3]{2} \in F_N = F_{N-1}(\sqrt{k})$, then $\sqrt[3]{2} \in F_{N-1}$. Repeating this argument inductively implies that $\sqrt[3]{2} \in \mathbb{Q}$, which is false. (See Exercise 5.13.) □

39.2 Trisecting the Angle

A similar approach is used to show that trisecting an angle is impossible. Again, recall the problem:

Given an arbitrary angle, divide it into three equal parts.

The problem calls for a method to trisect all angles. If we can exhibit just one angle that can't be trisected, then such a method does not exist. The angle we will use is the 60° angle. As we've noted before, the 60° angle is constructible—as an angle of an equilateral triangle, for instance. Trisecting a 60° angle implies the construction of a 20° angle. But if we can construct an angle α, we can construct the cosine of α, as the figure below shows.



So being able to trisect a 60° angle implies that cos 20° is a constructible number; this is what we will show is impossible, using the Constructible Number Theorem 38.3.

We use the formulas for the sine and cosine of the sum of angles to perform the following derivation:

$$\begin{aligned}\cos 3\theta &= \cos(2\theta + \theta)\\ &= \cos 2\theta \cos\theta - \sin 2\theta \sin\theta\\ &= (\cos^2\theta - \sin^2\theta)\cos\theta - (2\sin\theta\cos\theta)\sin\theta\\ &= \cos^3\theta - 3\sin^2\theta\cos\theta\\ &= \cos^3\theta - 3(1 - \cos^2\theta)\cos\theta\\ &= 4\cos^3\theta - 3\cos\theta\end{aligned}$$

Setting $\theta = 20°$, $\cos 3\theta = \cos 60° = 1/2$, and so $\cos 20°$ is a solution to $4x^3 - 3x - 1/2 = 0$ or $8x^3 - 6x - 1 = 0$; thus, $2\cos 20°$ is a solution to $x^3 - 3x - 1 = 0$. So, if $\cos 20°$ were constructible then so would $2\cos 20°$, which is a root of $x^3 - 3x - 1$. Thus, if $\cos 20°$ were constructible, it would be possible to construct a real root of $x^3 - 3x - 1$. We will show that this is impossible, using the next lemma, which has the same flavor as Lemma 39.1 above.

Lemma 39.3 *Let $F(\sqrt{k})$ be a quadratic field extension of a field F. If the equation $x^3 - 3x - 1 = 0$ has a solution in $F(\sqrt{k})$, then it has a solution in F.*

Proof: Let $a + b\sqrt{k}$ be a root of $x^3 - 3x - 1$ in $F(\sqrt{k})$, a proper quadratic extension of the field F. If $b = 0$, then the root is a, which is in F. If $b \neq 0$, we will show that $-2a$ is a root; this will still mean that $x^3 - 3x - 1$ has a root in F. So, if $b \neq 0$,

$$\begin{aligned} 0 &= (a + b\sqrt{k})^3 - 3(a + b\sqrt{k}) - 1 \\ &= a^3 + 3a^2b\sqrt{k} + 3ab^2k + b^3k\sqrt{k} - 3a - 3b\sqrt{k} - 1 \\ &= (a^3 + 3ab^2k - 3a - 1) + (3a^2b + b^3k - 3b)\sqrt{k}. \end{aligned}$$

But $3a^2b + b^3k - 3b = 0$, for otherwise, $\sqrt{k} \in F$. But then $a^3 + 3ab^2k - 3a - 1 = 0$. After dividing the first equation by b (we know $b \neq 0$), we have $3a^2 + b^2k - 3 = 0$, and so $b^2k = 3 - 3a^2$. Substituting this into the second equation we have,

$$\begin{aligned} 0 &= a^3 + 3a(3 - 3a^2) - 3a - 1 \\ &= a^3 + 9a - 9a^3 - 3a - 1 \\ &= -8a^3 + 6a - 1 \\ &= (-2a)^3 - 3(-2a) - 1. \end{aligned}$$

In other words, $-2a$ is a root of $x^3 - 3x - 1$. □

Theorem 39.4 *It is not possible to trisect an arbitrary angle.*

Proof: As noted before, if we could trisect a 60° angle, we could construct a 20° angle. This means we could construct the number $\cos 20°$, and this implies that we can construct a root of $x^3 - 3x - 1$. Using Lemma 39.3 and the Constructible Number Theorem 38.3 and arguing as in the previous theorem, we see that this implies that there is a rational root of $x^3 - 3x - 1$. But, by the Root Theorem 4.3, this implies that $x^3 - 3x - 1$ factors in $\mathbb{Q}[x]$. However, we can see that this polynomial is irreducible in $\mathbb{Z}[x]$ and so, by Gauss's Lemma 5.5, is also irreducible in $\mathbb{Q}[x]$.

▷ **Quick Exercise.** Why is $x^3 - 3x - 1$ irreducible in $\mathbb{Z}[x]$? ◁

Hence, $x^3 - 3x - 1$ has no rational root, and so we cannot trisect a 60° angle. □

39.3 Squaring the Circle

Finally, we turn our attention to squaring the circle. Let's state the problem carefully:

Given a circle, construct a square with the same area.

We are unable to give the full proof here of the impossibility of squaring the circle, because one step is quite difficult. Given a circle with radius 1, its area is π, and so to square the circle we must be able to construct the number $\sqrt{\pi}$ (to be the side of the required square). Obviously, this is possible, if we could construct the number π itself. (This is the argument for Exercise 37.e.)

In the 18th century the German mathematician Johann Heinrich Lambert managed to prove that π is not a rational number, by showing that if x is rational, then $\tan x$ cannot be; because $\tan(\pi/4) = 1$, Lambert's theorem means that $\pi/4$ (and hence π) cannot be rational. Lambert conjectured that π is a **transcendental** number; that is, a number that is not the root of any polynomial in $\mathbb{Q}[x]$. This was finally proved by another German mathematician, Ferdinand Lindemann, in 1882; he made heavy use of the work of the Frenchman Charles Hermite, who had proved the transcendence of e a decade earlier. As we shall see shortly, Lindemann's theorem finally laid to rest the last of the great constructibility problems of the ancient Greeks.

Theorem 39.5 Lindemann's Theorem *π is transcendental.*

Proof: The proof is long, difficult, and analytic, rather than algebraic. For an accessible version, see *Field Theory and its Classical Problems*, by Charles Hadlock. □

To use Lindemann's Theorem to prove the impossibility of squaring the circle, we first need a little more machinery.

Lemma 39.6 *Suppose $F(\sqrt{k})$ is a quadratic field extension of F. If α is a root of a polynomial in $F(\sqrt{k})[x]$ of degree n, then it is the root of a polynomial in $F[x]$ of degree $2n$.*

Proof: Assume that α is a solution to an equation of the form

$$(a_n + b_n\sqrt{k})x^n + \cdots + (a_0 + b_0\sqrt{k}) = 0$$

where all the a_i's and b_i's are in the field F and a_n and b_n are not both zero. Moving the terms with $\sqrt{k}$ to the right side of the equation, we get

$$a_nx^n + \cdots + a_0 = -\sqrt{k}(b_nx^n + \cdots + b_0).$$

Squaring both sides and moving all terms to the left gives a polynomial in $F[x]$ that also has α as a root. Now the leading term of this polynomial is $(a_n^2 - kb_n^2)x^{2n}$. This coefficient is not zero, because otherwise $k = a_n^2/b_n^2$, contradicting the fact that $\sqrt{k} \notin F$. □

Theorem 39.7 *If α is constructible, then α is the root of a polynomial in $\mathbb{Q}[x]$ of degree 2^r, for some $r \in \mathbb{N}$.*

Proof: By the Constructible Number Theorem 38.3, $\alpha \in F_N$, where $\mathbb{Q} = F_0 \subset F_1 \subset \cdots \subset F_N$ is a sequence of quadratic field extensions. But α is the root of a linear equation in F_N, namely, $x - \alpha$. By repeated application of Lemma 39.6, α is the root of a polynomial in $\mathbb{Q}[x]$ of degree 2^N. □

So, if π were constructible, it would have to be the root of some polynomial in $\mathbb{Q}[x]$; Lindemann's Theorem says this is not true, and so we have:

Theorem 39.8 *It is not possible to square the circle.*

Historical Remarks

Our proofs that it is impossible to duplicate the cube and to trisect an arbitrary angle are similar in flavor to the first such proofs, by Pierre Wantzel, which appeared in 1837. His version of the Constructible Number Theorem asserted that any constructible number is the root of an irreducible polynomial with degree a power of two, and hence numbers like $\sqrt[3]{2}$ and $\cos(20°)$ are not constructible.

Another important impossibility result had been obtained a decade earlier by the Norwegian mathematician Niels Abel, who showed that it is impossible to solve an arbitrary fifth-degree equation, using only elementary algebra and the extraction of roots. We made reference to this result in the Historical Remarks following Chapter 9.

It turns out that both these achievements can be viewed most elegantly as part of a general theory of field extensions of the rational numbers, and it was the French mathematician Evariste Galois who laid the important groundwork for this theory. In the remainder of this book we will look into this important area of algebra.

Chapter Summary

We proved that it is impossible to double the cube, trisect an angle, or square the circle using only a compass and a straightedge. The outline of the proofs are the same: If the construction were possible, then we could construct a certain number. (For us, the numbers are $\sqrt[3]{2}$ for doubling the cube, $\cos 20°$ for trisecting an angle, and π for squaring the circle.) But the Constructible Number Theorem or its corollaries show that each number is not constructible; therefore, the three constructions are impossible.

Warm-up Exercises

a. We have finally discussed some real numbers that are *not* constructible! Give some examples.

b. During our discussion of trisecting the angle, we proved that if an angle θ is constructible, then $\cos\theta$ is constructible. Explain why the converse of this statement is true too.

c. Lambert proved in the 18th century that π is not rational; why is this *not* sufficient to show that the circle cannot be squared?

d. Explain why it is possible to double the square.

e. Explain why it is possible to 'octuple' the cube.

Exercises

1. (a) The number $\sqrt{2} + \sqrt{3}$ is constructible. It is an element of $F(\sqrt{3})$ where $F = \mathbb{Q}(\sqrt{2})$. In fact, $\sqrt{2} + \sqrt{3}$ is the root of the polynomial $x - (\sqrt{2} + \sqrt{3})$ in $F(\sqrt{3})$. Find a polynomial in $\mathbb{Q}[x]$ for which $\sqrt{2} + \sqrt{3}$ is a root.

   (b) Find a polynomial in $\mathbb{Q}[x]$ for which $\sqrt{4 + \sqrt{7}}$ is a root.

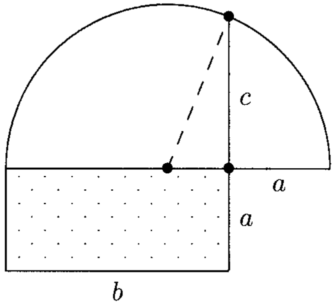
2. Show that $\sqrt[3]{2}+\sqrt{2}$ is a root of the polynomial

$$x^3 - 3\sqrt{2}x^2 + 6x - (2+2\sqrt{2})$$

in $\mathbb{Q}(\sqrt{2})[x]$. Then obtain a sixth-degree polynomial in $\mathbb{Q}[x]$ that has this number as a root.

3. While it is impossible to square the circle, it is possible to square the rectangle. That is, given a rectangle, it is possible to construct a square of the same area. Do this.

Hint: Consider the diagram below. Show that $c^2 = ab$. Thus, if one constructed a square with sides of length c, it would have the same area as the given rectangle with sides a and b.



4. (a) Show how to square an arbitrary triangle.
   (b) Consider a **polygon** in the plane (by this, we mean just a bounded figure with edges that are line segments). How could you use part a to square such a figure?

5. Suppose that p and q are elements of F, a subfield of the field of real numbers. Let $F(\sqrt{k})$ be a proper quadratic field extension of the field F. Prove that if the equation

$$x^3 + px + q = 0$$

has a solution in $F(\sqrt{k})$, then it has a solution in F.

6. Use the trig identity

$$\cos 3\theta = 4\cos^3\theta - 3\cos\theta,$$

and the previous exercise to prove that the angle $2\pi/9$ is not constructible.

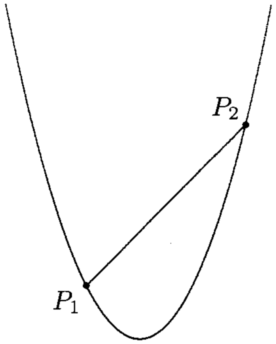
7. Suppose that n is a positive integer, with no integer cube root. Let $F(\sqrt{k})$ be a quadratic field extension of a field F. Prove that if $\sqrt[3]{n}\in F(\sqrt{k})$, then $\sqrt[3]{n}\in F$.

8. Let n be a positive integer, with no integer cube root. Use the previous exercise to prove that we cannot construct a cube with volume equal to n times that of a given cube. (Thus, we cannot triple or quadruple the cube with compass and straightedge.)

9. Consider a parabola (by analytic geometry, we may consider a curve of the form $f(x) = ax^2 + bx + c$). Pick two points

$$P_1(x_1, f(x_1)) \text{ and } P_2(x_2, f(x_2))$$

on the parabola, and consider the region bounded by the parabola and the line segment between the two points. This is called a **segment of the parabola**.



   (a) Prove (using calculus) that the area of the segment is equal to $\frac{4}{3}$ times the area of the triangle $P_1P_2P_3$, where P_3 is the point on the parabola with x coordinate $(x_1+x_2)/2$.
   (b) Explain why this means that we can square the segment of the parabola. This result was known to Archimedes (ca. 250 B.C.); he proved it using the geometry of the parabola, rather than calculus.

10. In the proof of Lemma 39.1 we needed to show that $x^3 - 2$ has only one real root; we did this using algebra. Prove this result using calculus.

11. Suppose that θ is any fixed angle with positive radian measure. In this problem you will show that it is possible to construct an angle arbitrarily close in size to θ.

    (a) Suppose that ϵ is an arbitrary positive number. Describe how to construct an angle ψ whose angle measure is smaller than ϵ.

    (b) Explain why an integer multiple $n\psi$ of the angle constructed in part a must be larger than θ.

    (c) Consider the *smallest* positive integer n so that $n\psi$ is larger than θ (why must n exist?). Use n to find an angle within ϵ radians of θ.

    (d) Why does part c mean that we can come arbitrarily close to constructing the trisection of any given angle?

    (e) What is the philosophical difference between the construction in part d and the sought-for (impossible) trisection construction?

Section VII in a Nutshell

This section presents three famous compass and straightedge construction problems of the ancient Greeks: doubling the cube, trisecting an angle, and squaring the circle. These problems were unsolved by the Greeks. We show that it is not in principle possible to make these constructions, using modern algebraic (and not geometric) techniques.

When starting with a line segment of length 1, we call the length of any line segment we can construct after a finite number of compass and straightedge construction steps to be a *constructible number.* First, we show that the set of constructible numbers is a field (Corollary 37.3). We can construct all rational numbers (Lemma 37.1 and Theorem 37.2) and all square roots of constructible numbers (Theorem 37.4). This development leads to the *Constructible Number Theorem* (Theorem 38.3), which asserts that a number α is constructible exactly if the following condition holds:

There exists a finite sequence of fields

$$\mathbb{Q} = F_0 \subset F_1 \subset \cdots \subset F_N$$

with $\alpha \in F_N$ and $F_{i+1} = F_i(\sqrt{k_i})$ for some $k_i \in F_i$, with $k_i > 0$ for $i = 0, \ldots, N-1$.

We show that it is impossible to double the cube by showing that $\sqrt[3]{2}$ is not constructible (Lemmas 39.1 and Theorem 39.2). We show that it is impossible to trisect a 60° angle (that is, construct a 20° angle) by showing that to do so would imply being able to construct a solution to $x^3 - 3x - 1 = 0$, which we show is impossible (Lemma 39.3 and Theorem 39.4). Finally, we consider the problem of squaring the circle. If this were possible, then π would be a constructible number. *Lindemann's Theorem* (Theorem 39.5 — which we do not prove) says that π is transcendental: that is, π is not the root of any polynomial with rational coefficients. But we show that any constructible number is the root of a polynomial in $\mathbb{Q}[x]$ of degree 2^n (Theorem 39.7) and so is not transcendental. Thus it is impossible to square the circle.

VIII

Vector Spaces and Field Extensions

Chapter 40

Vector Spaces I

The three previous chapters showed the impossibility of the three famous Greek constructibility problems. Some of the grubbier parts of the proofs in the previous chapters can be replaced by more elegant arguments—provided we know more about the algebraic structures involved. The next five chapters will develop the machinery needed for these more sophisticated arguments. The proofs we presented in Chapters 37–39 are correct and have the advantage of not needing a great deal of sophistication in order to understand them. However, the arguments to be presented next have the advantage of being much more elegant and concise (at the expense of being less accessible). This is not very surprising as the more we know, the easier it is to express ourselves. As an added bonus, our additional machinery will enable us to prove another impossibility result, regarding the solution of polynomial equations using arithmetic and root extraction.

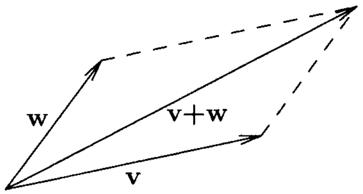
Putting our applications aside, the topics covered in these next few chapters are important in their own right. The first such topic we need to discuss is the notion of *vector space*; this will allow us to better understand the ideas of field extensions. As we shall see, a vector space (like a ring, group, or field) is just a set, equipped with operations satisfying certain nice rules.

The study of vector spaces is a subject in its own right, called *linear algebra*, and you may have the opportunity to take an entire course about this topic, if you haven't already. In the next two chapters we will develop only enough of the theory from this subject in order to better understand fields. There is much more (both computational and theoretical) to linear algebra than we will be able to present here.

40.1 Vectors

In calculus you studied vectors in two and three dimensions. Such vectors provide a very useful way of looking at the geometry of the plane and three-dimensional space. In this chapter, we wish to generalize the properties of such vectors, to cases that are not quite so easily visualized. We will emphasize a more abstract and algebraic approach to vectors, but you should keep in mind that the familiar vectors from calculus are important motivating examples.

Let's start by examining some of the algebraic properties of the familiar three-dimensional vectors. We denote this set of vectors by $\mathbb{R}^3$. Recall that $\mathbb{R}^3 = \{(r_1, r_2, r_3) : r_i \in \mathbb{R}\}$. Two of these vectors may be added coordinate-wise to get another vector. Addition of two vectors has a nice geometric interpretation, as illustrated below.



You may not have given much thought to it at the time, but this vector addition has the following properties: for $\mathbf{v}, \mathbf{w}$ and $\mathbf{u} \in \mathbb{R}^3$,

i. $\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}$,

ii. $\mathbf{v} + (\mathbf{w} + \mathbf{u}) = (\mathbf{v} + \mathbf{w}) + \mathbf{u}$,

iii. there exists a **zero vector** $\mathbf{0}$, with the property that $\mathbf{v} + \mathbf{0} = \mathbf{v}$, and

iv. every vector $\mathbf{v}$ has an **additive inverse** $-\mathbf{v}$, with the property that $\mathbf{v} + (-\mathbf{v}) = \mathbf{0}$.

You should recognize these as the same defining properties possessed by addition in a ring (see Chapter 6) or the properties of the operation in an abelian group (see Chapter 24). Here, the additive identity is the zero vector $\mathbf{0}$. In $\mathbb{R}^3$, $\mathbf{0} = (0, 0, 0)$ and if $\mathbf{v} = (v_1, v_2, v_3)$, then $-\mathbf{v} = (-v_1, -v_2, -v_3)$.

The other arithmetic operation in the algebra of vectors is scalar multiplication, in which vectors are multiplied by scalars to give other vectors. Here, our scalars come from the field $\mathbb{R}$. Scalar multiplication has the following properties: for $r, s \in \mathbb{R}$ and $\mathbf{v}, \mathbf{w} \in \mathbb{R}^3$,

v. $(r + s)\mathbf{v} = r\mathbf{v} + s\mathbf{v}$,

vi. $(rs)\mathbf{v} = r(s\mathbf{v})$,

vii. $r(\mathbf{v} + \mathbf{w}) = r\mathbf{v} + r\mathbf{w}$, and

viii. $1\mathbf{v} = \mathbf{v}$.

Note that the 1 in property (viii) is the scalar 1. You can easily show that these eight properties hold in $\mathbb{R}^2$ as well as $\mathbb{R}^3$. Note that when we considered $\mathbb{R}^3$ as a ring we also defined coordinate-wise multiplication. However, when thinking of elements of $\mathbb{R}^3$ as vectors, we will consider no such operation, because coordinate-wise multiplication of two vectors has no simple geometric interpretation.

40.2 Vector Spaces

There are other algebraic objects for which the above eight properties hold. A **vector space V over a field F** is a set with a binary operation called addition that satisfies properties (i) through (iv) above, together with a scalar multiplication of vectors from V by scalars from F so that if $r, s \in F$ and $\mathbf{v}, \mathbf{w} \in V$, then properties (v) through (viii) listed above hold.

Note that scalar multiplication is not a binary operation in the sense we've discussed before. Instead, scalar multiplication by the field element r maps each vector in V to another vector in V. We write $r\mathbf{v}$ to denote the vector that $\mathbf{v}$ gets mapped to and speak of 'multiplying $\mathbf{v}$ by r'.

It is important to think carefully about the meaning of the axioms (v) through (viii) and in particular where the operations are taking place. For example, the + in axiom (v) denotes the addition in the field F, while the + in axiom (vii) denotes the addition of vectors in V. The juxtaposition of r with s in (vi) denotes multiplication in the field, while the juxtaposition of (rs) with $\mathbf{v}$ denotes the scalar multiplication. You should be careful to keep track of which operation is which.

▷ **Quick Exercise.** Look at the other axioms, and make sure you understand which of the operations are meant in each case. ◁

We now examine some examples of vector spaces:

Example 40.1

The field $\mathbb{C}$ of complex numbers is a vector space over $\mathbb{R}$, the field of real numbers. Here, the vectors are complex numbers and the scalars are real numbers.

To show that this is a vector space, we first note that the set of vectors is closed with respect to addition (because $\mathbb{C}$ is a field, which is closed under addition) and that properties (i) through (iv) are properties of addition in all fields. Now, let's look at property (v). Let r and s be real numbers and $\mathbf{v} = a + bi \in \mathbb{C}$. Then

$$(r+s)\mathbf{v} = (r+s)(a+bi) = r(a+bi) + s(a+bi),$$

because multiplication in a field enjoys the distributive property. Thus, we have shown that $(r+s)\mathbf{v} = r\mathbf{v} + s\mathbf{v}$, as desired. The remaining properties are just as straightforward to show.

▷ **Quick Exercise.** Show that properties (vi) through (viii) hold here. ◁

We will generalize the idea of this example in Exercise 40.9.

Example 40.2

The set of complex numbers $\mathbb{C}$ forms a vector space over itself! Here, the vector space axioms are just properties that hold for addition and multiplication in a field.

▷ **Quick Exercise.** What does axiom vi mean in this context? ◁

In the next example we provide a natural generalization of the vector space $\mathbb{R}^3$ (over the field $\mathbb{R}$).

Example 40.3

The set F^n of n-tuples with coordinates from the arbitrary field F is a vector space over F. Here, a typical vector is of the form $(a_1, a_2, \ldots, a_n)$ where $a_i \in F$ and scalars are elements of F. The addition is coordinate-wise, and multiplication by a scalar merely multiplies each coordinate by that scalar. In Exercise 40.3, you will verify the details of this.

To be more concrete, the set $\mathbb{Z}_3 \times \mathbb{Z}_3$ of ordered pairs with coordinates from the field $\mathbb{Z}_3$ is a vector space, over the field $\mathbb{Z}_3$.

▷ **Quick Exercise.** List all the elements of this vector space. ◁

Example 40.4

The set of polynomials $\mathbb{Q}[x]$ is a vector space over $\mathbb{Q}$. The vectors are polynomials and the scalars are rational numbers. (See Exercise 40.1.)

Example 40.5

Let $\mathbb{Q}_n[x]$ be the set of polynomials of degree no more than n with coefficients from $\mathbb{Q}$. This is a vector space over $\mathbb{Q}$. Here, vectors are polynomials of degree no more than n and scalars are rational numbers. This example illustrates the fact that in a vector space, there need be no multiplication of vectors, but only addition of vectors: $\mathbb{Q}_n[x]$ is closed under addition but not multiplication. (See Exercise 40.2.)

Example 40.6

$\mathbb{Q}(\sqrt{2}) = \{a + b\sqrt{2} : a, b \in \mathbb{Q}\}$, is a vector space over $\mathbb{Q}$. Here, vectors are elements from $\mathbb{Q}(\sqrt{2})$ and scalars are from $\mathbb{Q}$.

▷ **Quick Exercise.** This example bears certain similarities with Example 40.1; explain this. *Hint*: What kind of ring is $\mathbb{Q}(\sqrt{2})$? ◁

The point we've made above in Example 40.5 bears repeating: In many of these examples of vector spaces it is natural also to define multiplication of vectors, as in our first example $\mathbb{R}^3$. Indeed, in many cases the vector space is a ring or field in its own right. But when we look at these as vector spaces over some field, we look at their structure in a different way and multiplication of vectors is not one of the operations considered. This change of point of view is important to keep in mind.

We now prove some basic properties of vector spaces. These properties should seem familiar, both from your previous experience with ring and group theory, and also any previous experience you have had with vectors, either in calculus or linear algebra.

Theorem 40.1 *Let V be a vector space over F with $\mathbf{v} \in V$ and $r \in F$. Then, the additive inverse $-\mathbf{v}$ of $\mathbf{v}$ is unique. Also,* $0\mathbf{v} = \mathbf{0}$, $r\mathbf{0} = \mathbf{0}$, *and*

$$(-r)\mathbf{v} = r(-\mathbf{v}) = -(r\mathbf{v}).$$

Proof: Because $(V, +)$ is a group, the uniqueness of the additive inverse follows from Theorem 25.1.d.

Note the difference between the scalar 0 and the zero vector $\mathbf{0}$. The first identity says that multiplying any vector by the zero scalar yields the zero vector. But,

$$0\mathbf{v} = (0+0)\mathbf{v} = 0\mathbf{v} + 0\mathbf{v},$$

and so adding $-(0\mathbf{v})$, the additive inverse of $0\mathbf{v}$, to both sides, we get that $\mathbf{0} = 0\mathbf{v}$, as desired.

Note that in proving that $0\mathbf{v} = \mathbf{0}$, we used the fact that 0 was the additive identity of F and one of the distributive laws (v). To prove the next identity, we will use the fact that $\mathbf{0}$ is the additive identity of V and the other distributive law (vii). So,

$$r\mathbf{0} = r(\mathbf{0}+\mathbf{0}) = r\mathbf{0} + r\mathbf{0},$$

and so $\mathbf{0} = r\mathbf{0}$, as before.

To show that $(-r)\mathbf{v} = -(r\mathbf{v})$, we will show that $(-r)\mathbf{v}$ is the additive inverse of $r\mathbf{v}$. That is, it must be equal to $-(r\mathbf{v})$. We do this by adding $(-r)\mathbf{v}$ to $r\mathbf{v}$ and showing that the sum is $\mathbf{0}$. Doing so, we have

$$(-r)\mathbf{v} + r\mathbf{v} = (-r+r)\mathbf{v} = 0\mathbf{v} = \mathbf{0},$$

as desired. You can show in a similar manner that $r(-\mathbf{v}) = -(r\mathbf{v})$.

▷ **Quick Exercise.** Show that $r(-\mathbf{v}) = -(r\mathbf{v})$. ◁ □

The proofs of the results of Theorem 40.1 might seem familar, and should indeed be compared to the solutions to Exercises 6.1 and 6.2.

Chapter Summary

In this chapter we defined *vector space*, looked at a number of examples of vector spaces, and examined some elementary properties.

Warm-up Exercises

a. Consider the vectors $\mathbf{v} = (2, -1)$ and $\mathbf{w} = (2, 3)$ in $\mathbb{R}^2$. Compute the following, and draw diagrams to interpret these computations geometrically:

$$\mathbf{v}+\mathbf{w}, \quad \mathbf{v}-\mathbf{w}, \quad 2\mathbf{w}, \quad \frac{1}{2}\mathbf{v}, \quad -\mathbf{v}$$

b. Explain the difference between 0 and $\mathbf{0}$.

c. Let $V = \{f \in \mathbb{Z}_3[x] : \deg(f) \leq 2\}$; explain why this is a vector space over $\mathbb{Z}_3$. List all vectors and all scalars in this case.

d. Let $V = \{a + b\pi : a, b \in \mathbb{Q}\}$. Is this a vector space over $\mathbb{Q}$? Over $\mathbb{R}$?

e. Let V be a vector space over $\mathbb{R}$, and let $\mathbf{v} \in V$. Explain what $-\mathbf{v}$ and $(-1)\mathbf{v}$ mean, according to the definitions of our notation; why are they equal?

f. Is $\mathbb{C}$ a vector space over $\mathbb{Q}$? Is $\mathbb{Q}$ a vector space over $\mathbb{C}$?

g. Is $\mathbb{Z}[x]$ a vector space over $\mathbb{Z}$?

Exercises

1. Check Example 40.4: That is, prove that $\mathbb{Q}[x]$ is a vector space over $\mathbb{Q}$.

2. Check Example 40.5: That is, prove that $\mathbb{Q}_n[x]$ is a vector space over $\mathbb{Q}$, although it is not a subring of $\mathbb{Q}[x]$.

3. Check Example 40.3: That is, let F be a field, and let F^n be the set of n-tuples with entries from F. Prove that F^n is a vector space over F.

4. Let V be a vector space over the field F. Suppose that $r, s \in F$ and $\mathbf{0} \neq \mathbf{v} \in V$. Prove that if $r\mathbf{v} = s\mathbf{v}$, then $r = s$.

5. Let V be a vector space over the field F. Suppose that $0 \neq r \in F$. Define the function
$$\varphi_r : V \to V$$
by $\varphi_r(\mathbf{v}) = r\mathbf{v}$. Prove that φ_r is a one-to-one onto function that preserves addition. That is, φ_r is an *additive group isomorphism* from $(V, +)$ onto itself.

6. Let V be a vector space over the field F. Suppose that $r, s \in F$ and $\mathbf{v} \in V$. Prove that $r(s\mathbf{v}) = s(r\mathbf{v})$.

7. Show that $M_2(\mathbb{R})$, the two-by-two matrices with entries from $\mathbb{R}$, is a vector space over $\mathbb{R}$.

8. Show that $M_{m,n}(F)$, the m-by-n matrices with entries from the field F, is a vector space over F.

9. If F and E are fields with $F \subseteq E$, show that E is a vector space over F. This is an important example of a vector space in subsequent chapters.

10. Let $\mathbb{Q}(\sqrt[3]{2}) = \{a + b\sqrt[3]{2} + c\sqrt[3]{4} : a, b, c \in \mathbb{Q}\}$. (In Exercise 38.5 you showed this is a field.) Show that $\mathbb{Q}(\sqrt[3]{2})$ is a vector space over $\mathbb{Q}$.

11. Let V be the set of real-valued functions with addition defined by
$$(f + g)x = f(x) + g(x).$$
For $c \in \mathbb{R}$ and $f \in V$, the scalar multiple of f by c is $cf(x) = c(f(x))$. Show V is a vector space over $\mathbb{R}$.

12. Prove that $\mathbb{Z}$ is not a vector space over $\mathbb{Z}_p$, where p is a positive prime integer.

13. Prove that $\mathbb{Z}$ is not a vector space over $\mathbb{Q}$.

Chapter 41

Vector Spaces II

In the previous chapter we defined vector spaces. In this chapter we examine subsets of vector spaces that in some way generate the entire vector space. This idea gives rise to a way of measuring the size of a vector space.

41.1 Spanning Sets

If V is a vector space over F and
$$\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\} \subseteq V,$$
then a **linear combination** of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is the vector
$$a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n,$$
where the a_i are scalars. The collection of vectors that can be written as linear combinations of a given set of vectors $\mathcal{V}$ is the set **spanned** by $\mathcal{V}$.

Example 41.1

Consider the vectors $\mathbf{v} = (1, 0, 0)$ and $\mathbf{w} = (0, 1, 0)$ in $\mathbb{R}^3$. The vector
$$\left(\frac{1}{3}, -1, 0\right) = \frac{1}{3}\mathbf{v} + (-1)\mathbf{w},$$
and so is a linear combination of $\mathbf{v}$ and $\mathbf{w}$. Evidently, the set of all vectors in $\mathbb{R}^3$ spanned by $\mathbf{v}$ and $\mathbf{w}$ is exactly $\{(x, y, 0) : x, y \in \mathbb{R}\}$.

▷ **Quick Exercise.** Why is the set of vectors spanned by $\mathbf{v}$ and $\mathbf{w}$ equal to $\{(x, y, 0) : x, y \in \mathbb{R}\}$? ◁

Note that if we include the vector $(0, 0, 1)$, we then obtain all vectors in $\mathbb{R}^3$ as linear combinations of these three vectors.

Example 41.2

It is not so evident which vectors in $\mathbb{R}^3$ are spanned by $(1,0,-1)$ and $(3,2,1)$. We can certainly conclude that the set spanned by them consists of all vectors of the form

$$(x+3y, 2y, -x+y), \quad \text{where} \quad x, y \in \mathbb{R},$$

but what vectors are these? For example, does the vector $(1,1,1)$ belong to this set? If it did, there would be a simultaneous solution to the three equations

$$\begin{aligned} x+3y &= 1 \\ 2y &= 1 \\ -x+y &= 1. \end{aligned}$$

It's pretty easy to see that there is no such solution.

▷ **Quick Exercise.** Verify that there is no solution to this system of equations. ◁

Thus, $(1,1,1)$ is not a linear combination of $(1,0,-1)$ and $(3,2,1)$.

If every vector of V can be written as a linear combination of vectors from a subset $\mathcal{V}$ of V, we say that $\mathcal{V}$ **spans** (or **generates**) V. Note that $\mathcal{V}$ may be infinite or finite, but any given linear combination of vectors from $\mathcal{V}$ involves only a finite number of vectors.

Example 41.3

For instance, from our discussion above it is clear that

$$\{(1,0,0), (0,1,0), (0,0,1)\}$$

spans $\mathbb{R}^3$.

Example 41.4

Similarly,

$$\{(1,0,0), (2,3,0), (0,1,0), (-1,0,-1), (0,0,1)\}$$

spans $\mathbb{R}^3$ because

$$\begin{aligned}(x,y,z) = x(1,0,0) + 0(2,3,0) + y(0,1,0) + \\ 0(-1,0,-1) + z(0,0,1).\end{aligned}$$

Note that this spanning set contains more vectors than necessary. Furthermore, there may be more than one way of expressing a given vector as a linear combination of vectors in the spanning set. For instance, we can express $(-2,-2,-5)$ as

$$-2(1,0,0) + 0(2,3,0) - 2(0,1,0) + 0(-1,0,-1) - 5(0,0,1),$$

or as

$$2(1,0,0) - 1(2,3,0) + 1(0,1,0) + 2(-1,0,-1) - 3(0,0,1).$$

Example 41.5

Now let's consider the set

$$\{(1,0,0), (2,3,0), (-1,0,-1)\}.$$

It is not so obvious that this also spans $\mathbb{R}^3$. To show this, we consider a arbitrary vector (x,y,z) from $\mathbb{R}^3$ and show that we can express it as a linear combination of the three vectors in the set. That is, we wish to find scalars a_1, a_2, and a_3 such that

$$(x,y,z) = a_1(1,0,0) + a_2(2,3,0) + a_3(-1,0,-1).$$

This vector equation is equivalent to the following three linear equations, which we must solve simultaneously:

$$\begin{aligned} x &= a_1 + 2a_2 - a_3 \\ y &= \phantom{a_1 + {}} 3a_2 \\ z &= - a_3 \end{aligned}$$

By the usual techniques we find the solutions for a_1, a_2, and a_3 are

$$\begin{aligned} a_1 &= x - \frac{2}{3}y - z \\ a_2 &= \frac{1}{3}y \\ a_3 &= -z. \end{aligned}$$

So, for instance,

$$(3,1,-2) = \frac{13}{3}(1,0,0) + \frac{1}{3}(2,3,0) + 2(-1,0,-1).$$

Example 41.6

Consider the vector space $\mathbb{Q}[x]$ over $\mathbb{Q}$. Note that $\{1, x, x^2, x^3, \ldots\}$ spans $\mathbb{Q}[x]$.

▷ **Quick Exercise.** Show that $\{1, x, x^2, x^3, \ldots\}$ spans $\mathbb{Q}[x]$. ◁

In this case, we have an *infinite* spanning set.

A vector space is said to be **finite dimensional** if there is a finite set of vectors that spans the vector space. So, $\mathbb{R}^3$ is finite dimensional over $\mathbb{R}$.

The vector space $\mathbb{C}$ over $\mathbb{R}$ is finite dimensional because $\{1, i\}$ spans $\mathbb{C}$.

▷ **Quick Exercise.** Verify that the vector space $\mathbb{C}$ over $\mathbb{R}$ is finite dimensional. ◁

Likewise, $\mathbb{Q}(\sqrt{2})$ over $\mathbb{Q}$ is finite dimensional because $\{1, \sqrt{2}\}$ spans $\mathbb{Q}(\sqrt{2})$.

▷ **Quick Exercise.** Verify that $\mathbb{Q}(\sqrt{2})$ over $\mathbb{Q}$ is finite dimensional. ◁

However, $\mathbb{Q}[x]$ is not finite dimensional over $\mathbb{Q}$. In Example 41.6 we gave an infinite spanning set for $\mathbb{Q}[x]$; this alone does not suffice in showing that $\mathbb{Q}[x]$ is not finite dimensional—it may be that there exists some finite set of vectors that does indeed span $\mathbb{Q}[x]$. To show that $\mathbb{Q}[x]$ is not finite dimensional, we must show that *every* finite subset of $\mathbb{Q}[x]$ fails to span $\mathbb{Q}[x]$. This is Exercise 41.7. For the most part, we will confine our study here to finite dimensional vector spaces.

41.2 A Basis for a Vector Space

In the remainder of this chapter we will examine spanning sets that are minimal in the sense that removing any one vector from the set will result in a set that does not span the entire vector space.

The following is an important definition for us. A set of vectors

$$\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$$

is **linearly independent** if

$$a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n = \mathbf{0}$$

implies that $a_1 = a_2 = \cdots = a_n = 0$. A set of vectors that is not linearly independent is **linearly dependent**.

Example 41.7

We can easily see that $\{(1,0,0), (0,1,0), (0,0,1)\}$ is linearly independent in $\mathbb{R}^3$ because if

$$a_1(1,0,0) + a_2(0,1,0) + a_3(0,0,1) = (0,0,0), \text{ then}$$

$(a_1, a_2, a_3) = (0,0,0)$, or $a_1 = a_2 = a_3 = 0$. The set

$$\{(1,0,0),\ (2,3,0),\ (-1,0,-1)\}$$

is also linearly independent, which we can verify by showing that the resulting system of equations has a unique solution of $a_1 = a_2 = a_3 = 0$.

▷ **Quick Exercise.** Show that the above sysytem of equations has a unique solution of $a_1 = a_2 = a_3 = 0$. ◁

Example 41.8

By contrast,

$$\{(1,0,0),\ (2,3,0), (0,1,0),\ (-1,0,-1),\ (0,0,1)\}$$

is linearly dependent because

$$(0,0,0) = -2(1,0,0) + 1(2,3,0) - 3(0,1,0) + \\ 0(-1,0,-1) + 0(0,0,1).$$

(There are other possible values here for the a_i.)

▷ **Quick Exercise.** Find a different set of values for the a_i. ◁

Example 41.9

The set of vectors $\{1+\sqrt{2},\ 2-\sqrt{2}\}$ is linearly independent in the vector space $\mathbb{Q}(\sqrt{2})$ over $\mathbb{Q}$. But if we include the third vector $6\sqrt{2}$ in this set, it becomes dependent, because

$$2(1+\sqrt{2}) + (-1)(2-\sqrt{2}) - \frac{1}{2}(6\sqrt{2}) = 0.$$

The following theorem conveniently characterizes linearly independent sets of vectors:

Theorem 41.1 *$\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is a set of linearly dependent vectors in the vector space V if and only if one of the vectors from this set can be written as a linear combination of the others.*

Proof: If $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is linearly dependent then we can write

$$a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n = \mathbf{0},$$

where the a_i's are not all 0. By renumbering the vectors, if necessary, let's assume that a_1 is not zero. But then,

$$\mathbf{v}_1 = -\frac{a_2}{a_1}\mathbf{v}_2 - \frac{a_3}{a_1}\mathbf{v}_3 - \cdots - \frac{a_n}{a_1}\mathbf{v}_n,$$

as desired. Note that division by scalars is permissible, because they come from a field.

Conversely, suppose one of the vectors (say, $\mathbf{v}_1$) can be expressed as a linear combination of the others:

$$\mathbf{v}_1 = a_2\mathbf{v}_2 + a_3\mathbf{v}_3 + \cdots + a_n\mathbf{v}_n.$$

But then,

$$\mathbf{0} = -1\mathbf{v}_1 + a_2\mathbf{v}_2 + a_3\mathbf{v}_3 + \cdots + a_n\mathbf{v}_n.$$

And so the set $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is linearly dependent. □

Looking at this another way, this theorem says that a set of vectors is linearly independent if and only if no vector in the set can be expressed as a linear combination of the others. Thus, if a set of vectors $\mathcal{V}$ is linearly independent and spans the vector space V, no proper subset of $\mathcal{V}$ can span V. In particular, any vector removed from $\mathcal{V}$ cannot be written as a linear combination of the remaining vectors. In other words, $\mathcal{V}$ is a *minimal* spanning set for V. We have a name for such a spanning set: A **basis** for a vector space V is a linearly independent set of vectors that spans V.

▷ **Quick Exercise.** We have thus argued above that any *basis* is a *minimal spanning set.* The converse is also true: any minimal spanning set is linearly independent, and so a basis. What is the argument for this? ◁

Example 41.10

$\{(1,0,0), (0,1,0), (0,0,1)\}$ is a basis for $\mathbb{R}^3$, as is

$$\{(1,0,0), (2,3,0), (-1,0,-1)\}.$$

Example 41.11

$\{1, i\}$ is a basis for $\mathbb{C}$ over $\mathbb{R}$. Note in this case, where our basis has two elements, linear independence is particularly easy to check: By Theorem 41.1 we need only verify that neither of the vectors is a scalar (in this case, real) multiple of the other. This is certainly true for 1 and i.

Example 41.12

$\{1+i, \sqrt{2}+4i\}$ is a basis for $\mathbb{C}$ over $\mathbb{R}$. To check this requires a solution of two equations in two unknowns that we leave as Exercise 41.12. This example should make evident the fact that there are infinitely many distinct bases for $\mathbb{C}$ over $\mathbb{R}$.

Example 41.13

$\{1, \sqrt{2}\}$ is a basis for $\mathbb{Q}(\sqrt{2})$ over $\mathbb{Q}$.

In the Quick Exercise above, we emphasized the fact that being a linearly independent spanning set (that is, a basis) is actually equivalent to being a minimal spanning set. We can change perspective a bit, and also prove that being a basis is equivalent to being a *maximal independent set*; you will pursue this in Exercises 41.16 and 41.17.

The following is another important equivalent characterization of a basis, which you will prove in Exercise 41.8.

Theorem 41.2 *$\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is a basis for a vector space V if and only if every element of V can be uniquely written as a linear combination of the $\mathbf{v}_i$.*

The last important goal of this chapter is to show that for a finite dimensional vector space, all bases have the same number of vectors. (This is also true for infinite dimensional vector spaces, but the proof requires more sophisticated set theory than we wish to use here.) The number of elements in a basis for a vector space will then become an *invariant* of the space; that is, the number of elements in a basis is *independent* of the particular basis determined. The number of basis vectors consequently provides us with a good measure of the 'size' of the space. It is exactly this measure of size that we will need when we return to the theory of field extensions.

But first, we make a simple observation about linear combinations: If $\mathbf{v}$ is a linear combination of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, and each $\mathbf{v}_i$ is a linear combination of $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m$, then $\mathbf{v}$ is a linear combination of $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m$. To show this, simply write $\mathbf{v}$ as a linear combination of the $\mathbf{v}_i$ and substitute for each $\mathbf{v}_i$ the appropriate linear combination of the $\mathbf{w}_j$. (This is Exercise 41.15.)

41.3 Finding a Basis

Now suppose V is a finite dimensional vector space spanned by $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$. Then we will argue below that we can find a subset of $\mathcal{V}$ that is a basis for V. That is, *any* finite spanning set for a vector space contains a linearly independent set, which is then necessarily a basis.

We build this set in the following way. We may as well assume that V does not have zero dimension and none of the $\mathbf{v}_i$ is the zero vector. First, we start with $\mathbf{v}_1$. We add each $\mathbf{v}_i$ in succession until we find a $\mathbf{v}_i$ that is a linear combination of the previous $\mathbf{v}_j$, $j < i$. We discard this $\mathbf{v}_i$. We continue in this manner, adding or discarding vectors until we've run through the vectors in $\mathcal{V}$. Call the set of vectors that remains $\mathcal{B}$. This set also spans V because we have removed only those vectors that can be expressed as linear combinations of the vectors in $\mathcal{V}$ that come before it, and so, by Exercise 41.15, $\mathcal{B}$ also spans V. It remains to show that $\mathcal{B}$ is linearly independent.

By way of contradiction, suppose $\mathcal{B} = \{\mathbf{v}_{i_1}, \mathbf{v}_{i_2}, \ldots, \mathbf{v}_{i_k}\}$ is linearly dependent where $i_1, i_2, \ldots, i_k$ are indices of the vectors in $\mathcal{V}$ with $i_1 < i_2 < \cdots < i_k$. Then

$$0 = a_1\mathbf{v}_{i_1} + a_2\mathbf{v}_{i_2} + \cdots + a_k\mathbf{v}_{i_k},$$

where not all the a_k are zero. Pick the largest j with $a_j \neq 0$. Then, solving the above for $\mathbf{v}_{i_j}$, we can express $\mathbf{v}_{i_j}$ as a linear combination of vectors in $\mathcal{V}$ with smaller indices, which is a contradiction. Therefore, $\mathcal{B} \subseteq \mathcal{V}$ is linearly independent, and thus a basis for $\mathcal{V}$. We summarize this in the following theorem.

Theorem 41.3 *Every finite spanning set of a vector space contains a subset that is a basis for the vector space.*

We demonstrate this technique by finding a subset of

$$\{(1,0,0), (2,3,0), (0,1,0), (-1,0,-1), (0,0,1)\}$$

that is a basis for $\mathbb{R}^3$. We have already shown in Examples 41.4 and 41.8 that this set is linearly dependent and spans $\mathbb{R}^3$.

We start with the set $\{(1,0,0)\}$. Now we add the vector $(2,3,0)$, if $(2,3,0)$ is not a linear combination of $\{(1,0,0)\}$. Because this is clearly not the case (a linear combination of a single vector is a scalar multiple of that vector), we add it to our set, which is now $\{(1,0,0), (2,3,0)\}$.

The next vector is $(0,1,0)$. To see if $(0,1,0)$ is a linear combination of $(1,0,0)$ and $(2,3,0)$, we attempt to solve

$$(0,1,0) = a(1,0,0) + b(2,3,0),$$

and find that there is a solution of $a = -2/3$ and $b = 1/3$. Thus, $(0,1,0)$ is discarded and our set remains $\{(1,0,0), (2,3,0)\}$.

We next see if the vector $(-1,0,-1)$ is a linear combination of $(1,0,0)$ and $(2,3,0)$ by attempting to solve

$$(-1,0,-1) = a(1,0,0) + b(2,3,0).$$

But we easily see that there is no solution to this equation. (There is no way to make the third coordinate of the right-hand side non-zero.) Hence, we add this vector, making our set

$$\{(1,0,0), (2,3,0), (-1,0,-1)\}.$$

Finally, we consider the last vector and try to solve

$$(0,0,1) = a(1,0,0) + b(2,3,0) + c(-1,0,-1).$$

We find that there is a solution of $a = c = -1$ and $b = 0$. Thus, $(0,0,1)$ is discarded, leaving the set $\{(1,0,0), (2,3,0), (-1,0,-1)\}$ as a basis for $\mathbb{R}^3$. Note that a reordering of the vectors in our original set might produce a different basis for $\mathbb{R}^3$. You will see this explicitly in Exercise 41.14.

41.4 Dimension of a Vector Space

We are now ready to prove that any two bases for a finite dimensional vector space have the same number of vectors.

Theorem 41.4 *Suppose that V is a finite dimensional vector space with basis $\mathcal{W} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m\}$ and $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n, \ldots\}$ is another linearly independent spanning set. Then $\mathcal{V}$ has no more than m elements. Thus every basis for V has exactly m elements.*

Proof: We start by considering the spanning set

$$\{\mathbf{v}_1, \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m\}.$$

This set is linearly dependent, because $\mathbf{v}_1$ is a linear combination of the $\mathbf{w}_i$. So, it contains a subset that is a basis. (Of course, $\mathcal{W}$ is such a subset, but we're going to force $\mathbf{v}_1$ into a basis.) We proceed to find this subset using the method described above. Here, we first include $\mathbf{v}_1$ and then proceed through the $\mathbf{w}_i$ and will discard at least one of the $\mathbf{w}_i$. We claim that only one is discarded. Suppose that $\mathbf{w}_i$ and $\mathbf{w}_j$ are both discarded, where $i < j$. Then,

$$\mathbf{w}_i = a_0\mathbf{v}_1 + a_1\mathbf{w}_1 + \cdots + a_{i-1}\mathbf{w}_{i-1}.$$

Now $a_0 \neq 0$, else we've expressed $\mathbf{w}_i$ as a linear combination of the vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{i-1}$, contradicting the linear independence of $\mathcal{W}$. So we can write

$$\mathbf{v}_1 = -\frac{a_1}{a_0}\mathbf{w}_1 - \frac{a_2}{a_0}\mathbf{w}_2 - \cdots - \frac{a_{i-1}}{a_0}\mathbf{w}_{i-1} + \frac{1}{a_0}\mathbf{w}_i. \tag{41.1}$$

Likewise, we can write $\mathbf{w}_j$ as

$$\mathbf{w}_j = b_0\mathbf{v}_1 + b_1\mathbf{w}_1 + \cdots + b_{j-1}\mathbf{w}_{j-1}. \tag{41.2}$$

But if we substitute expression (41.1) for $\mathbf{v}_1$ into equation (41.2), we will have expressed $\mathbf{w}_j$ as a linear combination of the other vectors in $\mathcal{W}$, again contradicting the linear independence of $\mathcal{W}$.

Thus, the linearly independent subset of

$$\{\mathbf{v}_1, \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m\}$$

has precisely one fewer vector in it, and that discarded vector was one of the $\mathbf{w}_i$. By renumbering, if necessary, let's suppose that the discarded vector was $\mathbf{w}_1$. So, the new basis we obtained is $\{\mathbf{v}_1, \mathbf{w}_2, \mathbf{w}_3, \ldots, \mathbf{w}_m\}$.

Now add $\mathbf{v}_2$ to this basis so that our set is now

$$\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{w}_2, \ldots, \mathbf{w}_m\}.$$

Again, this is a spanning set that is linearly dependent and so contains a linearly independent subset which we find by the now familiar method. Note that neither $\mathbf{v}_1$ nor $\mathbf{v}_2$ will be discarded in the process, because $\mathbf{v}_1, \mathbf{v}_2$ is a linearly independent set. So, the discarded vector will be one of the $\mathbf{w}_i$. (There will be only one discarded by an argument similar to the one above.) Again, by renumbering if necessary, we assume that $\mathbf{w}_2$ was the one discarded, leaving the set $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{w}_3, \mathbf{w}_4, \ldots, \mathbf{w}_m\}$ as the basis.

We repeat this process, successively adding $\mathbf{v}_3, \mathbf{v}_4, \ldots,$ and $\mathbf{v}_n$ and each time discarding one of the $\mathbf{w}_i$ to get yet another basis for V. We will never run out of the $\mathbf{w}_i$'s before exhausting the $\mathbf{v}_j$'s, because then we would have a basis for V consisting of a proper subset of $\mathcal{V}$. (This can't be, because by the Quick Exercise in Section 41.2 a basis is a *minimal* spanning set.) This means that $\mathcal{V}$ must actually be a finite set, with no more than m elements.

By interchanging the roles of $\mathcal{V}$ and $\mathcal{W}$, and repeating the above argument, we obtain that these two sets have the same (finite) number of elements. □

The **dimension** of a given vector space is the number of vectors in any basis (this definition for our purposes applies only for finite-dimensional vector spaces, although it can be extended to the infinite case).

Example 41.14

The dimension of the vector space $\mathbb{R}^3$ over $\mathbb{R}$ is 3.

Example 41.15

The vector space $\mathbb{C}$ over $\mathbb{R}$ has dimension 2, and the vector space $\mathbb{Q}(\sqrt{2})$ over $\mathbb{Q}$ has dimension 2.

Example 41.16

Consider the vector space

$$V = \{f \in \mathbb{Z}_3[x] : \deg(f) \leq 2\}$$

over $\mathbb{Z}_3$, which we considered in Exercise 40.c. This vector space has dimension 3, because $\{1, x, x^2\}$ is a basis.

▷ **Quick Exercise.** Check this. ◁

We can easily obtain two important corollaries from Theorem 41.4:

Corollary 41.5 *Suppose that V is a finite dimensional vector space, and $\mathcal{U}$ is a linearly independent subset of V. Then $\mathcal{U}$ has finitely many elements, and there is a basis $\mathcal{W}$ for V with $\mathcal{U} \subseteq \mathcal{W}$.*

Proof: Choose a basis $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ for V with n elements. Now if $\mathcal{U}$ is already a spanning set, then it is a basis, and by the previous theorem it has exactly n elements.

If $\mathcal{U}$ is not a spanning set, we run through the vectors of $\mathcal{V}$ one at a time as follows, to build the basis $\mathcal{W}$ containing $\mathcal{U}$: Begin by letting $\mathcal{W} = \mathcal{U}$. For step i, check whether $\mathbf{v}_i$ is spanned by $\mathcal{W}$. If it is, we add nothing to $\mathcal{W}$. If, however, $\mathbf{v}_i$ is not in the span of $\mathcal{W}$, we add it to $\mathcal{W}$; the set $\mathcal{W}$ remains linearly independent. After n steps, we will have augmented $\mathcal{U}$, building a larger set $\mathcal{W}$ that remains linearly independent but now must span V (because $\mathcal{V}$ does). But the previous theorem asserts that this set is finite, with exactly n elements. Thus the original set $\mathcal{U}$ must have been finite, and it is indeed a subset of a basis for V. □

Corollary 41.6 *Every set in an n-dimensional vector space with more than n elements is linearly dependent.*

Proof: Any linearly independent set can by the previous corollary be augmented to obtain a basis for the vector space. But any basis for the vector space has no more than n elements. □

Chapter Summary

In this chapter we defined *linear combination*, *spanning set*, *linear independence*, *linear dependence*, and *basis*.

We showed that any spanning set contains a subset that is a basis and that all bases for a given finite dimensional vector space have the same number of vectors. The number of vectors in a basis for a vector space is the *dimension* of the vector space.

Warm-up Exercises

a. Give examples of the following, or explain why such an example cannot exist. Be sure to specify explicitly the vector space, the field of scalars, and the vectors required:

(a) A spanning set that isn't a basis.

(b) A basis that isn't a spanning set.

(c) Two distinct bases for the same vector space.

(d) Two distinct bases for the same vector space, one of which is a proper subset of the other.

(e) A set of three linearly dependent vectors, any two of which are independent.

(f) A set of four linearly dependent vectors, any three of which are independent.

b. Why could the zero vector never belong to a basis (for a vector space with more than one element)?

c. What is the dimension of $\mathbb{C}$ as a vector space over $\mathbb{R}$? What is the dimension of $\mathbb{C}$ as a vector space over $\mathbb{C}$?

Exercises

1. Prove that $\sqrt{2}$ and $\sqrt{3}$ are linearly independent elements of the vector space $\mathbb{R}$, over $\mathbb{Q}$.

2. (Here, we extend Exercise 1.) Prove that $\{1, \sqrt{2}, \sqrt{3}\}$ is an independent subset of the vector space $\mathbb{R}$ over $\mathbb{Q}$. Is it a basis for $\mathbb{R}$?

3. Consider the set V, consisting of all polynomials of degree 0 or 1, with coefficients from $\mathbb{C}$. Prove that this is both a vector space over $\mathbb{C}$, and over $\mathbb{R}$. Find a basis in each case. Does its dimension stay the same, when we change the scalar field?

4. Find a basis for the vector space given in Exercise 40.7.

5. Find a basis for the vector space given in Exercise 40.8.

6. Give a basis for the vector space given in Exercise 40.10.

7. Show that $\mathbb{Q}[x]$ over $\mathbb{Q}$ is not finite dimensional by showing that no finite subset of $\mathbb{Q}[x]$ is a spanning set.

8. Prove Theorem 41.2.

9. (a) Define a **subspace** of a vector space.

   (b) Let $\mathcal{S}$ be a subset of the vector space V. Show that the set of linear combinations of vectors from $\mathcal{S}$ is a subspace of V. (Note that this subspace is all of V if and only if $\mathcal{S}$ spans V.)

10. Determine which of the following are subspaces of the vector space $M_2(\mathbb{R})$ discussed in Exercise 40.7.

    (a) Matrices with determinant 1.

    (b) Matrices with determinant 0.

    (c) The diagonal matrices.

    (d) Matrices with integer entries.

    (e) Matrices of the form

$$\begin{pmatrix} a & a+b \\ b & 0 \end{pmatrix}.$$

    (f) Matrices with first row, first column entry of 0.

11. Determine which of the following are subspaces of the vector space of all real-valued functions discussed in Exercise 40.11.

    (a) All differentiable functions.

    (b) All functions f with $f(2) = 0$.

    (c) All linear functions. (That is, functions of the form $f(x) = a + bx$.)

    (d) All polynomial functions.

12. Check Example 41.12: that is, prove that $\{1 + i, \sqrt{2} + 4i\}$ is a spanning set for $\mathbb{C}$ over $\mathbb{R}$.

13. Find a subset of the set

$$\{(1,1,0),(2,3,-1),(0,1,-1),(1,4,2),(0,0,3)\}$$

that is a basis for $\mathbb{R}^3$ using the technique described in this chapter.

14. Find a subset of the set

$$\{(-1,0,-1),(0,0,1),(1,0,0),(2,3,0),(0,1,0)\}$$

that is a basis for $\mathbb{R}^3$ using the technique in this chapter; this is the same set we used in Section 41.3 to demonstrate this technique. Take the vectors in the order given here, to see explicitly that we obtain a different subset than in the text.

15. Let V be a vector space. Show that if $\mathbf{v}$ is a linear combination of the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \in V$ and each $\mathbf{v}_i$ is in turn a linear combination of $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m$, then $\mathbf{v}$ is a linear combination of $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m$.

16. Prove that a basis for a vector space V is a *maximal independent set*: That is, it is a set $\mathcal{B} \subseteq V$ of independent vectors, so that $\mathcal{B} \cup \{w\}$ is a dependent set, whenever $w \in V \backslash \mathcal{B}$.

17. Prove that a maximal independent set of vectors of a vector space is necessarily a basis (that is, prove the converse of Exercise 16).

18. Use the technique in the proof of Corollary 41.5 to build a basis for $\mathbb{R}^4$ that contains the linearly independent set

$$\{(1,0,-1,0),\ (-2,1,2,0)\}.$$

Chapter 42

Field Extensions and Kronecker's Theorem

Consider the polynomial $x^2 + 1 \in \mathbb{R}[x]$. This polynomial clearly has no roots in $\mathbb{R}$, but we know that there exists a larger field—namely, $\mathbb{C}$—in which this polynomial does have roots. In this chapter we develop a general method for constructing bigger fields, with the aim of being able to further factor polynomials.

42.1 Field Extensions

A field E is an **extension field** of a field F if $E \supseteq F$. That is, E is an extension field of F exactly if F is a subfield of E; in this situation, we call the field F the **base field**. For example, $\mathbb{R}$ is an extension field of $\mathbb{Q}$, $\mathbb{C}$ is an extension field of both $\mathbb{R}$ and $\mathbb{Q}$, and $\mathbb{Q}(\sqrt{2})$ is an extension field of $\mathbb{Q}$.

If E is an extension field of a field F and $\alpha \in E$ is a root of a polynomial in $F[x]$, we say α is **algebraic over** F. Otherwise, α is **transcendental over** F.

Example 42.1

Note first that if E is an extension field of a field F and $\alpha \in F$, then trivially α is algebraic over F because we can easily find a polynomial $p \in F[x]$ such that $p(\alpha) = 0$.

▷ **Quick Exercise.** Find such a polynomial. ◁

Example 42.2

The complex field $\mathbb{C}$ is an extension field of $\mathbb{Q}$, and because $\sqrt{2}$ is a root of $x^2 - 2 \in \mathbb{Q}[x]$, $\sqrt{2}$ is algebraic over $\mathbb{Q}$. Also, i is algebraic over $\mathbb{Q}$ because i is a root of $x^2 + 1 \in \mathbb{Q}[x]$.

Example 42.3

The real field $\mathbb{R}$ is an extension field of $\mathbb{Q}$. As we mentioned in Chapter 39 (see Theorem 39.5), the numbers e and π are transcendental over $\mathbb{Q}$; that is, neither e nor π are roots of any polynomial in $\mathbb{Q}[x]$. These famous and difficult results are due to Hermite and Lindemann, respectively.

42.2 Kronecker's Theorem

We now prove an extremely important theorem due to Leopold Kronecker, an eminent German mathematician of the latter part of the 19th century. Given a polynomial f in $F[x]$, is there a field extension of the field F that contains a root of f? Kronecker's Theorem provides us with an affirmative answer to this question.

Note that this theorem will apply to any polynomial over any field. Furthermore, while the description of the extension field will seem somewhat abstract—we will find that it is a ring of cosets of $F[x]$—later theorems will allow us to explicitly describe some of these extension fields in a way that is more palatable.

Theorem 42.1 Kronecker's Theorem *Let F be a field and f a polynomial in $F[x]$ of degree at least 1. Then there exists an extension field E of F and an $\alpha \in E$ such that $f(\alpha) = 0$.*

Proof: If f has a root in the field F, then F itself is the required extension field. So we may as well assume that f has no roots in F. From Chapter 9 we know that f can be factored into irreducible polynomials in $F[x]$; note that each of these polynomials has degree at least two. Let p be one of these irreducible factors. We will show that there is an extension field E of F and an $\alpha \in E$ with $p(\alpha) = 0$.

From Theorem 13.3 we know that because p is irreducible in $F[x]$, the principal ideal $\langle p \rangle$ is maximal in $F[x]$. Hence, from Theorem 20.1, $F[x]/\langle p \rangle$ is a field. This is our field E.

We now need to show two things. First, we must show that E can be viewed as an extension field of F; that is, we will obtain an isomorphism from F into E. Second, we need to find an element of E that is a root of p.

We show that there is an isomorphism from F into $F[x]/\langle p \rangle$ by considering the function $\psi : F \longrightarrow F[x]/\langle p \rangle$ defined by $\psi(a) = \langle p \rangle + a$, where $a \in F$. That is, an element in F is mapped to its coset in $F[x]/\langle p \rangle$. (Note this function makes a subtle shift in viewing a. First, we think of a as an element of F while the image of a treats a as a polynomial in $F[x]$.) It is evident that ψ is a ring homomorphism, because of the way the ring operations are defined on the ring of cosets.

▷ **Quick Exercise.** Check that ψ is a ring homomorphism. ◁

We now claim that ψ is one-to-one; to show this, we will verify that its kernel is $\{0\}$. For that purpose, suppose that

$$\psi(a) = \langle p \rangle + a = \langle p \rangle + 0.$$

Then $a \in \langle p \rangle$. But the ideal $\langle p \rangle$ is precisely the set of multiples of p and because p has degree at least 1, so do all the non-zero elements of $\langle p \rangle$. But a (viewed as a polynomial in $F[x]$, of course) has degree 0, and so a must be the zero polynomial. That is, the kernel of ψ is trivial, and so ψ is one-to-one. We consequently may as well assume that F is a subfield of the field $F[x]/\langle p \rangle$.

Finally, we must find an element of $F[x]/\langle p \rangle$ that is a root of our polynomial p. Consider the element $\alpha = \langle p \rangle + x$. Suppose $p = a_0 + a_1 x + \cdots + a_n x^n$. Then

$$p(\alpha) = a_0 + a_1(\langle p \rangle + x) + \cdots + a_n(\langle p \rangle + x)^n \in F[x]/\langle p \rangle.$$

But we do arithmetic in $F[x]/\langle p \rangle$ by choosing any coset representative we wish. If we pick x as the coset representative of $\langle p \rangle + x$, then

$$p(\alpha) = \langle p \rangle + a_0 + a_1 x + \cdots + a_n x^n = \langle p \rangle + p = \langle p \rangle + 0.$$

But $\langle p \rangle + 0$ is the zero element of the field $F[x]/\langle p \rangle$, and so we have found our desired element of $F[x]/\langle p \rangle$. □

Let's perform the construction of the theorem in two specific cases.

Example 42.4

Consider the polynomial $f = x^3 - 5 \in \mathbb{Q}[x]$; it has no roots in $\mathbb{Q}$.

▷ **Quick Exercise.** Verfiy that $f = x^3 - 5$ has no roots in $\mathbb{Q}$? ◁

Thus, f is irreducible in $\mathbb{Q}[x]$, and so $\langle f\rangle$ is a maximal ideal. This makes $E = \mathbb{Q}[x]/\langle f\rangle$ a field. Furthermore, we may identify $\mathbb{Q}$ with the subfield

$$\{\langle x^3-5\rangle + q : q \in \mathbb{Q}\}$$

of E. The element $\langle f\rangle + x$ of E is then the required root for the polynomial f, because

$$f(\langle f\rangle + x) = (\langle f\rangle + x)^3 - 5 = \langle f\rangle + x^3 - 5 = \langle f\rangle + f = \langle f\rangle + 0.$$

Example 42.5

Consider the polynomial $p = x^2+x+1$ in $\mathbb{Z}_2[x]$. It has no roots in $\mathbb{Z}_2$, because the only possibilities are 0 and 1, and neither works. Thus, p is irreducible in $\mathbb{Z}_2[x]$, and so $\langle p\rangle$ is a maximal ideal. This makes $E = \mathbb{Z}_2[x]/\langle p\rangle$ a field. Furthermore, we may identify $\mathbb{Z}_2$ with the subfield

$$\{\langle p\rangle + 0, \langle p\rangle + 1\}$$

of E. The element $\langle p\rangle + x$ of E is then the required root for the polynomial p.

In the next chapter, we prove a theorem that gives a much more explicit description of $F[x]/\langle p\rangle$. For example, x^2+1 is irreducible in $\mathbb{R}[x]$, and so x^2+1 has a root in the field $\mathbb{R}[x]/\langle x^2+1\rangle$. But we also know that

$$\mathbb{R}(i) = \{a + bi : a, b \in \mathbb{R}\} = \mathbb{C},$$

is a field in which i is a root of x^2+1. We will see that these two fields are in fact isomorphic.

42.3 The Characteristic of a Field

It turns out that every field has a unique smallest subfield, which we call the *prime* subfield. This fact provides us with many nice examples of field extensions.

To show that this is true, suppose that K is an arbitrary field, with multiplicative identity 1. Consider what happens when we repeatedly add 1 to itself: We get the elements

$$1, \quad 1+1=2, \quad 1+2=3 \quad \cdots.$$

In the case of a field like $\mathbb{R}$, this process goes on forever, and we obtain in this way all the positive integers. But in the case of a field like $\mathbb{Z}_{13}$, we obtain 0, once we have added 1 to itself thirteen times. The **characteristic** of the field K is the least positive integer n so that $n\cdot 1 = 0$ (if such exists). If no such n exists, we set the characteristic to 0. (If you did Exercise 8.11, you have already encountered this concept, in a more general context.)

Example 42.6

The characteristic of such fields as $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{Q}(\sqrt{2})$ is zero, while the characteristic of the field $\mathbb{Z}_p$ is p.

Example 42.7

Consider now the field E constructed in Example 42.5. What is the characteristic of this field? The multiplicative identity in E is the coset $\langle p\rangle + 1$, and clearly

$$(\langle p\rangle + 1) + (\langle p\rangle + 1) = \langle p\rangle + 0.$$

This means that the characteristic of E is 2.

Although we defined the characteristic in terms of the number of times 1 could be added to itself, notice that the properties of arithmetic in a field means that we can in fact define characteristic in terms of *any* non-zero element. For if $n\cdot 1 = 0$, then

$$n\cdot r = (r + r + \cdots + r) = r(1 + 1 + \cdots + 1) = r(n\cdot 1),$$

and so $n \cdot 1 = 0$ if and only if $n\cdot r = 0$.

▷ **Quick Exercise.** Why is the fact that a field has no zero divisors relevant to this remark? ◁

Notice that the only values for the characteristic we have obtained in our examples are 0 and positive prime integers. It is easy to prove that these are the only possible cases:

Theorem 42.2 *The characteristic of any field is either 0 or a positive prime integer.*

Proof: Suppose that a field F has characteristic $n > 0$, and that n has a non-trivial factorization $n = rs$. Then

$$(r1)(s1) = (1 + \cdots + 1)(1 + \cdots + 1) = (1 + 1 + \cdots + 1) = (rs)1 = 0$$

(where we have added 1 to itself r times, s times, and then rs times, in the previous computation). But because a field has no zero divisors, this means that either $r1$ or $s1$ is zero, and this contradicts the minimality of n. □

Suppose now that F is a field with characteristic zero. We define a function $\iota : \mathbb{Q} \to F$ as follows: Given $a/b \in \mathbb{Q}$ (where a and b are integers, with $b \neq 0$), let $\iota(a/b) = (a \cdot 1)/(b \cdot 1)$. Notice first that the quotient makes sense, because in a field of characteristic zero, $b \cdot 1 \neq 0$. It is now easy to prove that this function preserves both addition and multiplication, and so is a ring homomorphism.

▷ **Quick Exercise.** Check these two facts. ◁

But the kernel of ι is evidently $\{0\}$, because if $a \in \mathbb{Z}$ and $a \neq 0$, then $a \cdot 1 \neq 0$ (because the field is characteristic zero).

Thus, any field of characteristic zero contains (an isomorphic copy of) the field $\mathbb{Q}$. And so, such a field is necessarily a vector space over the rational numbers. This copy of $\mathbb{Q}$ is the unique smallest subfield of F. (You will prove this fact in Exercise 42.7a.) This copy of $\mathbb{Q}$ is called the **prime subfield** of F.

Example 42.8

In Example 42.4, we identified the prime subfield of the field E as the set $\{\langle f \rangle + q : q \in \mathbb{Q}\}$.

What about fields with non-zero characteristic? If a field F has characteristic p, where p is a positive prime integer, then we define $\iota : \mathbb{Z}_p \to F$ by setting $\iota([1]) = 1$, and extending this function additively. That is, we set $\iota([2]) = 1 + 1$, and so forth. This function is also a one-to-one ring homomorphism (see Exercise 42.6), and so any field with characteristic p contains (an isomorphic copy of) the field $\mathbb{Z}_p$. This copy of $\mathbb{Z}_p$ is the unique smallest subfield of F. (Again: You will prove this fact in Exercise 42.7b.) This copy of $\mathbb{Z}_p$ is called the **prime subfield** of F.

Example 42.9

In Example 42.5, we identified the prime subfield of the field E as the set $\{\langle p \rangle + 0, \langle p \rangle + 1\}$.

▷ **Quick Exercise.** What can you say about a field which has dimension 1 as a vector space over its prime field? ◁

We can view the prime subfield of a field as the result of applying all the field operations to the multiplicative identity 1. If the field has characteristic zero, this gives (an isomorphic copy of) $\mathbb{Q}$; if the field has characteristic p, this gives (an isomorphic copy of) $\mathbb{Z}_p$. Note that if we know the characteristic of a field F and $E \supseteq F$ is any extension field, then E will automatically have the same characteristic as F, because its prime subfield will be the same.

Alternatively, we could do the following: Given a field K, consider the set of all subfields $E \subseteq K$; this set is non-empty, because K itself is such a subfield. Now consider the intersection of all such subfields. The result is certainly non-empty, because it necessarily contains the element 1. It is also a field (as you prove in Exercise 42.8). It is necessarily the smallest subfield of K, which is then the prime subfield.

Chapter Summary

In this chapter we defined the phrase *E is an extension field of a field F*. In this case we also defined what it means for an element of E to be *algebraic over F* and *transcendental over F*.

Next, we proved *Kronecker's Theorem* which says that for a field F, if $f \in F[x]$, then there exists an extension field of F that contains a root of f.

Finally, we defined the *characteristic* of a field and showed that all fields have characteristic 0, or p, where p is a positive prime integer.

Warm-up Exercises

a. In each of the following cases, show that the given element is algebraic over the given base field, by finding an appropriate polynomial with the element as a root.

(a) $\sqrt[3]{2}$ is algebraic over $\mathbb{Q}$.

(b) $\sqrt[3]{2}$ is algebraic over $\mathbb{R}$.

(c) $1+\sqrt{2}$ is algebraic over $\mathbb{Q}$.

(d) $\sqrt{1+\sqrt{2}}$ is algebraic over $\mathbb{Q}$.

(e) $1+2i$ is algebraic over $\mathbb{Q}$.

(f) $\pi + i$ is algebraic over $\mathbb{R}$.

b. If F is a field, is F an extension field of F? If so, does F have any transcendental elements over F?

c. What extension field do we get if we apply Kronecker's Theorem 42.1 to a degree 1 polynomial?

d. Recall the Fundamental Theorem of Algebra 9.1, which asserts that all non-constant polynomials in $\mathbb{C}[x]$ can be factored completely into linear factors. What does this say about finding elements outside $\mathbb{C}$ that are algebraic over $\mathbb{C}$?

e. Give examples of the following, or else explain why no such example exists:

(a) An element in $\mathbb{C}$ transcendental over $\mathbb{R}$.

(b) An element in $\mathbb{C}$ transcendental over $\mathbb{Q}$.

(c) An element in $\mathbb{R}$, but not in $\mathbb{Q}(\sqrt{2})$, which is algebraic over $\mathbb{Q}(\sqrt{2})$.

f. For each of the following fields, identify its characteristic, and determine its prime subfield:

$$\mathbb{C},\quad \mathbb{Q},\quad \mathbb{Q}(\sqrt{2}),\quad \mathbb{Z}_{11},\quad \mathbb{Q}[x]/\langle x^3-2\rangle,\quad \mathbb{Z}_3[x]/\langle x^2+x+2\rangle.$$

g. "Every polynomial of degree at least 1 has a root." Discuss the truth of this statement.

h. Why is cos 20° algebraic over $\mathbb{Q}$? *Hint:* Find the appropriate result in Chapter 39.

i. Give some examples of real numbers that are algebraic over $\mathbb{Q}$, but not constructible.

Exercises

1. Let F be a field and $f \in F[x]$ a polynomial of degree at least 1. Prove that there exists a field extension E of F in which f can be factored into linear factors. Such a field extension is called a **splitting extension for f**.

2. In the proof of Kronecker's Theorem 42.1, we prove that the function $\psi : F \to F[x]/\langle p\rangle$ is a one-to-one ring homomorphism. Show that ψ is onto if and only if the irreducible polynomial p is linear.

3. If R is a commutative ring with unity, and R is a subring of a field F, we can speak of the elements in F that are **algebraic** over R, meaning simply those elements of the field which are roots of polynomials from $R[x]$.

   (a) Explain why all rational numbers are algebraic over $\mathbb{Z}$.

   (b) Prove that the set of elements of $\mathbb{R}$ that are algebraic over $\mathbb{Z}$ is the same as the set of elements of $\mathbb{R}$ which are algebraic over $\mathbb{Q}$.

4. Suppose that $\alpha \in \mathbb{C}$ is an algebraic number over $\mathbb{Q}$, and $r \in \mathbb{Q}$. Prove that $\alpha + r$ and $r\alpha$ are also algebraic over $\mathbb{Q}$. (Actually, this exercise is a special case of Exercise 44.10, where you will show that the set $\mathbb{A}$ of complex numbers algebraic over $\mathbb{Q}$ is a field, called the **field of algebraic numbers**. Why is our exercise a special case?)

5. Prove that $\sin 1°$ is algebraic over $\mathbb{Q}$.

6. Let F be a field with characteristic p. Prove that the function $\iota : \mathbb{Z}_p \to F$ discussed in Section 42.3 is a one-to-one ring homomorphism.

7. (a) For a field F with characteristic zero, we defined in Section 42.3 a one-to-one ring homomorphism $\iota : \mathbb{Q} \to F$. Prove that $\iota(\mathbb{Q})$ is the smallest subfield of F.

(b) For a field F with characteristic p, we defined in Section 42.3 a one-to-one ring homomorphism $\iota : \mathbb{Z}_p \to F$. (You completed the proof of this in Exercise 6.) Prove that $\iota(\mathbb{Z}_p)$ is the smallest subfield of F.

8. Let K be a field, and consider the set of all subfields of K (this set may well have infinitely many elements). Consider the intersection of all these subfields. Prove that the result is a field, by checking explicitly the field axioms.

Chapter 43

Algebraic Field Extensions

In this chapter we give a nicer description for certain field extensions. This will in particular allow us to exhibit some new finite fields. We will focus our attention on field extensions in which the additional elements are algebraic over the base field.

43.1 The Minimal Polynomial for an Element

As we've seen, $\mathbb{R}$ is an extension field of $\mathbb{Q}$ and $\sqrt{2}$ is algebraic over $\mathbb{Q}$ because $\sqrt{2}$ is a root of $x^2 - 2 \in \mathbb{Q}[x]$. Of course, $\sqrt{2}$ is a root of many other polynomials in $\mathbb{Q}[x]$; for instance, $3x^2 - 6$, $(x^2 - 2)(x + 1)$, and $(x^2 - 2)(x^5 + 2x - 3)$. But every polynomial listed is a multiple of the irreducible polynomial $x^2 - 2$. This is in fact true in general, as the next important theorem shows.

Theorem 43.1 *If E is an extension field of a field F and $\alpha \in E$ is algebraic over F, then there exists an irreducible polynomial $p \in F[x]$ such that $p(\alpha) = 0$. Furthermore, if $f \in F[x]$ and $f(\alpha) = 0$, then p divides f.*

Proof: Because α is algebraic over F, there is at least one polynomial in $F[x]$ with α as a root. Let $p \in F[x]$ be a non-zero polynomial of minimal degree such that $p(\alpha) = 0$. We claim that p is irreducible in $F[x]$. For if $p = tq$ where t and q are polynomials in $F[x]$, then $0 = p(\alpha) = t(\alpha)q(\alpha)$, and so α would be a root of at least one of t and q. Let's assume $t(\alpha) = 0$. But p is of minimal degree among those polynomials with α as a root. Hence, $\deg(t) = \deg(p)$, and so q is a constant polynomial, as required.

It remains to show that p divides every polynomial $f \in F[x]$, where $f(\alpha) = 0$. By the Division Theorem we can write $f = pq + r$ where

$q, r \in F[x]$ and $\deg(r) < \deg(p)$. But

$$0 = f(\alpha) = p(\alpha)q(\alpha) + r(\alpha) = 0q(\alpha) + r(\alpha) = r(\alpha).$$

Therefore, α is a root of r. But because $\deg(r) < \deg(p)$ and p was of minimal degree among those polynomials with α as a root, r must be the zero polynomial. Thus, p divides f. □

Note that the above proof shows that the irreducible polynomial in $F[x]$ with α as a root is unique up to constant multiple. So, there exists a unique *monic* polynomial in $F[x]$ that is irreducible with α as a root. (Recall that a monic polynomial is one whose leading coefficient is 1.) We call this polynomial the **minimal polynomial of α over F**, and the degree of this polynomial the **degree of α over F**.

Example 43.1

The minimal polynomial for $\sqrt{2}$ over $\mathbb{Q}$ is $x^2 - 2$. Thus, $\sqrt{2}$ is of degree 2 over $\mathbb{Q}$.

Example 43.2

The minimal polynomial for i over $\mathbb{R}$ is x^2+1. Thus, i is of degree 2 over $\mathbb{R}$. (The same can be said if $\mathbb{R}$ is replaced by $\mathbb{Q}$, because $x^2 + 1 \in \mathbb{Q}[x]$.)

Example 43.3

The minimal polynomial for $\sqrt[3]{2}$ over $\mathbb{Q}$ is $x^3 - 2$. Thus, $\sqrt[3]{2}$ is of degree 3 over $\mathbb{Q}$.

Example 43.4

We claim that $\sqrt{2+\sqrt{2}}$ has minimal polynomial $x^4 - 4x^2 + 2$ over $\mathbb{Q}$ and so is of degree 4 over $\mathbb{Q}$.

▷ **Quick Exercise.** Check that $\sqrt{2+\sqrt{2}}$ is indeed a root of $x^4 - 4x^2 + 2$. ◁

But $x^4 - 4x^2 + 2$ is clearly irreducible over $\mathbb{Z}[x]$, by Eisenstein's criterion 5.7, and so is irreducible over $\mathbb{Q}[x]$, by Gauss's Lemma 5.5.

43.2 Simple Extensions

Suppose E is an extension field of the field F and $\alpha \in E$. We wish to obtain the smallest subfield of E that contains both α and all the elements of F. To do this, we consider the set of all such subfields of E. There is at least one such subfield—namely, E itself; there may be infinitely many others. We now consider

$$F(\alpha) = \cap\{\text{fields } K : F \subseteq K \subseteq E, \alpha \in K\}.$$

This set clearly contains the elements of F and also α. But is it a subfield? To show that this set is closed under subtraction, choose $a, b \in F(\alpha)$. Then $a, b \in K$, for each of the subfields K containing F and α. But then because K is a subfield, $a - b \in K$, for every such K. Thus, $a - b$ must be an element of the intersection of all such K.

▷ **Quick Exercise.** The rest of the proof that $F(\alpha)$ is a subfield follows the same lines; complete this proof. Note that this argument is essentially the same as that for Exercise 42.8. ◁

Because $F(\alpha)$ is contained in all subfields of E that contain both F and α, it is certainly the *smallest* such subfield. If $E = F(\alpha)$, we then say E is a **simple extension** of F. If α is algebraic over F, we say that $F(\alpha)$ is an **algebraic simple extension** of F.

Example 43.5

One example of an algebraic simple extension we've seen before is $\mathbb{Q}(\sqrt{2}) = \{a + b\sqrt{2} : a, b \in \mathbb{Q}\}$. In Section 38.1 we argued directly that every element of this field belongs to *any* subfield of $\mathbb{C}$ that contains the rational numbers and $\sqrt{2}$; furthermore, the set of such elements forms a field. This was a more concrete way of seeing that this field is the smallest field extension of $\mathbb{Q}$ containing $\sqrt{2}$.

In the previous example we have an explicit description of the elements of a simple extension; of course, this explicit description depended on knowing the specific arithmetic properties of $\sqrt{2}$. Our goal now is to find such a description for the general simple algebraic extension $F(\alpha)$ of a field F. We get the answer in the next two theorems. Here, the minimal polynomial of α over F will provide us with the arithmetic information we need about α.

Theorem 43.2 *Let F be a subfield of a field E and α an element of E that is algebraic over F; let $p \in F[x]$ be the minimal polynomial for α. Then the simple extension $F(\alpha)$ is isomorphic to the field $F[x]/\langle p\rangle$. Consequently, any two simple algebraic extensions of F by a root of p are isomorphic.*

Proof: Consider the evaluation homomorphism $\psi : F[x] \to E$ defined by $\psi(f) = f(\alpha)$. It is now evident that the image of ψ is contained in $F(\alpha)$. From the proof of Kronecker's Theorem 42.1 it is clear that p divides every polynomial in the kernel of ψ; it follows that the kernel of the homomorphism ψ is exactly $\langle p\rangle$. The Fundamental Isomorphism Theorem for Rings 19.1 then tells us that $F[x]/\langle p\rangle$ is isomorphic to the image of ψ in F. Because p is irreducible, $F[x]/\langle p\rangle$ is a field (Theorems 13.3c and 20.1.). The isomorphic image of this field in E contains both F and α. Hence, the image of ψ is exactly $F(\alpha)$, the smallest subfield of E containing F and α.

Because this is true regardless of the field E, this means that any two simple algebraic extensions of this form are isomorphic to $F[x]/\langle p\rangle$, and hence to each other. □

Apparently the simple algebraic extension $F(\alpha)$ depends heavily on the nature of the larger field E in which we are doing our computations. However, this is not actually the case: The theorem (and its proof) asserts that $F(\alpha)$ is entirely determined by F, and by the minimal polynomial for α.

The previous theorem now allows us to describe the elements of an algebraic simple extension quite explicitly.

Theorem 43.3 *Consider the simple extension $F(\alpha)$ of F where α is algebraic over F. Let $n \geq 1$ be the degree of α over F. Then every element β of $F(\alpha)$ can be uniquely written as*

$$\beta = b_0 + b_1\alpha + b_2\alpha^2 + \cdots + b_{n-1}\alpha^{n-1},$$

where $b_i \in F$.

Proof: Consider the evaluation homomorphism ψ_α from $F[x]$ to $F(\alpha)$: If $f \in F[x]$, then $\psi_\alpha(f) = f(\alpha)$. (The previous theorem shows us that this is onto.) So, if $f = a_0 + a_1x + \cdots + a_mx^m$, then $\psi_\alpha(f) = a_0 + a_1\alpha + a_2\alpha^2 + \cdots + a_m\alpha^m$, an element of $F(\alpha)$. Let $p = c_0 + c_1x + c_2x^2 + \cdots + x^n$ be the minimal polynomial of α over F. (Recall that the minimal polynomial is monic.) Because $p(\alpha) = 0$, we have $\alpha^n = -c_0 - c_1\alpha - c_2\alpha^2 - \cdots - c_{n-1}\alpha^{n-1}$. Note that we can use this equation to write every α^m for $m \geq n$ in terms of powers of α less than n. For example,

$$\begin{aligned}\alpha^{n+1} &= \alpha\alpha^n = -c_0\alpha - c_1\alpha^2 - \cdots - c_{n-1}\alpha^n \\ &= -c_0\alpha - c_1\alpha^2 - \cdots - c_{n-1}(-c_0 - c_1\alpha - \cdots - c_{n-1}\alpha^{n-1}).\end{aligned}$$

Higher powers of α are whittled down in this manner. So, every element β of $F(\alpha)$ can be written

$$\beta = b_0 + b_1\alpha + b_2\alpha^2 + \cdots + b_{n-1}\alpha^{n-1}.$$

We need only show that this expression is unique. So, suppose

$$\beta = b_0 + b_1\alpha + \cdots + b_{n-1}\alpha^{n-1} = d_0 + d_1\alpha + \cdots + d_{n-1}\alpha^{n-1}$$

for $b_i, d_i \in F$. Consider the polynomial

$$f = (b_0 - d_0) + (b_1 - d_1)x + \cdots + (b_{n-1} - d_{n-1})x^{n-1}.$$

Then $f \in F[x]$ and $f(\alpha) = 0$. Furthermore, the degree of f is less than the degree of p. But we know that p is of minimal degree among those polynomials with α as a root. Thus, f must be the zero polynomial. In other words, $b_i = d_i$ for $i = 0, 1, \ldots, n-1$, as desired. □

Example 43.6

Consider the real number $\alpha = \sqrt{2} + \sqrt[3]{2}$. It turns out that this element has degree 6 over $\mathbb{Q}$, with minimal polynomial $x^6 - 6x^4 - 4x^3 + 12x^2 - 24x - 4$. (See Exercise 43.3.) This means that

$$\mathbb{Q}(\alpha) = \{b_0 + b_1\alpha + b_2\alpha^2 + b_3\alpha^3 + b_4\alpha^4 + b_5\alpha^5 \in \mathbb{R} : b_i \in \mathbb{Q}\}.$$

That is, every element of $\mathbb{Q}(\alpha)$ can be expressed as a fifth-degree (or smaller) polynomial from $\mathbb{Q}[x]$, evaluated at α.

But consider the real number $\frac{1}{\alpha}$. This is obviously an element of the field $\mathbb{Q}(\alpha)$, and so we must be able to express it in terms of a fifth degree polynomial evaluated at α. But we know that $\alpha^6 - 6\alpha^4 - 4\alpha^3 + 12\alpha^2 - 24\alpha = 4$. If we divide by α, we then have that

$$\frac{4}{\alpha} = -24 + 12\alpha - 4\alpha^2 - 6\alpha^3 + \alpha^5.$$

Division by the integer 4 gives us the required expression for $\frac{1}{\alpha}$. In Exercise 43.5 you will do some more computations of this sort in this field.

Example 43.7

Let's return to Example 42.5: the polynomial $p = x^2 + x + 1$ is irreducible in $\mathbb{Z}_2[x]$. Kronecker's Theorem 42.1 says that there is an extension field E of $\mathbb{Z}_2$ that contains a root α of p. But p is degree 2, and so Theorem 43.3 says that $\mathbb{Z}_2(\alpha) = \{a + b\alpha : a, b \in \mathbb{Z}_2\}$. That is,

$$\mathbb{Z}_2(\alpha) = \{0 + 0\alpha, 0 + 1\alpha, 1 + 0\alpha, 1 + 1\alpha\}.$$

Thus, $\mathbb{Z}_2(\alpha)$ is a field with four elements—something we have not seen before (unless you did Exercise 8.12). It is easy to give the addition and multiplication tables for this field. The critical fact when computing these tables is that $\alpha^2 + \alpha + 1 = 0$. We leave the actual computations as Exercise 43.1.

Example 43.8

As another example, we return to $\mathbb{R}[x]/\langle x^2 + 1\rangle$, a field extension of $\mathbb{R}$ that contains a root for $x^2 + 1$. Specifically, the root is the coset $\alpha = \langle x^2 + 1\rangle + x$. Then

$$\mathbb{R}(\alpha) = \mathbb{R}[x]/\langle x^2 + 1\rangle$$

and

$$\mathbb{R}(\alpha) = \{a + b\alpha : a, b \in \mathbb{R}\},$$

where $\alpha^2 + 1 = 0$. Of course,

$$\mathbb{C} = \{a + bi : a, b \in \mathbb{R}\},$$

where $i^2 + 1 = 0$. It is easy to show that these two fields are isomorphic via the function $\varphi : \mathbb{R}(\alpha) \to \mathbb{C}$ given by $\varphi(a + b\alpha) = a + bi$. (This is because α and i play the same roles in their respective fields. Specifically, $\alpha^2 = -1$ and $i^2 = -1$.)

$\triangleright$ **Quick Exercise.** Show that the map φ given above is an isomorphism. $\triangleleft$

Example 43.9

Consider the irreducible polynomial $x^3 - 2 \in \mathbb{Q}[x]$. This polynomial has three distinct roots in $\mathbb{C}$, namely $\sqrt[3]{2}$, $\sqrt[3]{2}\zeta$, $\sqrt[3]{2}\zeta^2$, where

$$\zeta = e^{\frac{2\pi i}{3}} = -\frac{1}{2} + \frac{\sqrt{3}i}{2}$$

is the primitive cube root of unity (see Exercise 9.25).

$\triangleright$ **Quick Exercise.** Verify that ζ and ζ^2 are cube roots of 1, and so the three given numbers are in fact roots of $x^3 - 2$. $\triangleleft$

Now consider the three simple extensions

$$\mathbb{Q}(\sqrt[3]{2}), \quad \mathbb{Q}(\sqrt[3]{2}\zeta), \quad \mathbb{Q}(\sqrt[3]{2}\zeta^2)$$

of $\mathbb{Q}$. According to Theorem 43.2, all three of these fields are isomorphic, even though the first is a subfield of the real numbers $\mathbb{R}$, while the latter two fields obviously include complex numbers.

It is clear that $\sqrt[3]{2}\zeta, \sqrt[3]{2}\zeta^2 \notin \mathbb{Q}(\sqrt[3]{2})$, but we claim that it is also the case that $\sqrt[3]{2} \notin \mathbb{Q}(\sqrt[3]{2}\zeta)$. If it were, then by Theorem 43.3,

$$\sqrt[3]{2} = a + b(\sqrt[3]{2}\zeta) + c(\sqrt[3]{2}\zeta)^2,$$

for some rational numbers a, b, c. By setting real and imaginary parts of this equation equal, we obtain

$$\sqrt[3]{2} = a - \frac{1}{2}\sqrt[3]{2}b - \frac{1}{2}\sqrt[3]{4}c$$

and

$$0 = \frac{\sqrt{3}}{2}\sqrt[3]{2}b - \frac{\sqrt{3}}{2}\sqrt[3]{4}c.$$

The second equation implies that $b = c = 0$, because otherwise we could infer that $\sqrt[3]{2}$ is rational. But then the first equation implies that $\sqrt[3]{2}$ is rational. By similar arguments, you can check (see Exercise 43.9) that none of the three simple extensions

$$\mathbb{Q}(\sqrt[3]{2}),\ \mathbb{Q}(\sqrt[3]{2}\zeta),\ \mathbb{Q}(\sqrt[3]{2}\zeta^2)$$

of $\mathbb{Q}$ contain either of the other two roots for $x^3 - 2$.

Let's actually find an explicit isomorphism φ between $\mathbb{Q}(\sqrt[3]{2})$ and $\mathbb{Q}(\sqrt[3]{2}\zeta)$. Since 1 must be sent to 1, it is easy to see that $\varphi(q) = q$ for all rational numbers q. Now if α is a root of $x^3 - 2$, then $\alpha^3 - 2 = 0$, and so $\varphi(\alpha)^3 - \varphi(2) = \varphi(0)$, or in other words, $\varphi(\alpha)$ must also be a root of $x^3 - 2$. But because we have argued that $\sqrt[3]{2}\zeta$ is the *only* root of $x^3 - 2$ belonging to $\mathbb{Q}(\sqrt[3]{2}\zeta)$, it then follows that $\varphi(\sqrt[3]{2}) = \sqrt[3]{2}\zeta$. Thus, the isomorphism is easily described in terms of the unique representation of Theorem 43.3 as follows: Suppose that $\beta = a + b\sqrt[3]{2} + (\sqrt[3]{2})^2$ is an arbitrary element of $\mathbb{Q}(\sqrt[3]{2})$; then $\varphi(\beta) = a + b\sqrt[3]{2}\zeta + c(\sqrt[3]{2}\zeta)^2$.

In Chapters 45 and 47 we will return to this example and examine this isomorphism (and others), from a more sophisticated point of view.

43.3 Simple Transcendental Extensions

Notice that Theorem 43.3 describing simple extensions applies only to *algebraic* extensions. To see how nice this description is, let's examine a *transcendental* simple extension:

Example 43.10

Consider the simple extension $\mathbb{Q}(\pi)$ of $\mathbb{Q}$ by the transcendental element π. The general intersection argument we gave above for obtaining the simple extension still applies in this case, and so we do have the smallest subfield $\mathbb{Q}(\pi)$ of $\mathbb{R}$, which contains the rational numbers and the transcendental π as well. Thus, such real numbers as

$$\frac{1}{\pi} \quad \text{and} \quad \frac{1/2 + 3\pi^2}{2 + 3\pi + (5/4)\pi^2}$$

must belong to this field because they are obtained just by field operations. But in this case we *cannot* express such elements as a rational polynomial in π. Indeed,

$$\mathbb{Q}(\pi) = \left\{ \frac{f(\pi)}{g(\pi)} \in \mathbb{R} : f, g \in \mathbb{Q}[x],\ g \neq 0 \right\}.$$

In Exercise 43.6 you will prove that this set of real numbers is actually a field and consequently is the simple extension required.

For our purposes, algebraic simple extensions are much more important than transcendental ones, which is fortunate, considering how much nicer our description is in the algebraic case.

43.4 Dimension of Simple Algebraic Extensions

Notice that the description of the elements of algebraic simple extensions in Theorem 43.3 has a distinctive vector space flavor. Indeed, the theorem says that when the simple algebraic extension $F(\alpha)$ is viewed as a vector space over F, the set

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

is a basis for this vector space. Let us restate this important fact: *If the degree of the algebraic element α over a field* F *is* n, *then* n *is the dimension of* F(α) *as a vector space over* F.

Example 43.11

We now rephrase Example 43.6: $\mathbb{Q}(\sqrt[3]{2} + \sqrt{2})$ is a vector space of dimension 6 over $\mathbb{Q}$. A basis consists of

$$1,\ \sqrt[3]{2} + \sqrt{2},\ (\sqrt[3]{2} + \sqrt{2})^2,\ \cdots,\ (\sqrt[3]{2} + \sqrt{2})^5.$$

Example 43.12

The field with four elements described in Example 43.7 is a vector space of dimension 2 over $\mathbb{Z}_2$. A basis consists of 1 and α.

The following is an interesting corollary to the fact just stated.

Corollary 43.4 *If α is algebraic over F and $F(\alpha)$ has dimension n over F and $\beta \in F(\alpha)$, then β is also algebraic over F. Furthermore, the degree of β over F is at most n.*

Proof: Because $F(\alpha)$ has dimension n over F, any collection of more than n elements of $F(\alpha)$ is linearly dependent (Corollary 41.6). Specifically, consider the $n + 1$ elements

$$1, \beta, \beta^2, \ldots, \beta^n.$$

Because this set is linearly dependent, there exist $b_0, b_1, \ldots, b_n \in F$, such that

$$b_0 + b_1\beta + b_2\beta^2 + \cdots + b_n\beta^n = 0.$$

That is, β is a root of the polynomial $b_0+b_1x+b_2x^2+\cdots+b_nx^n \in F[x]$. So, β is algebraic over F of degree no more than n. □

To rephrase: Every element of an algebraic simple extension is algebraic over the base field.

What about transcendental simple extensions? The reason we have no nice element-wise description for such a field is exactly this: The vector space dimension of such a field over its base field is infinite! You will prove this fact in Exercise 43.8.

Chapter Summary

For an element that is algebraic over a field we defined the *minimal polynomial* for that element over the field. The degree of this polynomial is the *degree* of the element over the field. We showed that such a polynomial always exists.

We also examined *simple extensions* and showed that in the case of a simple extension of a field by an algebraic element, the extension is in fact a vector space over the base field with dimension equal to the degree of the element. We showed this by explicitly displaying a basis for this vector space. Using this theorem, we were then able to construct a field with four elements.

Warm-up Exercises

a. Return to Exercise a in the previous chapter. In each case, determine a basis for the simple extension obtained by adjoining the element to the field in question.

b. Express the following field elements as a linear combination of the basis elements you determined in Exercise a.

(a) $\dfrac{1}{2+\sqrt{2}} \in \mathbb{Q}(1+\sqrt{2})$.

(b) $\dfrac{\sqrt[3]{2}}{1+2\sqrt[3]{2}} \in \mathbb{Q}(\sqrt[3]{2})$.

c. Explain why the following statements are true, or else give a counterexample. In each case F is a subfield of the field E, and $\alpha \in E\backslash F$:

(a) $F(\alpha)$ is finite dimensional over F.

(b) Let α be algebraic over F. Then $F(\alpha)$ is finite dimensional over F.

(c) Every element of $F(\alpha)$ is algebraic over F.

(d) Every element of $F(\alpha)$ is algebraic over E.

(e) Suppose that α is algebraic over F. Then every element of $F(\alpha)$ is algebraic over F.

d. Explain why the minimal polynomial is unique.

Exercises

1. (a) Compute the addition and multiplication tables for the field $\mathbb{Z}_2(\alpha)$ described in Example 43.7. Use your table to determine explicitly the multiplicative inverse of each non-zero element of this field. (This exercise is essentially a repeat of Exercise 8.12.)

   (b) In Exercise 17.18, you considered the Frobenius isomorphism, defined on any finite field. Compute this homomorphism explicitly in this case.

2. Consider the polynomial

$$f = 1 + x^2 + x^3 \in \mathbb{Z}_2(\alpha)[x],$$

where $\mathbb{Z}_2(\alpha)$ is the field considered in the previous exercise.

   (a) Use the Root Theorem to show that this polynomial is irreducible.

   (b) By Kronecker's Theorem 42.1 we can construct an extension field of $\mathbb{Z}_2(\alpha)$ so that f has a root β. How many elements does this field have?

   (c) If you're not doing anything this weekend, construct a multiplication table for this field.

3. In this problem you will prove that

$$f = x^6 - 6x^4 - 4x^3 + 12x^2 - 24x - 4$$

is the minimal polynomial for $\alpha = \sqrt{2}+\sqrt[3]{2}$, as stated in Example 43.6. (We will return to this example in Example 44.3.)

(a) Check that $f(\alpha) = 0$.

(b) Show that f is irreducible in $\mathbb{Q}[x]$, by using Eisenstein's Criterion 5.7, after the linear change of variables $x = y + 2$ (see Exercise 5.14).

(c) Show that f can be factored into irreducibles in $\mathbb{R}[x]$ as follows:

$$\left(x - \sqrt{2} - \sqrt[3]{2}\right)\left(x^2 - (2\sqrt{2} + \sqrt[3]{2})x + (2 + \sqrt{2}\sqrt[2]{2} + \sqrt[3]{4})\right)$$
$$\left(x + \sqrt{2} - \sqrt[3]{2}\right)\left(x^2 + (2\sqrt{2} - \sqrt[3]{2})x + (2 - \sqrt{2}\sqrt[2]{2} + \sqrt[3]{4})\right).$$

(d) How would you factor f into irreducibles in $\mathbb{C}[x]$? (You need not feel obliged to actually carry this out!)

4. Determine the minimal polynomial of $\sqrt{2} + \sqrt{3}$ over the following three fields: $\mathbb{Q}$, $\mathbb{Q}(\sqrt{2})$, $\mathbb{R}$.

5. Let $\alpha = \sqrt{2} + \sqrt[3]{2}$, and f be its minimal polynomial, as in Example 43.6, and in Exercise 3. Show explicitly that the following elements of $\mathbb{Q}(\alpha)$ are linear combinations of

$$1, \alpha, \alpha^2, \cdots, \alpha^5.$$

(a) $\dfrac{1}{\alpha + 1}$.

(b) $\dfrac{\alpha}{1 + \alpha^2}$.

6. Let

$$K = \left\{ \frac{f(\pi)}{g(\pi)} \in \mathbb{R} : f, g \in \mathbb{Q}[x],\ g \neq 0 \right\},$$

as given in Example 43.10.

(a) Why do the quotients defined above always make sense? (That is, why are the denominators always non-zero?)

(b) Show that K is a subfield of $\mathbb{R}$.

(c) Argue that $K = \mathbb{Q}(\pi)$.

(d) Are the elements in K uniquely represented, as we've presented them in the definition of K?

7. Let F be a subfield of the field E, and suppose that $\alpha \in E$ is transcendental over F. Using Exercise 6 as a model, formulate an explicit description of the members of the simple extension $F(\alpha)$, and prove that your formulation works.

8. Let F be a subfield of the field E, and suppose that $\alpha \in E$ is transcendental over F.

(a) Show that

$$\{1, \alpha, \alpha^2, \alpha^3, \cdots\}$$

is a linearly independent set of the vector space $F(\alpha)$ (over F).

(b) Why does part a mean that the dimension of $F(\alpha)$ over F is infinite?

(c) Show that the set in part a is *not* a basis for $F(\alpha)$ over F.

9. In this exercise we perform some additional verifications related to Example 43.9. Show that $\sqrt[3]{2}, \sqrt[3]{2}\zeta \notin \mathbb{Q}(\sqrt[3]{2}\zeta^2)$ and $\sqrt[3]{2}\zeta^2 \notin \mathbb{Q}(\sqrt[3]{2}\zeta)$. Thus, in each case the field isomorphisms between the three simple extensions of Example 43.9 are very simply described. How?

Chapter 44

Finite Extensions and Constructibility Revisited

In the last chapter we obtained a good description for a simple extension of a field by an algebraic element. Specifically, if α is algebraic over F, we showed that $F(\alpha)$ is a vector space over F with dimension equal to the degree of α over F. Furthermore, every element of $F(\alpha)$ is algebraic over F. In this chapter we are interested in field extensions that are not necessarily simple, but in which every element is algebraic. We'll then use our further results to provide more elegant proofs of the impossibility of two of the classical construction problems of the ancient Greeks.

44.1 Finite Extensions

We say that a field extension E of a field F is an **algebraic extension** if every element of E is algebraic over F. That is, every element of E is a root of some polynomial in $F[x]$. We proved in the last chapter that every element of a simple extension by an algebraic element is in fact an algebraic element itself; using this new terminology, this means that every such extension is algebraic.

We say E is a **finite extension** of the field F if it has finite dimension as a vector space over F. We shall use $[E : F]$ to stand for the dimension of E over F, also called the **degree of E over F**.

Example 44.1

> We know that $\mathbb{Q}(\sqrt[3]{3})$ is a finite extension of $\mathbb{Q}$ and $[\mathbb{Q}(\sqrt[3]{3}) : \mathbb{Q}] = 3$. Likewise, $\mathbb{C}$ is a finite extension of $\mathbb{R}$, and $[\mathbb{C} : \mathbb{R}] = 2$.

We showed in the last chapter that any simple extension by an algebraic element is a finite extension (Section 43.4). The following theorem asserts that every finite extension is algebraic; its proof uses an argument similar to that we used in Corollary 43.4 when we proved that elements of algebraic simple extensions are algebraic.

Theorem 44.1 *A finite extension of a field is an algebraic extension.*

Proof: Suppose that E is a finite extension of F and $[E : F] = n$. Let $\alpha \in E$. Then the $n + 1$ elements

$$1, \alpha, \alpha^2, \ldots, \alpha^n$$

cannot be linearly independent, by Corollary 41.6. Therefore, there exist

$$a_0, a_1, \ldots, a_n \in F$$

such that

$$a_0 + a_1\alpha + a_2\alpha^2 + \cdots + a_n\alpha^n = 0.$$

In other words, α is a root of the polynomial $a_0+a_1x+a_2x^2+\cdots+a_nx^n$ in $F[x]$. □

What about the converse of this theorem? In Exercise 44.5 you will explore an example of a field extension that is algebraic but not finite, thus showing that the converse is false.

The next theorem is the most important result of this chapter; it is this theorem that will make our new approach to the constructibility problems quite easy. It is a counting theorem, relating the dimensions of field extensions to one another. Counting theorems are always important. Despite the theorem's importance, its proof is easy. In fact, before reading the proof, you might attempt it yourself.

Theorem 44.2 *If K is a finite extension of a field E and E is a finite extension of a field F, then K is a finite extension of the field F. Furthermore,*

$$[K : F] = [K : E][E : F].$$

Proof: Let $\{\alpha_i : i = 1, \ldots, n\}$ be a basis for E over F and $\{\beta_j : j = 1, \ldots, m\}$ be a basis for K over E. We will show that

$$\{\alpha_i\beta_j : i = 1, \ldots, n,\ j = 1, \ldots, m\}$$

is a basis for K over F. Note that there are mn distinct elements in this set.

First, we show that the mn elements $\alpha_i\beta_j$ span K. Accordingly, let $k \in K$. Because the β_j span K as a vector space over E, there exist elements $a_1, \ldots, a_m$ of E such that

$$k = \sum_{j=1}^{m} a_j\beta_j.$$

Likewise, because the α_i span E as a vector space over F, for each a_j there exist elements $b_{1j}, \ldots, b_{nj}$ of F such that

$$a_j = \sum_{i=1}^{n} b_{ij}\alpha_i.$$

Substituting these into the sum for k, we get

$$k = \sum_{j=1}^{m}\left(\sum_{i=1}^{n} b_{ij}\alpha_i\right)\beta_j = \sum_{i,j} b_{ij}(\alpha_i\beta_j).$$

Thus, the mn elements $\alpha_i\beta_j$ span K. (This argument is really a repeat of Exercise 42.15.)

We now show that the $\alpha_i\beta_j$ are linearly independent. So, suppose that

$$\sum_{i,j} c_{ij}(\alpha_i\beta_j) = 0, \quad c_{ij} \in F.$$

We will show that $c_{ij} = 0$ for all i and j. But,

$$0 = \sum_{i,j} c_{ij}(\alpha_i\beta_j) = \sum_{j=1}^{m}\left(\sum_{i=1}^{n} c_{ij}\alpha_i\right)\beta_j.$$

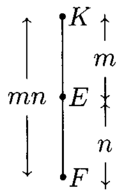
The β_j form a basis for K over E; therefore,

$$\sum_{i=1}^{n} c_{ij}\alpha_i = 0$$

for each j. But the α_i form a basis for E over F, so $c_{ij} = 0$ for each i and j. Thus, the mn elements $\alpha_i\beta_j$ are linearly independent, and so form a basis for K as a vector space over F. □

A rough paraphrase of this theorem is this: A finite extension of a finite extension is a finite extension. Notice that we have explicitly constructed a basis for K over F from the bases given for K over E

and E over F. We will explore some specific examples of this procedure below. The theorem is pictured in the following diagram:



Before looking at examples, we will sharpen our result for the case of simple algebraic extensions. Suppose that E is an extension field of F and $\alpha \in E$ is algebraic over F. Let $\beta \in F(\alpha)$. Thus, $F(\beta) \subseteq F(\alpha)$. We have seen from Corollary 43.4 that β is algebraic over F, and so it follows from the last theorem that $[F(\beta) : F]$ must divide $[F(\alpha) : F]$. That is, the degree of β over F must divide the degree of α over F. We restate this in the following corollary.

Corollary 44.3 *Suppose E is an extension field of the field F and $\alpha \in E$ is algebraic over F. If $\beta \in F(\alpha)$, then the degree of β over F divides the degree of α over F.*

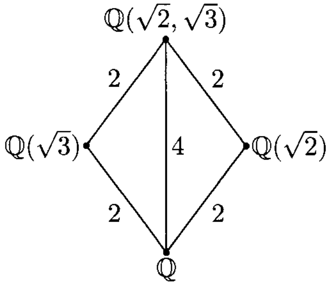
We now illustrate the theorem with a couple of examples.

Example 44.2

Consider $\mathbb{Q}(\sqrt{2}, \sqrt{3})$, the smallest subfield of $\mathbb{R}$ containing $\mathbb{Q}$, $\sqrt{2}$, and $\sqrt{3}$, as a vector space over $\mathbb{Q}$. Now, $\sqrt{3} \notin \mathbb{Q}(\sqrt{2})$ (see Exercise 44.1 or Example 38.2), and so $x^2 - 3$ is the minimal polynomial for $\sqrt{3}$ over $\mathbb{Q}(\sqrt{2})$. Therefore, $[\mathbb{Q}(\sqrt{2}, \sqrt{3}) : \mathbb{Q}(\sqrt{2})] = 2$ and $\{1, \sqrt{3}\}$ is a basis for $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ over $\mathbb{Q}(\sqrt{2})$. Because $[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 2$ and $\{1, \sqrt{2}\}$ is a basis for $\mathbb{Q}(\sqrt{2})$ over $\mathbb{Q}$, we have that $[\mathbb{Q}(\sqrt{2}, \sqrt{3}) : \mathbb{Q}] = 4$ and $\{1, \sqrt{2}, \sqrt{3}, \sqrt{6}\}$ is a basis for $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ over $\mathbb{Q}$. Notice that here we are thinking of $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ as a simple extension of $\mathbb{Q}(\sqrt{2})$, which in turn is a simple extension of $\mathbb{Q}$.

Alternatively, we could think of $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ as an extension of $\mathbb{Q}(\sqrt{3})$. Then, $[\mathbb{Q}(\sqrt{2}, \sqrt{3}) : \mathbb{Q}(\sqrt{3})] = 2$ and $[\mathbb{Q}(\sqrt{3}) : \mathbb{Q}] = 2$. These are two ways of stepping up from $\mathbb{Q}$ to $\mathbb{Q}(\sqrt{2}, \sqrt{3})$: First, we adjoin $\sqrt{2}$ to $\mathbb{Q}$ and then adjoin $\sqrt{3}$, or first adjoin $\sqrt{3}$ to $\mathbb{Q}$ and then adjoin $\sqrt{2}$.

The following diagram illustrates this situation.



You should refer to Example 38.2 for an account of this example that does not use the theory of vector spaces.

Example 44.3

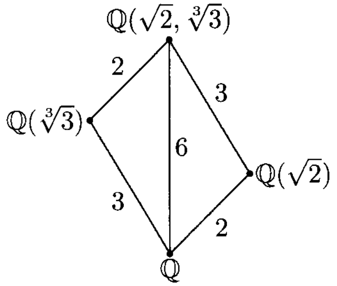
Now consider $\mathbb{Q}(\sqrt{2}, \sqrt[3]{2})$. Because $[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3$ and $[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 2$, it follows from Corollary 44.3 that $\sqrt[3]{2} \notin \mathbb{Q}(\sqrt{2})$. Thus, $[\mathbb{Q}(\sqrt{2}, \sqrt[3]{2}) : \mathbb{Q}(\sqrt{2})] = 3$. So,

$$[\mathbb{Q}(\sqrt{2}, \sqrt[3]{2}) : \mathbb{Q}] = [\mathbb{Q}(\sqrt{2}, \sqrt[3]{2}) : \mathbb{Q}(\sqrt{2})][\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 3 \cdot 2 = 6.$$

Again, we could build $\mathbb{Q}(\sqrt{2}, \sqrt[3]{2})$ another way. Because $\sqrt{2} \notin \mathbb{Q}(\sqrt[3]{2})$ (from Corollary 44.3), we have that $[\mathbb{Q}(\sqrt{2}, \sqrt[3]{2}) : \mathbb{Q}(\sqrt[3]{2})] = 2$, and so

$$[\mathbb{Q}(\sqrt{2}, \sqrt[3]{2}) : \mathbb{Q}] = [\mathbb{Q}(\sqrt{2}, \sqrt[3]{2}) : \mathbb{Q}(\sqrt[3]{2})][\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 2 \cdot 3 = 6.$$

The following diagram illustrates this.


$\mathbb{Q}(\sqrt{2}, \sqrt[3]{3})$
2
3
$\mathbb{Q}(\sqrt[3]{3})$
6
$\mathbb{Q}(\sqrt{2})$
3
2
$\mathbb{Q}$

But now consider the simple algebraic extension $\mathbb{Q}(\alpha)$, where $\alpha = \sqrt{2} + \sqrt[3]{2}$. It is evident that $\alpha \in \mathbb{Q}(\sqrt{2}, \sqrt[3]{2})$, and so $\mathbb{Q}(\alpha) \subseteq \mathbb{Q}(\sqrt{2}, \sqrt[3]{2})$; we claim that this last field extension is trivial. In Exercise 43.3 we proved that $\mathbb{Q}(\alpha)$ is of degree 6 over $\mathbb{Q}$. So by the theorem, we have that $[\mathbb{Q}(\sqrt{2}, \sqrt[3]{2}) : \mathbb{Q}(\alpha)] = 1$, as we require.

But let's prove that $\mathbb{Q}(\sqrt{2}, \sqrt[3]{2}) = \mathbb{Q}(\alpha)$ again, without making use of Exercise 43.3. We know that $\sqrt[3]{2} = \alpha - \sqrt{2}$. Cubing both sides of this equation gives

$$2 = \alpha^3 - 3\sqrt{2}\alpha^2 + 6\alpha - 2\sqrt{2}.$$

Solving this equation for $\sqrt{2}$ gives us

$$\sqrt{2} = \frac{\alpha^3 + 6\alpha - 2}{3\alpha^2 + 2}.$$

▷ **Quick Exercise.** Check this. ◁

But this means that $\sqrt{2} \in \mathbb{Q}(\alpha)$; but then $\sqrt[3]{2} = \alpha - \sqrt{2} \in \mathbb{Q}(\alpha)$ too, and so $\mathbb{Q}(\sqrt{2}, \sqrt[3]{2}) \subseteq \mathbb{Q}(\alpha)$. Because we've already noted the reverse inclusion, we have proved again that $\mathbb{Q}(\alpha) = \mathbb{Q}(\sqrt{2}, \sqrt[3]{2})$. This means of course that the minimal polynomial for α must be of degree 6. We thus arrive at the result of Exercise 43.3 again, but by vector space considerations.

Note furthermore that the finite extension $\mathbb{Q}(\sqrt{2}, \sqrt[3]{2})$ over $\mathbb{Q}$ is actually simple. It turns out that *all* finite extensions over the rational field are simple; we will prove this important fact in Theorem 45.5.

It is clear that an induction argument extends Theorem 44.2 to any number of steps. For completeness, we include this as a corollary. (The proof of this is left as Exercise 44.2.)

Corollary 44.4 *If F_i, $i = 1, \ldots, n$ are fields with F_{i+1} a finite extension of F_i, for $i = 1, \ldots, n-1$, then F_n is a finite extension of F_1 and*

$$[F_n : F_1] = [F_n : F_{n-1}][F_{n-1} : F_{n-2}] \cdots [F_2 : F_1].$$

44.2 Constructibility Problems

Let's start by recalling the three Greek construction problems: doubling the cube, trisecting an arbitrary angle, and squaring the circle. The proofs that these constructions are impossible depend in an essential way on the Constructible Number Theorem 38.3, which we re-state here:

Constructible Number Theorem *The number α is constructible if and only if there exists a finite sequence of fields*

$$\mathbb{Q} = F_0 \subset F_1 \subset \cdots \subset F_N,$$

with $\alpha \in F_N$ and $F_{i+1} = F_i(\sqrt{k_i})$ for some $k_i \in F_i$, with $k_i > 0$ for $i = 0, \ldots, N-1$.

Let's re-consider this theorem, in light of our theory of field extensions. We now know that the degree of each of the extensions in the Constructible Number Theorem is 2. Thus, $[F_N : \mathbb{Q}]$ must be a power of 2. Now clearly the extension $\mathbb{Q}(\alpha) \subseteq F_N$, and so by Theorem 44.2, we must have that $[\mathbb{Q}(\alpha) : \mathbb{Q}]$ is also a power of 2. We thus obtain the following corollary of the Constructible Number Theorem:

Corollary 44.5 *If a number α is constructible, then $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 2^n$, for some positive integer n.*

Note that the converse to this corollary is false; $[\mathbb{Q}(i) : \mathbb{Q}] = 2$ is a counterexample, because constructible numbers are necessarily real. But the converse is also false for real numbers. In Exercise 49.7 we will present an irreducible fourth degree polynomial in $\mathbb{Q}[x]$ with a real root α that is not constructible, even though $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 4$.

Let's now return to the impossibility proofs of the construction problems. Recall that to double the cube, we are to construct an edge of a cube whose volume is twice that of a given cube. (Actually, we are given the *edge* of the original cube.) If we consider the edge of the original cube to be length 1, then we are required to construct a line segment of length $\sqrt[3]{2}$. But $[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3$, and so Corollary 44.5 tells us that $\sqrt[3]{2}$ is not constructible. Thus, it is impossible to double the cube with only a compass and straightedge.

In our impossibility proof for the trisection problem, we showed that it is in fact impossible to trisect a 60° angle. If we could do so, we could construct an angle of 20° which in turn implies the construction of a root of the polynomial $x^3 - 3x - 1$, which is irreducible in $\mathbb{Q}[x]$. But any root of this polynomial has degree 3 over $\mathbb{Q}$; once again, Corollary 44.5 shows that this is not constructible.

We are not able to use the theory of this chapter to more easily prove the impossibility of squaring the circle, because this problem involves considering the *transcendental* simple extension $\mathbb{Q}(\sqrt{\pi})$. We must still rely on Lindemann's difficult Theorem 39.5 asserting that π is transcendental over $\mathbb{Q}$.

Chapter Summary

In this chapter we showed that every *finite extension* is an *algebraic extension.* If K is a finite extension of E, which is a finite extension of F, then K is a finite extension of F with $[K : F] = [K : E][E : F]$.

We then used this theorem to give alternate proofs for the impossibility of doubling the cube and trisecting an arbitrary angle.

Warm-up Exercises

a. If $[K : F] = 1$, what can you say about the fields K and F?

b. If $[K : F]$ is prime and E is a field where $F \subseteq E \subseteq K$, what can you say about the field E?

c. If $F \subseteq E \subseteq K$ is a sequence of finite field extensions and $[K : F] = [E : F]$, what can you say about the fields K and E?

d. Explain why the following statements are true, or else give a counterexample.

   (a) Every finite extension is algebraic.

   (b) Every simple extension is algebraic.

   (c) Every simple extension is finite.

   (d) Every algebraic extension is simple.

   (e) Every algebraic extension is finite.

e. How many elements belong to a field that is a degree 2 extension of a degree 3 extension of $\mathbb{Z}_7$?

f. Give examples of the following, or else explain why the example does not exist:

   (a) A degree 2 extension of $\mathbb{Q}(\sqrt{2})$.

   (b) A degree 3 extension of $\mathbb{R}$.

   (c) A degree 3 extension of $\mathbb{Z}_2$.

Exercises

1. Show that $\sqrt{2} \notin \mathbb{Q}(\sqrt{3})$, using Theorem 43.2. (See Example 38.2 for a more elementary solution of this exercise.)

2. Prove Corollary 44.4. That is, suppose that each F_i is a field, and $F_1 \subseteq F_2 \subseteq \cdots \subseteq F_n$ is a sequence of finite extensions. Show that

$$[F_n : F_1] = [F_n : F_{n-1}][F_{n-1} : F_{n-2}] \cdots [F_2 : F_1].$$

3. Prove that $\mathbb{Q}(\sqrt{2}+\sqrt{3}) = \mathbb{Q}(\sqrt{2}, \sqrt{3})$.

4. (a) Determine $[\mathbb{Q}(\sqrt[3]{2}, i) : \mathbb{Q}]$; include a proof of your result.

   (b) Find $\alpha \in \mathbb{C}$ so that $\mathbb{Q}(\alpha) = \mathbb{Q}(\sqrt[3]{2}, i)$.

5. In this exercise you will show that not every algebraic extension is a finite extension by considering the field F, constructed as follows. Let

$$F_1 = \mathbb{Q}(\sqrt{2}),\ F_2 = F_1(\sqrt[3]{2}),\ F_3 = F_2(\sqrt[4]{2}),\ F_4 = F_3(\sqrt[5]{2}),\ \cdots,$$

and continue inductively. Then let

$$F = \bigcup_{n=1}^{\infty} F_n.$$

Argue that F is field. Then prove that F is not a finite extension, thus showing that the converse of Theorem 44.1 is false.

6. Give an example to show that if $[F(a) : F] = n$ and $[F(b) : F] = m$, it does not necessarily follow that $[F(a, b) : F] = mn$.

7. Prove that an algebraic extension of an algebraic extension is an algebraic extension.

8. Consider the primitive cube root of unity

$$\zeta = -\frac{1}{2} + i\frac{\sqrt{3}}{2} = e^{\frac{2\pi i}{3}} = \cos\left(\frac{2\pi}{3}\right) + i \sin\left(\frac{2\pi}{3}\right).$$

Compute $[\mathbb{Q}(\zeta) : \mathbb{Q}]$.

9. Consider the field $F = \mathbb{Z}_2(\alpha, \beta)$ constructed in Exercise 43.2.

   (a) Use Theorem 44.2 to determine the possible degrees $[F : K]$, where K is a subfield of F. Find a subfield with each of these degrees, and draw a field extension diagram similar to that in Examples 44.2 and 44.3.

   (b) Prove that $\mathbb{Z}_2(\alpha + \beta) = \mathbb{Z}_2(\alpha, \beta)$.

10. Prove that $\mathbb{A}$, the set of all complex numbers algebraic over $\mathbb{Q}$, is a field.

Section VIII in a Nutshell

This section presents some more sophisticated mathematical ideas that allow us to prove more elegantly the impossibility of the constructibility problems of the previous section. This more powerful approach to field theory is interesting in its own right and also plays a vital role in Section IX.

First, we define a *vector space* V over a field F. This is a set of objects (called *vectors*) equipped with a binary operation called addition, making the set an additive group. In addition a vector space has a *scalar multiplication* whereby vectors are multipied by elements (*scalars*) from the field F. The complete set of axioms follows, where $\mathbf{v}, \mathbf{w}, \mathbf{u} \in V$ and $r, s \in F$:

1. $\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}$,

2. $\mathbf{v} + (\mathbf{w} + \mathbf{u}) = (\mathbf{v} + \mathbf{w}) + \mathbf{u}$,

3. there exists a **zero vector** $\mathbf{0}$, with the property that $\mathbf{v} + \mathbf{0} = \mathbf{v}$, and

4. every vector $\mathbf{v}$ has an **additive inverse** $-\mathbf{v}$, with the property that $\mathbf{v} + (-\mathbf{v}) = \mathbf{0}$.

5. $(r + s)\mathbf{v} = r\mathbf{v} + s\mathbf{v}$,

6. $(rs)\mathbf{v} = r(s\mathbf{v})$,

7. $r(\mathbf{v} + \mathbf{w}) = r\mathbf{v} + r\mathbf{w}$, and

8. $1\mathbf{v} = \mathbf{v}$.

Classic examples of vector spaces are the set F^n of n-tuples from the field F, and the polynomial ring $F[x]$. If F and E are fields and $F \subseteq E$, then the field E is actually a vector space over F. (In this case we say E is an *extension field* of the field F.) This eventually allows us to apply vector space theory to field theory.

A *basis* for a vector space V is a linearly independent set of vectors that spans V. The number of vectors in a basis is the *dimension* of the vector space. (See Theorem 41.4.) Thus F^n has dimension n and $F[x]$ has infinite dimension over F. If B is a basis for vector space V then every element of V can be uniquely written as a linear combination of elements in B (Theorem 41.2). Every spanning set contains a subset that is a basis (Theorem 41.3).

Kronecker's Theorem 42.1 says that if F is a field and $f \in F[x]$ of degree at least 1, we can always construct an extension field E of F such that $f(\alpha) = 0$ for some $\alpha \in E$. The proof of this theorem reveals that $F[x]/\langle p \rangle$ is the desired field, where p is an irreducible factor of f in $F[x]$.

If E is an extension field of F and $\alpha \in E$ is a root of a polynomial in $F[x]$, then α is algebraic over F. The smallest subfield $F(\alpha)$ of E containing F and α is essentially unique, and isomorphic to the field $F[x]/\langle p \rangle$, where p is an irreducible polynomial $p \in F[x]$ with $p(\alpha) = 0$ (Theorem 43.2). Then p divides all $f \in F[x]$ where $f(\alpha) = 0$ (Theorem 43.1), and if we further require that p be monic, it is unique and called the *minimal polynomial of α over F*.

Fields of the form $F(\alpha)$, are called *simple extensions* of F. If α happens to be algebraic over F, then $F(\alpha)$ is called an *algebraic simple extension* of F. Furthermore, if $\deg(p) = n$, the $F(\alpha)$ is a vector space over F of dimension n with basis $\{1, \alpha, \alpha^2, \ldots, \alpha^{n-1}\}$. It follows that every element in $F(\alpha)$ is algebraic over F (Corollary 43.4).

If E is an extension field of F and has finite dimension as a vector space over F, we call E a *finite extension* of F and denote the degree of the extension by $[E : F]$. Every finite extension is algebraic (Theorem 44.1). Degrees multiply: if $F_1 \subseteq F_2 \subseteq F_3$ form a sequence of finite extensions, then $[F_3 : F_1] = [F_3 : F_2][F_2 : F_1]$. This extends to any finite sequence of finite extensions. If E is an extension field of F and $\alpha \in E$ is algebraic over F and $\beta \in F(\alpha)$, then the degree of β over F divides the degree of α over F (Corollary 44.3).

It follows from the Constructible Number Theorem that if α is constructible, then $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 2^n$. This allows us to show that it is impossible to double the cube (because $[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3$), it is impossible to trisect a 60° angle (because to do so means we could construct a root of $x^3 - 3x - 1$, which is irreducible in $\mathbb{Q}[x]$ of degree 3), and it is impossible to square the circle (because $\mathbb{Q}(\sqrt{\pi})$ is a transcendental simple extension of $\mathbb{Q}$).

IX

Galois Theory

Chapter 45

The Splitting Field

We are all familiar with the way in which the quadratic formula gives us an explicit formula for the roots of any degree two polynomial in $\mathbb{Q}[x]$ (and, this formula works in $\mathbb{C}[x]$ too – see Exercises 9.1 and 9.2). In Exercises 9.12–9.19, you can explore the cubic formula that performs the same task as the quadratic formula; it is a good bit more complicated. In the case of the quadratic formula, we need to perform the field operations from $\mathbb{Q}$, and in addition extract a square root. In the case of the cubic formula, we need to perform the field operations from $\mathbb{Q}$, extract cube roots, and also have to extract one or more square roots. Can this process be extended to higher and higher degree polynomials from $\mathbb{Q}[x]$? The answer for quartic (that is, fourth degree) equations is yes; it is a nightmare of a formula, which involves not only the extraction of a fourth root, but also cube and square roots as well. In essence, the cubic problem reduces finally to a quadratic, and the quartic in turn finally reduces to a cubic! You can explore the quartic formula in Exercise 9.20.

It seems natural to suppose that this project could be carried on indefinitely (at the cost of more and more complicated calculations), to obtain formulas solving fifth, sixth and higher degree polynomial equations over $\mathbb{C}$. This supposition is false: It was one of the triumphs of nineteenth century abstract algebra to prove that such formulas for degree 5 and higher are impossible. This result can and should be compared to the results we have encountered in Chapters 39 and 44, showing that the classical construction problems are impossible.

In order to prove the impossibility of solving all polynomial equations by radicals, we will need some more field theory, which we will begin exploring in this chapter. To complete our project, we will eventually make use of the group theory we encountered in Chapter 36. The exciting interplay between groups and fields makes this subject an outstanding example of the power of abstract algebra.

45.1 The Splitting Field

In Chapter 42 we encountered Kronecker's Theorem 42.1: Given a field F and a non-constant polynomial $f \in F[x]$, we can always build a (potentially) bigger field than F in which f has at least one root. By an easy inductive argument (Exercise 42.1) on the degree of f, we can then build a (potentially) larger field than F, in which f can be completely factored into linear polynomials; such a field is called a **splitting extension** over F for f. In this chapter we will show that this can be done minimally in essentially only one way.

We make a definition to capture this idea. Suppose that F is a field and $f \in F[x]$, with $\deg(f) > 0$. Then the field K is a **splitting field for f over** F, if $F \subseteq K$, f factors into linear polynomials in $K[x]$, and if L is any other field with $F \subseteq L \subset K$, then f cannot be factored into linear polynomials in $L[x]$. Any field containing K is a splitting extension of f over F. The splitting field of f over F is a minimal splitting extension. When the polynomial f does factor into linear polynomials in a field K, we say f **splits in** K.

Example 45.1

If the polynomial $f \in F[x]$ can already be factored into linear polynomials in $F[x]$, then F itself is a splitting field for f.

Example 45.2

The field $\mathbb{C}$ is a splitting field for the polynomial $x^2 + 1 \in \mathbb{R}[x]$ over $\mathbb{R}$.

Example 45.3

Let $f \in \mathbb{Q}[x]$ be any quadratic polynomial that is irreducible in $\mathbb{Q}[x]$, and $\alpha \in \mathbb{C}$ be a root of f given by the quadratic formula. Then the quadratic extension $\mathbb{Q}(\alpha)$ is a splitting field for f over $\mathbb{Q}$. This is true because the second root for f in $\mathbb{C}$ is the complex conjugate $\bar{\alpha}$, and $\bar{\alpha} \in \mathbb{Q}(\alpha)$.

▷ **Quick Exercise.** Why is $\bar{\alpha} \in \mathbb{Q}(\alpha)$?

Example 45.4

Consider the polynomial $f = x^3 - 2 \in \mathbb{Q}[x]$ that we looked at in Example 43.9. The field $\mathbb{Q}\left(\sqrt[3]{2}\right)$ certainly contains a root for f, namely $\sqrt[3]{2}$. But it is not a splitting field, because there are two other cube roots of 2 not belonging to this field, namely $\sqrt[3]{2}\zeta$ and $\sqrt[3]{2}\zeta^2$, where $\zeta = e^{\frac{2\pi i}{3}} = -\frac{1}{2} + \frac{\sqrt{3}i}{2}$ is the primitive cube root of unity. Thus, a splitting field for f is

$$\mathbb{Q}\left(\sqrt[3]{2}\right)(\zeta) = \mathbb{Q}\left(\sqrt[3]{2}\right)(\sqrt{3}i).$$

Example 45.5

Consider the polynomial $f = x^3 + x + 1 \in \mathbb{Z}_2[x]$. It is evident that f is an irreducible polynomial. By Kronecker's Theorem 42.1 we can build the extension field

$$\begin{aligned}\mathbb{Z}_2[x]/\langle f\rangle &= \mathbb{Z}_2(\alpha)\\ &= \{0,\ 1,\ \alpha,\ 1+\alpha,\ \alpha^2,\ 1+\alpha^2,\ \alpha+\alpha^2,\ 1+\alpha+\alpha^2\}\end{aligned}$$

where $\alpha = \langle f\rangle + x$. This is a splitting field for f, because

$$f = (x+\alpha)\left(x+\alpha^2\right)\left(x+\alpha+\alpha^2\right).$$

You will check this factorization in Exercise 45.1 and also show that

$$\mathbb{Z}_2(\alpha) = \mathbb{Z}_2(\alpha^2) = \mathbb{Z}_2(\alpha+\alpha^2).$$

We will now prove that a non-constant polynomial over a field always has a splitting field; furthermore, such a field is unique (up to isomorphism, of course). We will prove existence (and more) in Theorem 45.1 and uniqueness in Theorem 45.2.

Theorem 45.1 *Let F be a field and $f \in F[x]$ a non-constant polynomial. Then there exists a splitting field K for f over F. Furthermore, if the linear factorization of f in $K[x]$ is given by*

$$f = \beta(x-\alpha_1)(x-\alpha_2)\cdots(x-\alpha_n),$$

Then K is equal to the finite extension $F(\alpha_1, \alpha_2, \cdots, \alpha_n)$.

Proof: Let F be a field and $f \in F[x]$ a non-constant polynomial. We know by Exercise 42.1 that we can construct a splitting extension M of F for f. Now consider the finite field extension $K = F(\alpha_1, \alpha_2, \cdots, \alpha_n)$, where the elements α_i are the roots of f in M. It is then evident that K is a splitting extension of F for f. But if L is any other splitting extension of F contained in M, then L must contain the roots of f, and so must contain K. This means that K is minimal among such extensions, and so is a splitting field. □

We now prove that splitting fields are unique, up to isomorphism. The following theorem carefully sets up what we mean by this uniqueness:

Theorem 45.2 *Suppose that F is a field and $f \in F[x]$ is a non-constant polynomial. Suppose that K and $\bar{K}$ are two splitting fields of F for f. Then there exists an isomorphism $\varphi : K \to \bar{K}$ that leaves F fixed.*

Proof: Assume that F, f, K, and $\bar{K}$ are as in the hypotheses of the theorem. We know from Theorem 45.1 that $K = F(\alpha_1, \alpha_2, \alpha_3, \cdots, \alpha_n)$, where the α_i are the (not necessarily distinct) roots of f in K. Furthermore, $\bar{K} = F(\beta_1, \beta_2, \cdots, \beta_n)$, where the β_i are the (not necessarily distinct) roots of f in $\bar{K}$. We will proceed by induction on the degree n. To facilitate this, define the fields K_k inductively, by letting $K_0 = F$, and $K_{k+1} = K_k(\alpha_{k+1})$. We will define the fields $\bar{K}_k$ in the same manner; we will let $\bar{K}_0 = F$ and define the subfields of $\bar{K}$ at each inductive step, by reordering the roots $\beta_1, \cdots, \beta_n$, which we will describe below. We will extend the field isomorphism φ from each of the extension fields K_k onto the extension fields $\bar{K}_k$, one step at a time.

We start with $\varphi : K_0 \to \bar{K}_0$ being the identity isomorphism (since $K_0 = \bar{K}_0 = F$). We now assume by induction that φ has in fact been defined and is an isomorphism from K_k onto $\bar{K}_k$. Furthermore, this isomorphism leaves F fixed and has been constructed so that $\varphi(\alpha_i) = \beta_i$, for $i = 1, 2, \cdots, k$.

Now consider the root α_{k+1} of the polynomial $f \in F[x] \subseteq K_k[x]$. If $\alpha_{k+1} \in K_k$, then $f \in F[x] \subseteq \bar{K}_k[x]$ must have a root in $\bar{K}_k$ other than $\beta_1, \beta_2, \cdots, \beta_k$, and, by renumbering, we may assume that this is the root β_{k+1}. We consequently have φ already defined on the field $K_{k+1} = K_k$ onto the field $\bar{K}_{k+1} = \bar{K}_k$.

We thus may assume that the root $\alpha_{k+1} \notin K_k$. We then know by Theorem 43.2 that $K_k(\alpha_{k+1})$ is isomorphic to the field $K_k[x]/\langle p \rangle$, where p is the irreducible factor of f in $K_k[x]$ for which α_{k+1} is a root. Now because the fields K_k and $\bar{K}_k$ are isomorphic, we know that the polynomial rings $K_k[x]$ and $\bar{K}_k[x]$ are isomorphic and there is an irreducible factor $\bar{p}$ of f in $\bar{K}_k[x]$ corresponding to p. (See Exercise 19.17.) Since $\bar{p}$ divides f in $\bar{K}[x]$, the roots of $\bar{p}$ are also roots of f in $\bar{K}$. So choose any root of $\bar{p}$ in $\bar{K}$, which we can (by renumbering) call β_{k+1} and let $\bar{K}_{k+1} = \bar{K}_k(\beta_{k+1})$. Then $\bar{K}_{k+1}$ is isomorphic to $\bar{K}_k[x]/\langle \bar{p} \rangle$. But by Exercise 19.20, it is clear that $K_k[x]/\langle p \rangle$ and $\bar{K}_k[x]/\langle \bar{p} \rangle$ are isomorphic, and the isomorphism between them extends the assumed isomorphism for K_k to $\bar{K}_k$, and it takes $\langle p \rangle + x$ to $\langle \bar{p} \rangle + x$. This means that we have an isomorphism from $K_{k+1} = K_k(\alpha_{k+1})$ to $\bar{K}_{k+1} = \bar{K}_k(\beta_{k+1})$, which takes α_{k+1} to β_{k+1}.

Thus, by induction, we have that there is an isomorphism from K to $\bar{K}$ that extends the isomorphism φ from F to $\bar{F}$, as required. □

Example 45.6

Suppose that $f \in \mathbb{Q}[x] \subseteq \mathbb{C}[x]$. Recall the Fundamental Theorem of Algebra 9.1, which implies that every polynomial with complex coefficients can be factored completely into linear factors. That is, we have in this case that all of the roots $\alpha_1, \cdots, \alpha_n$ of f belong to $\mathbb{C}$. Consequently, $\mathbb{C}$ contains a splitting field for f over $\mathbb{Q}$ (that is, $\mathbb{C}$ is a splitting extension for f over $\mathbb{Q}$), and it consists precisely of $\mathbb{Q}(\alpha_1, \cdots, \alpha_n)$. By Theorem 45.2, we know that this is essentially the only splitting field.

The next theorem will give us some insight into just how nice splitting fields are. It says that if K is a splitting field over F for f and K contains *any* root for an irreducible polynomial $g \in F[x]$, then K will contain *all* roots for g. That is, g will also split in $K[x]$. The proof of this theorem relies heavily on the uniqueness of splitting fields.

Theorem 45.3 *Let K be the splitting field over the field F for the polynomial f. Suppose that $g \in F[x]$ is an irreducible polynomial over F and $\gamma \in K \backslash F$ is a root for g. Then g factors completely into linear factors in $K[x]$.*

Proof: Suppose that K is the splitting field over the field F for the polynomial f. We know that $K = F(\alpha_1, \alpha_2, \cdots, \alpha_n)$, where the α_i are the roots of f in K. Consider an irreducible polynomial $g \in F[x]$, and suppose that $\gamma \in K \backslash F$ with $g(\gamma) = 0$. Let's assume by way of contradiction that g does not completely factor into linear factors inside $K[x]$. But then by Kronecker's Theorem 42.1 we can construct a strictly larger extension field $K(\beta)$ of K, for which β is a root of g. Now because γ and β are both roots of the irreducible polynomial g, we know by Theorem 43.2 that there is an isomorphism between the fields $F(\gamma)$ and $F(\beta)$ that leaves F fixed. Now $K(\beta) = F(\alpha_1, \cdots, \alpha_n)(\beta) = F(\alpha_1, \cdots, \alpha_n, \beta) = F(\beta)(\alpha_1, \cdots, \alpha_n)$, and this is the splitting field of f over $F(\beta)$. But because $F(\beta)$ and $F(\gamma)$ are isomorphic, there is an isomorphism between $K(\beta)$ and $K = K(\gamma) = F(\gamma, \alpha_1, \cdots, \alpha_n)$, since the latter field is the splitting field for f over $F(\gamma)$.

But now we count degrees. By our isomorphism between K and $K(\beta)$ as splitting fields over the isomorphic fields $F(\gamma)$ and $F(\beta)$ we have that $[K : F(\gamma)] = [K(\beta) : F(\gamma)]$. But then by Theorem 44.2

$$[K : F] = [K : F(\gamma)][F(\gamma) : F] = [K(\beta) : F(\gamma)][F(\gamma) : F] = [K(\beta) : F],$$

and this means that $[K(\beta) : K] = 1$, and so $\beta \in K(\beta) = K$, which is as we wish. □

Suppose that K is an extension field of the field F, and that whenever an irreducible polynomial $f \in F[x]$ has one root in K, then it splits in K. In this case we say that K is a **normal extension** of F. With this terminology, we can rephrase Theorem 45.3 by saying that if K is a splitting field over the field F for the polynomial f, then K is in fact a normal extension. Later in this chapter we will encounter Theorem 45.6, which asserts that for a field with characteristic zero, being a splitting field is *equivalent* to being a finite normal extension.

In Chapter 48 we will discover that the concept of normal extension is actually closely related to the concept of normal subgroup!

45.2 Fields with Characteristic Zero

Theorems 45.1 and 45.2 together say that given any field F and a nonconstant polynomial f, we can always build a unique minimal extension field, in which f completely factors into linear factors. So, for any polynomial over a field, we obtain the unique (up to isomorphism) splitting field for the polynomial merely by adjoining the roots of the polynomial, by repeated application of Kronecker's Theorem 42.1. In general, some of these roots may already belong to the base field, and some of the roots may be repeated. However, if the polynomial is irreducible, the situation is particularly nice (for fields of characteristic zero).

Theorem 45.4 *Let F be a field of characteristic zero and f an irreducible polynomial in $F[x]$. Then the roots of f in its splitting field are all distinct.*

Proof: Suppose that f has a repeated root α in its splitting field K. This means that $f = (x - \alpha)^k g$, where k is an integer greater than one, and g is a polynomial over K. We will now make use of the formal derivative of this polynomial. We first encountered this idea in Exercise 4.7 in the polynomial ring $\mathbb{Q}[x]$, and in Exercise 45.3 you will check that the appropriate results hold, for any field (and in fact for any commutative ring). In particular, the product rule like you encountered in calculus works here, and so the formal derivative f' can be computed thus:

$$f' = k(x - \alpha)^{k-1} g + (x - \alpha)^k g'.$$

Now because K has characteristic zero, this is necessarily a non-zero polynomial. We then have that $x - \alpha$ is a factor of both f and f'. But if we use term-by-term differentiation instead, it is clear that $f' \in F[x]$. Since f is irreducible, we know that in $F[x]$ we have $\gcd(f, f') = 1$, and so by the GCD identity for $F[x]$ we can find polynomials $a, b \in F[x]$ for which $1 = af + bf'$. But if we evaluate this polynomial at α we obtain the following:

$$1 = 1(\alpha) = a(\alpha)f(\alpha) + b(\alpha)f'(\alpha) = 0 + 0 = 0,$$

a contradiction. Thus f must not have any repeated roots in the splitting field K. □

Theorem 45.4 remains true for all finite fields, and you will prove this fact in Exercise 46.8. The theorem is false, however, for infinite fields with characteristic p (although we will not pursue this topic in this book). We will find Theorem 45.4 of considerable use when we return to the problem of determining when polynomial equations can be solved with radicals.

We close this section by proving another important result about fields with characteristic zero. It is surprisingly the case that for such fields any finite algebraic extension is simple. That is, if we adjoin finitely many algebraic numbers to a field with characteristic zero, we only need to adjoin a single element. In particular, over such a field the splitting field is always a simple extension.

Example 45.7

An interesting example of this is Example 44.3, where we rather laboriously proved that $\mathbb{Q}(\sqrt{2}, \sqrt[3]{2})$ is a simple extension $\mathbb{Q}(\alpha)$ of $\mathbb{Q}$, where $\alpha = \sqrt{2} + \sqrt[3]{2}$. As we shall see in the proof of the theorem, it is no accident that α is a linear combination of the elements $\sqrt{2}, \sqrt[3]{2}$.

Theorem 45.5 *Let F be a field with characteristic zero, and K a finite algebraic extension of F. Then K is a simple algebraic extension of F.*

Proof: Suppose that α and β are algebraic elements over a field F, which has characteristic zero. We will show that $F(\alpha, \beta) = F(\mu)$, for some algebraic element μ. This reduction from two generators to one is clearly the induction step needed to show that any finite algebraic extension is simple (See Exercise 45.4).

Now α and β are roots of irreducible polynomials $f, g \in F[x]$ of degrees n and m, respectively. We may as well do all of our computations in a field K in which both f and g split into linear factors. Because F has characteristic zero, Theorem 45.4 tells us that the roots $\alpha = \alpha_1, \cdots, \alpha_n$ of f in K are all distinct from one another, and that the roots $\beta = \beta_1, \cdots, \beta_m$ of g in K are also all distinct from one another.

Consider the finite set of elements in the field K:

$$\left\{\frac{\alpha_j - \alpha}{\beta - \beta_i}\right\},$$

where $i > 1$ and $j > 1$. Since F is of characteristic zero, it is an infinite field, and so we can choose a non-zero element $a \in F$ not equal to any of these quotients.

We will now define $\mu = \alpha + a\beta$. It is quite evident that $F(\mu) \subseteq F(\alpha, \beta)$. We must show the reverse inclusion is true, by arguing that $\alpha, \beta \in F(\mu)$. For that purpose, consider the polynomial $h \in F(\mu)[x]$, defined by $h(x) = f(\mu - ax)$. Notice that

$$h(\beta) = f(\mu - a\beta) = f(\alpha) = 0.$$

Thus h and g share the root β.

However, h and g can share no other root. For if they did, it would be some β_i, with $i > 1$. But then

$$0 = h(\beta_i) = f(\mu - a\beta_i) = f(\alpha + a\beta - a\beta_i).$$

This would mean that $\alpha + a(\beta - \beta_i) = \alpha_j$, for some j, and we chose a precisely so that this cannot be true.

Because g factors into linear factors in K, it is now evident that the gcd of h and g in $K[x]$ is $x - \beta$. But $h, g \in F(\mu)[x]$, and so when we perform Euclid's algorithm to compute $\gcd(h, g)$, we will remain in the ring $F(\mu)[x]$. This means that $x - \beta \in F(\mu)[x]$, or in other words, $\beta \in F(\mu)$. But because $\alpha = \mu - a\beta$, $\alpha \in F(\mu)$, too. Thus, $F(\alpha, \beta) = F(\mu)$. □

Examples of this theorem appear in Example 44.3, and Exercises 44.3 and 44.4.

We can now use this theorem to show that for fields of characteristic zero, a finite extension is normal exactly if it is the splitting field for some irreducible polynomial.

Theorem 45.6 *Suppose that F is a field of characteristic zero, with finite extension K. Then K is the splitting field for some irreducible $f \in F[x]$ if and only if K is a normal extension of F.*

Proof: Suppose that K is a normal extension of F, a field with characteristic zero. Then by Theorem 45.5 $K = F(\alpha)$, where α is algebraic over F. Let $f \in F[x]$ be the minimal polynomial for α. Since K is normal, f splits in K. But clearly no smaller subfield of K contains all the roots of f, because $K = F(\alpha)$. Thus K is the splitting field for f over F.

The converse is just Theorem 45.3. □

Theorems 47.2 and 48.2 will give other conditions equivalent to being a finite normal extension.

Chapter Summary

In this chapter we proved that every polynomial over a base field admits a unique minimal field extension of that field, in which the polynomial factors into linear factors; this is called its splitting field. A splitting field for a polynomial actually has the stronger property that if an irreducible polynomial over the base field has even one root in the splitting field, it has all of its roots. We call such field extensions normal. We also showed that for a field of characteristic zero, finite extensions are necessarily simple extensions.

Warm-up Exercises

a. Answer the following true or false; if your answer is false, give a counterexample. Assume that F, K and L are fields.

(a) A splitting field for f over F is an algebraic extension of F.

(b) All finite extensions of $\mathbb{Q}$ are simple.

(c) A normal extension of $\mathbb{Q}$ is the splitting field for some $f \in \mathbb{Q}[x]$.

(d) A splitting field for f over F is a normal extension.

(e) If $f \in \mathbb{Q}[x]$ is irreducible, then all of the roots of f in $\mathbb{C}$ are distinct.

(f) Suppose that $f \in F[x]$ is irreducible and has degree bigger than 1. To construct the splitting field for f over F, it is always possible to use the field $F[x]/\langle f \rangle$ constructed by Kronecker's Theorem 42.1.

(g) Suppose that $f \in F[x]$ is irreducible and has degree bigger than 1. To construct the splitting field for f over F, it is never possible to use the field $F[x]/\langle f \rangle$ constructed by Kronecker's Theorem 42.1.

(h) All splitting fields are finite extensions.

(i) If $F \subseteq L \subseteq K$, and K is the splitting field for f over F, then K is the splitting field for f over L.

(j) Suppose K is a proper finite extension of F. Then K is the splitting field for some polynomial f over F.

b. Consider the polynomial $x^3 - 2$. Describe its splitting field over the following fields:

$$\mathbb{Q}, \quad \mathbb{R}, \quad \mathbb{Q}(i), \quad \mathbb{C}$$

c. Let F be a field, and suppose that $f \in F[x]$ has degree 1. What is the splitting fiuld of f over F?

Exercises

1. In this problem we check the claims made in Example 45.5.

(a) Show that the polynomial f factors as

$$f = (x+\alpha)\left(x+\alpha^2\right)\left(x+\alpha+\alpha^2\right).$$

(b) Prove that $\mathbb{Z}_2(\alpha) = \mathbb{Z}_2(\alpha^2)$.

(c) Prove that $\mathbb{Z}_2(\alpha) = \mathbb{Z}_2(\alpha+\alpha^2)$.

2. In this problem you will follow the proof of Theorem 45.5 in the particular case where $\alpha = \sqrt{2}$, $\beta = \sqrt{7}$ and $F = \mathbb{Q}$.

(a) Make a complete list of the elements of $\mathbb{Q}(\sqrt{2}, \sqrt{7})$ that are not allowed to be the element a. Choose your own value of a (there are many choices) and find an appropriate value for μ.

(b) Determine the polynomial $h \in \mathbb{Q}(\mu)[x]$ and check that β is a root.

(c) Compute the gcd of h and g in $\mathbb{Q}(\mu)[x]$ using Euclid's algorithm, and thus verify the claim about this gcd in the proof.

3. Suppose that R is a commutative ring, and $f \in R[x]$. Then we can write f as $f = a_n x^n + a_{n-1}x^{n-1} + a_1 x + a_0$, where the coefficients a_i are elements of R. Define the formal derivative of f as

$$f' = na_n x^{n-1} + (n-1)a_{n-1}x^{n-2} + \cdots + a_1,$$

which is just the formula we expect from calculus. (Note that we explored this notion for $R = \mathbb{Q}$ in Exercise 4.7.)

(a) Suppose that $a \in R$ and $f \in R[x]$. Prove that $(af)' = af'$. (The *constant multiple rule.*)

(b) Suppose that $f, g \in R[x]$. Prove that $(f+g)' = f' + g'$. (The *sum rule.*)

(c) Suppose that $f, g \in R[x]$. Then $(fg)' = f'g + fg'$. (The *product rule.*)

(d) Suppose that $a \in R$ and n is a positive integer. Prove that if $f = (x-a)^n$, then $f' = n(x-a)^{n-1}$. (The *power rule.*)

4. Complete the inductive proof for Theorem 45.5: Suppose $K = F(\alpha_1, \cdots, \alpha_n)$, for algebraic elements α_i. Show that there is an algebraic element μ so that $K = F(\mu)$.

5. Consider the polynomial $f = x^4 - 4x^2 + 2 \in \mathbb{Q}[x]$.

   (a) Argue that f is irreducible over $\mathbb{Q}$.

   (b) Factor f completely into linear factors in $\mathbb{C}[x]$.

   (c) Describe the splitting field of f over $\mathbb{Q}$ in $\mathbb{C}$.

   (d) This splitting field can actually be described in the form $\mathbb{Q}(\alpha)$, where α is one of the roots of f. Do so if you haven't done so already, justifying your assertion rigorously.

6. Consider the polynomial $f = x^4 - 2 \in \mathbb{Q}[x]$.

   (a) Argue that f is irreducible over $\mathbb{Q}$.

   (b) Factor f completely into linear factors in $\mathbb{C}[x]$.

   (c) Describe the splitting field of f over $\mathbb{Q}$ in $\mathbb{C}$. (The easiest such description is of the form $\mathbb{Q}(\alpha, \beta)$, where α, β are appropriately chosen elements of $\mathbb{C}$.)

   (d) This splitting field *cannot* be described in the form $\mathbb{Q}(\alpha)$, where α is one of the roots of f. Prove that this is true.

   (e) Why does Theorem 45.5 imply that this splitting field is a simple extension of $\mathbb{Q}$? Give such a description, justifying your assertion rigorously.

7. Suppose that F is a field, and K is a field extension of F that is *algebraically closed* (see Exercise 9.26 for a definition). Suppose that $f \in F[x]$. Argue that K contains a copy of the splitting field of f over F. Describe this splitting field more explicitly.

Chapter 46

Finite Fields

In the previous chapter we proved that any irreducible polynomial f over a field F has a unique splitting field: a minimal field extension of F in which f can be factored into linear factors. This provides a field inside of which we can explore whether or not the roots of f are obtainable by elementary algebraic operations. We will pursue this goal in the remaining chapters in this book.

But a wonderful bonus flows from the existence and uniqueness of splitting fields, and we will take a small detour from our goal to explore this bonus in the present chapter. We are now able to completely describe all finite fields. We have for a long time been familiar with the finite fields $\mathbb{Z}_p$, where p is a prime integer. We have also encountered various finite fields as finite extensions of such fields; for example, consider Example 42.5, Exercise 43.2 and Example 45.5. In this chapter we will be able to place these examples in a beautiful general context.

46.1 Existence and Uniqueness

Theorem 42.2 says that every field has characteristic zero or p, where p is a prime integer. Fields with characteristic zero have a subfield isomorphic to $\mathbb{Q}$ and so are infinite. Thus, any finite field has characteristic p, for some prime p. This means that every finite field contains (an isomorphic copy of) one of the fields $\mathbb{Z}_p$ as a subfield (recall that this is called the prime subfield). We use these considerations to prove our first result about arbitrary finite fields.

Theorem 46.1 *A finite field of characteristic p has p^n elements, for some positive integer n.*

Proof: With our knowledge of group theory, this becomes quite an easy theorem. Let F be a finite field, and consider its additive group structure. Saying that its characteristic is p means exactly that every non-zero element has (additive) order p. But then Theorem 31.5 says that p must be the only prime integer dividing the order of F. That is, every finite field has p^n elements, for some positive integer n. □

But do such fields exist, for every prime p, and every positive integer n? The next theorem says that they do. Here, field theory comes to the rescue.

Theorem 46.2 *For every prime integer p and every positive integer n, there exists a field with p^n elements.*

Proof: Consider the polynomial $f = x^{p^n} - x \in \mathbb{Z}_p[x]$. By Theorem 45.1 we know that there exists a splitting field F of $\mathbb{Z}_p$ for f. Because F is a field extension of $\mathbb{Z}_p$, it must be a field of characteristic p. We will show that F is exactly the required field of p^n elements.

Pick two roots $r, s \in F$ of the polynomial f. We claim that $r-s$ is also a root of f. To check this, we must evaluate $f(r-s) = (r-s)^{p^n} - (r-s)$. But in a field of characteristic p, we know that $(r-s)^{p^n} = r^{p^n} - s^{p^n}$ (see Exercise 46.1c). Thus,

$$f(r-s) = r^{p^n} - s^{p^n} - r + s = \left(r^{p^n} - r\right) - \left(s^{p^n} - s\right) = 0 - 0 = 0,$$

and so $r - s$ is a root of f, as claimed. That is, the set of roots of f is an (additive) group.

Now pick two roots $r, s \in F$ of the polynomial f, with $s \neq 0$. We claim that rs^{-1} is also a root of f. But

$$f(rs^{-1}) = r^{p^n} s^{-p^n} - rs^{-1} = rs^{-1} - rs^{-1} = 0.$$

Thus, the set of roots of f actually forms a *subfield* of F. But because a splitting field is minimal among all fields containing the roots, F consists exactly of the roots of f.

To show that F contains exactly p^n elements, it remains to check that f has no repeated roots. Pick any root r of f. Then

$$\begin{aligned} f(x-r) &= (x-r)^{p^n} - (x-r) \\ &= x^{p^n} - x - (r^{p^n} - r) = f(x) - 0 = f(x). \end{aligned}$$

Thus,

$$f = (x-r)((x-r)^{p^n-1} - 1).$$

But clearly $x - r$ is not a factor of $(x-r)^{p^n-1} - 1$ (because $x - r$ does not divide 1). Thus, r is not a repeated root. Consequently, f has p^n roots, and so the set of roots of f is equal to the splitting field F and is the required field with p^n elements. □

We shall now use Theorem 45.2 (the uniqueness of splitting fields) to show that any two finite fields with p^n elements are isomorphic:

Theorem 46.3 *All fields with p^n elements are isomorphic.*

Proof: Any finite field F with p^n elements has characteristic p, and so is a finite extension of $\mathbb{Z}_p$. Consider again the polynomial $f = x^{p^n} - x \in \mathbb{Z}_p[x]$. We will now show that the elements of F are exactly the distinct roots of f, and so is exactly the splitting field of f over $\mathbb{Z}_p$.

We know that the multiplicative group F^* has $p^n - 1$ elements. Now pick $a \in F^*$. By Lagrange's Theorem we know that the order of a divides $p^n - 1$; thus, $a^{p^n-1} = 1$. But then $a^{p^n} = a$, and so a is a root of of $x^{p^n} - x$. Thus, every element of F is a root of $x^{p^n} - x$. The field F has p^n elements and $x^{p^n} - x$ can have no more than p^n roots. Hence, F consists of exactly those roots.

But then the uniqueness of splitting fields (Theorem 45.2) means that all fields with p^n elements are isomorphic. □

Because the field with p^n elements is unique (up to isomorphism), we can unambiguously denote it by $GF(p^n)$. This stands for **Galois field of order** p^n, after the Frenchman Evariste Galois who was the first mathematician to consider finite fields. Of course, $GF(p)$ is merely a different notation for the familiar field $\mathbb{Z}_p$.

The uniqueness of finite fields was proved in 1893 by the American mathematician E. H. Moore, who was by that time at the University of Chicago. Moore played a crucial part in bringing American mathematics to the level of European mathematics. He was the founder of the mathematics department at the University of Chicago and was one of the founders of the American Mathematical Society. But most important was his work as an admired mathematician and as a teacher and mentor for the next generation of American mathematicians.

46.2 Examples

Example 46.1

Consider the polynomial $x^2+1 \in \mathbb{Z}_3[x]$. It is easy to see that this polynomial has no roots in $\mathbb{Z}_3$, by checking the three possibilities 0, 1, and 2. Thus, by the Root Theorem, x^2+1 is irreducible, and so the ideal $\langle x^2+1\rangle$ is maximal (Theorem 13.3). Thus, $\mathbb{Z}_3[x]/\langle x^2+1\rangle$ is a field. This field has degree 2 over $\mathbb{Z}_3$ and so has 9 elements. It is consequently the Galois field $GF(9)$. We should really write these elements as (additive) cosets. For simplicity's sake, we will replace $\langle x^2+1\rangle + x$ by α. With this convention, the elements of this field are

$$\{0,\ 1,\ 2,\ \alpha,\ \alpha+1,\ \alpha+2,\ 2\alpha,\ 2\alpha+1,\ 2\alpha+2\},$$

where the multiplication is determined by the law $\alpha^2+1=0$; that is, $\alpha^2=2$.

Let's look at the multiplicative structure of this field. That is, what sort of group is $GF(9)^*$? It is certainly an abelian group with 8 elements, and up to isomorphism there are three such groups. It turns out that this group is cyclic, and $\alpha+1$ is a generator. Let's verify this, by brute calculation:

$$(\alpha+1)^1=\alpha+1, \quad (\alpha+1)^2=\alpha^2+2\alpha+1=2\alpha,$$

$$(\alpha+1)^3=2\alpha(\alpha+1)=2\alpha^2+2\alpha=2\alpha+1,$$

$$(\alpha+1)^4=(\alpha+1)(2\alpha+1)=2\alpha^2+3\alpha+1=2,$$

$$(\alpha+1)^5=2(\alpha+1)=2\alpha+2,$$

$$(\alpha+1)^6=(\alpha+1)(2\alpha+2)=2\alpha^2+4\alpha+2=\alpha,$$

$$(\alpha+1)^7=(\alpha+1)\alpha=\alpha^2+\alpha=\alpha+2,$$

$$(\alpha+1)^8=(\alpha+1)(\alpha+2)=\alpha^2+3\alpha+2=1.$$

It turns out that the group of units of *any* finite field is cyclic! This is called the Primitive Root Theorem; we will not prove this theorem in this book.

Example 46.2

In order to construct the field $GF(8)$, we need only find an irreducible polynomial f in $\mathbb{Z}_2[x]$ of degree three; then form $\mathbb{Z}_2[x]/\langle f\rangle$. We actually performed this calculation in Example 45.5 (see also Exercise 45.1), using the polynomial $f=x^3+x+1$. In Exercise 46.9 you will use the polynomial x^3+x^2+1 instead, and show explicitly that the resulting field is isomorphic to the field constructed in Example 45.5.

Now consider the polynomial $g=x^2+x+1 \in GF(8)[x]$. We claim that this is an irreducible polynomial in $GF(8)[x]$. To show this, let's try by brute force to see whether any of the 8 elements of $GF(8)$ are roots of the polynomial. For example,

$$g(\alpha^2+\alpha)=(\alpha^2+\alpha)^2+(\alpha^2+\alpha)+1=$$

$$\alpha^4+\alpha^2+\alpha^2+\alpha+1=\alpha^4+\alpha+1=$$

$$\alpha(\alpha+1)+\alpha+1=\alpha^2+1\neq 0.$$

Thus, $\alpha^2+\alpha$ is not a root.

▷ **Quick Exercise.** Try the other seven elements of the field, and thus see that this polynomial is irreducible. ◁

But then we can construct the field

$$GF(8)/\langle x^2+x+1\rangle.$$

This is a degree 2 extension of $GF(8)$, which has 8 elements, and so is a field with 64 elements altogether. Thus we have constructed the essentially unique field with 64 elements $GF(2^6)=GF(64)$. Note of course that we could have also constructed this field, had we been able to find an irreducible polynomial in $\mathbb{Z}_2[x]$ of degree 6. You will find such a polynomial in Exercise 46.10.

The previous example shows that the unique field with 2^3 elements can be considered a subfield of the unique field with 2^6 elements. The full story is given by the following theorem, which you prove in Exercise 46.5.

Theorem 46.4 *The field $GF(p^m)$ is a subfield of $GF(p^n)$ if and only if m divides n.*

Chapter Summary

In this chapter we prove that for every positive prime integer p and positive integer n, there exists a unique finite field with exactly p^n elements. This is a nice consequence of the uniqueness of splitting fields.

Warm-up Exercises

a. Describe how you would construct the field $GF(25)$; specify a polynomial $p \in \mathbb{Z}_5[x]$ that you could use.

b. Consider the multiplicative group $GF(32)^*$. Why is this group obviously cyclic? (You need not refer to the Primitive Root Theorem.)

c. Give three non-isomorphic commutative rings with nine elements. How many are fields? Are they isomorphic as abelian groups, or not?

Exercises

1. Suppose that F is a field with characteristic p, and $a, b \in F$.

   (a) Prove that $(a+b)^p = a^p + b^p$. *Hint:* Use Exercise 2.19 and the binomial theorem Exercise 6.17.

   (b) For any positive integer n, prove that $(a+b)^{p^n} = a^{p^n} + b^{p^n}$.

   (c) For any positive integer n, prove that $(a-b)^{p^n} = a^{p^n} - b^{p^n}$.

2. Explicitly write out the elements of $GF(25)$, as constructed along the lines of Exercise c above. Find a generator for the cyclic multiplicative group $GF(25)^*$.

3. Define the function

$$\varphi : GF(p^n) \to GF(p^n)$$

by $\varphi(a) = a^p$. Prove that this is a ring isomorphism, called the **Frobenius** isomorphism (this is actually a repeat of Exercise 17.18).

4. Describe how to construct $GF(81)$ by using a specific polynomial $p \in GF(9)[x]$.

5. Prove Theorem 46.4. That is, show that $GF(p^m)$ is a subfield of $GF(p^n)$ if and only if m divides n.

6. Prove that every element in a finite field with characteristic p has a pth root.

7. Suppose that $f \in GF(p^n)[x]$ is a non-constant polynomial and suppose that its formal derivative is zero. Prove that f is not irreducible.

8. Suppose that F is a finite field and f is an irreducible polynomial in $F[x]$. Prove that the roots of f in its splitting field are all distinct. (That is, prove Theorem 45.4, replacing the hypothesis that the field is characteristic zero, with the hypothesis that it is finite.)

9. In this exercise we follow up on Example 46.2, where we constructed $GF(8)$ as $\mathbb{Z}_2[x]/\langle f \rangle$, with $f = x^3 + x + 1$. Then $GF(8) = \mathbb{Z}_2(\alpha)$, where $\alpha = \langle f \rangle + x$.

   (a) We know from the proof of Theorem 46.2 that $GF(8)$ consists exactly of the 8 distinct roots of the polynomial $x^8 - x \in \mathbb{Z}_2[x]$. Explain why f is a factor of this polynomial.

   (b) Factor $x^8 - x$ completely in $\mathbb{Z}_2[x]$ as a product of irreducible polynomials. (You should obtain a cubic irreducible polynomial g other than f as a factor.)

   (c) Factor g into linear factors in $\mathbb{Z}_2(\alpha)[x]$.

   (d) Why are $\mathbb{Z}_2[x]/\langle f \rangle$ and $\mathbb{Z}_2[x]/\langle g \rangle$ isomorphic as fields? (Your answer should just involve a citation of the appropriate theorem(s).)

   (e) Let $\beta = \langle g \rangle + x$. Construct an explicit isomorphism between the fields $\mathbb{Z}_2(\alpha)$ and $\mathbb{Z}_2(\beta)$. You should specify a function from one of these fields to the other, including precisely where each of the eight elements goes.

10. In Example 46.2 we constructed $GF(64)$ by building $GF(8)$ first, and then constructing a degree two extension of that field. The alternative is to find a degree six irreducible polynomial $h \in \mathbb{Z}_2[x]$, and then construct the field $\mathbb{Z}_2[x]/\langle h \rangle$.

(a) Give a complete list (with justification) of all degree 1, 2 or 3 irreducible polynomials in $\mathbb{Z}_2[x]$.

(b) Determine all fourth degree irreducible polynomials in $\mathbb{Z}_2[x]$.

(c) Determine all sixth degree irreducible polynomials in $\mathbb{Z}_2[x]$.

Chapter 47

Galois Groups

We now return to our goal of understanding whether the roots of an irreducible polynomial over a field can be obtained by elementary algebraic computations. In Chapter 45 we constructed the unique splitting field for such a polynomial, inside of which such computations must occur. In the present chapter we will look closely at what sort of field extension the splitting field must be. We will use group theory to do this.

In Chapters 22 and 23 we saw how geometry could be illuminated by considering functions leaving geometric properties fixed; we thus obtained groups of symmetries. Here we will illuminate field extensions (and splitting fields in particular) by considering functions leaving field properties fixed; we will thus obtain groups of automorphisms called Galois groups.

47.1 The Galois Group

To better understand a field F, it is useful to consider all functions that preserve essential algebraic structure. Such functions are one-to-one and onto ring homomorphisms from the field to itself; such homomorphisms are called **automorphisms**. We have denoted the set of all such automorphisms $\text{Aut}(F)$; this set is a group under functional composition (see Example 24.18).

Suppose that E and F are fields, and $E \supseteq F$. We may now consider the following subset of $\text{Aut}(F)$:

$$\text{Gal}(E|F) = \{\varphi \in \text{Aut}(E) : \varphi(f) = f, \text{for all } f \in F\}.$$

That is, we are considering only those automorphisms of the field E that leave all elements of the subfield F *fixed.* It is easy to check that $\text{Gal}(E|F)$ is a subgroup of $\text{Aut}(E)$, using the Subgroup Theorem 25.2.

▷ **Quick Exercise.** Show that $\mathrm{Gal}(E|F)$ is a subgroup of $\mathrm{Aut}(E)$. ◁

We call $\mathrm{Gal}(E|F)$ the **Galois Group of the field** E **over** F. Let's consider some examples.

Example 47.1

$\mathrm{Gal}(\mathbb{R}|\mathbb{Q})$ is the trivial group, because $\mathrm{Aut}(\mathbb{R})$ itself is the group with only one element (see Exercise 24.13).

Example 47.2

Let K be any field, with prime subfield F. The prime subfield is the field generated from 1 by field operations, and so any automorphism must leave it fixed. This means that $\mathrm{Gal}(K|F) = \mathrm{Aut}(K)$.

Example 47.3

$\mathrm{Gal}(\mathbb{C}|\mathbb{R})$ is the two element group $\{\iota, \varphi\}$, where ι is the identity automorphism, and φ is the complex conjugation map. This is true because $\mathrm{Aut}(\mathbb{C})$ has only these two elements (Exercise 24.14), and both such automorphisms leave the real subfield fixed.

Example 47.4

Let's compute the Galois group $\mathrm{Gal}(\mathbb{Q}(\sqrt{2})|\mathbb{Q})$. Because $\mathbb{Q}$ is the prime subfield of $\mathbb{Q}(\sqrt{2})$, it is left fixed by any automorphism (Example 47.2). Since any element of $\mathbb{Q}(\sqrt{2})$ is of the form $a + b\sqrt{2}$, where $a, b \in \mathbb{Q}$, it is clear than any automorphism φ of this field is determined by $\varphi(\sqrt{2})$. But

$$2 = \varphi(2) = \varphi\left(\sqrt{2}^2\right) = (\varphi(\sqrt{2}))^2,$$

and so $\varphi(\sqrt{2})$ must be a square root of 2. There are only two choices: $\varphi(\sqrt{2}) = \sqrt{2}$ and $\varphi(\sqrt{2}) = -\sqrt{2}$. The first of these choices leads to the identity automorphism. The second leads to $\varphi(a + b\sqrt{2}) = a - b\sqrt{2}$. You can check that this is an automorphism of $\mathbb{Q}(\sqrt{2})$ leaving $\mathbb{Q}$ fixed, and so $\mathrm{Gal}(\mathbb{Q}(\sqrt{2})|\mathbb{Q})$ is a two element group.

▷ **Quick Exercise.** Check explicitly that the set of these two automorphisms forms a group. ◁

We will generalize this example in Theorem 47.1 below.

Example 47.5

Consider the field $\mathbb{Q}(\sqrt{2}, \sqrt{3})$. Now every element of this field can be expressed as $a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6}$ (see Example 44.2). Thus every automorphism φ of $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ is determined by what it does to $\varphi(\sqrt{2}), \varphi(\sqrt{3}), \varphi(\sqrt{6})$. Since $\sqrt{6} = (\sqrt{2})(\sqrt{3})$, we actually need only know $\varphi(\sqrt{2})$ and $\varphi(\sqrt{3})$. By an argument similar to that in Example 47.4, we have two choices for each of these. These lead to exactly four possibilities:

$$\begin{aligned} \iota, \qquad \varphi_1(a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6}) &= a - b\sqrt{2} + c\sqrt{3} - d\sqrt{6}, \\ \varphi_2(a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6}) &= a - b\sqrt{2} - c\sqrt{3} + d\sqrt{6}, \\ \varphi_3(a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6}) &= a + b\sqrt{2} - c\sqrt{3} - d\sqrt{6}. \end{aligned}$$

You can check directly that these are all automorphisms. We thus have a group with four elements. It is easy to check that each of these elements has order 2, and so this group is (up to isomorphism) the Klein Four Group (see the discussion in Section 31.3).

▷ **Quick Exercise.** Check that each φ_i is in fact an automorphism, and that each such has order 2. ◁

We will now obtain a theorem generalizing the arguments made in Examples 47.3 and 47.4:

Theorem 47.1 *Let $F \subseteq K$ be fields, and $f \in F[x]$ an irreducible polynomial, and $\alpha \in K \backslash F$ a root of f. Suppose that $\varphi \in \mathrm{Gal}(F(\alpha)|F)$. Then φ is entirely determined by $\varphi(\alpha)$. Furthermore, $\varphi(\alpha)$ must be a root of f in K, and so*

$$|\mathrm{Gal}(F(\alpha)|F)| \leq \deg(f) = [F(\alpha) : F].$$

Proof: Let $\deg(f) = n$. We know from Theorem 43.3 that every element of $F(\alpha)$ can be written (uniquely) in the form

$$\beta = b_0 + b_1\alpha + \cdots + b_{n-1}\alpha^{n-1},$$

where $b_i \in F$. But then

$$\varphi(\beta) = \varphi(b_0) + \varphi(b_1)\varphi(\alpha) + \cdots + \varphi(b_{n-1})\varphi\left(\alpha^{n-1}\right) =$$

$$b_0 + b_1\varphi(\alpha) + \cdots + b_{n-1}\varphi(\alpha)^{n-1};$$

this follows because φ is a ring homomorphism, and $\varphi(b_i) = b_i$ since $b_i \in F$. It then follows that φ is entirely determined by what it does to α.

But α is a root of $f = a_0 + a_1x + \cdots + a_nx^n$, and so by a similar argument

$$f(\varphi(\alpha)) = a_0 + a_1\varphi(\alpha) + \cdots + a_n\varphi(\alpha)^n = \varphi(f(\alpha)) = 0.$$

This means that $\varphi(\alpha)$ is also a root of f. Since $\deg(f) = n$, there are at most n choices for $\varphi(\alpha)$. It then follows that $|\mathrm{Gal}(F(\alpha)|F)| \leq \deg(f)$. The irreducibility of f implies that $\deg(f) = [F(\alpha) : F]$. □

This theorem provides us with a practical approach for computing Galois groups of simple algebraic extensions. Because each element of the Galois group takes each root of f to a root of f, and since an automorphism is of course a one-to-one function, we can (up to isomorphism) view the Galois group as a subgroup of the group of permutations of the roots of f. In the examples that follow, we shall usually take this point of view.

▷ **Quick Exercise.** Review the previous two examples in light of this theorem and the observations that follow it. ◁

Example 47.6

Let's reconsider Example 45.4 in light of this theorem. Consider the irreducible polynomial $x^3 - 2 \in \mathbb{Q}[x]$ and the field extension $\mathbb{Q}(\sqrt[3]{2}) \subset \mathbb{R}$ of the rational numbers $\mathbb{Q}$. Theorem 47.1 says that $\mathrm{Gal}(\mathbb{Q}(\sqrt[3]{2})|\mathbb{Q})$ has at most three elements. But in this case the Galois group is trivial, because only one of the three roots of $x^3 - 2$ is a real number, and so is the only root in $\mathbb{Q}(\sqrt[3]{2}) \subset \mathbb{R}$.

47.2 Galois Groups of Splitting Fields

In light of the previous example and theorem, it seems natural to ask when the Galois group of an algebraic simple field extension is as large as possible (namely, is equal to the degree of the field extension). It turns out that we already have the appropriate concept at hand: this takes place exactly when the extension is normal (at least for fields of characteristic zero). This is the content of the next theorem:

Theorem 47.2 *Let K be a finite extension of the field F, which is of characteristic zero. Then $|\mathrm{Gal}(K|F)| = [K : F]$ if and only if K is a normal extension of F.*

Proof: Suppose first that K is a normal extension of the field F with characteristic zero. We know from Theorem 45.5 that there exists an algebraic element β so that $K = F(\beta)$. Because K is a normal extension, K is the splitting field for the minimal polynomial g for β over F; let's suppose that the degree of g is m. We know from Theorem 43.3 that $[K : F] = m$. Suppose that

$$\beta = \beta_1,\ \beta_2,\ \cdots,\ \beta_m$$

are the roots of g in K; these roots are distinct, because F has characteristic zero (Theorem 45.4).

Consider now an element $\varphi \in \mathrm{Gal}(K|F)$. By Theorem 47.1 it is entirely determined by $\varphi(\beta)$, and we have only m distinct choices β_i to consider. Now by the proof of Exercise 19.20, the function

$$a_0 + a_1\beta + \cdots + a_{m-1}\beta^{m-1} \mapsto a_0 + a_1\beta_i + \cdots + a_{m-1}\beta_i^{m-1}$$

is an isomorphism between $F(\beta)$ and $F(\beta_i)$ that leaves F fixed. Furthermore, $F(\beta) = F(\beta_i) = K$, since K is the unique splitting field for g and a normal extension of F. This means that each choice β_i leads to a distinct element of $\mathrm{Gal}(K|F)$, and so $|\mathrm{Gal}(K|F)| = m = [K : F]$.

Conversely, suppose that $[K : F] = |\mathrm{Gal}(K|F)|$. Since our fields are of characteristic zero, $K = F(\alpha)$, a simple extension of F, where α is a root of an irreducible polynomial $f \in F[x]$ and $\deg(f) = [K : F]$. But any $\varphi \in \mathrm{Gal}(K|F)$ is determined by its value $\varphi(\alpha)$, and $\varphi(\alpha)$ is a root of f. Since there are by assumption $[K : F]$ distinct elements of $\mathrm{Gal}(K|F)$, this means that all of the roots of f already belong to K. That is, K is the splitting field for f over F and so is a normal extension, by Theorem 45.3. □

Example 47.7

As we saw in Example 45.4, the splitting field for $x^3 - 2$ over $\mathbb{Q}$ is $\mathbb{Q}(\sqrt[3]{2}, \zeta)$. We note that $[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3$ and $[\mathbb{Q}(\zeta) : \mathbb{Q}] = 2$. This means that $[\mathbb{Q}(\sqrt[3]{2}, \zeta) : \mathbb{Q}] = 6$, and so by Theorem 47.2 the Galois group $G = \mathrm{Gal}(\mathbb{Q}(\sqrt[3]{2}, \zeta)|\mathbb{Q})$ has six elements. Since each element of the group G permutes the roots of $x^3 - 2$, we can (and should) think of G as a group of permutations. But since $3! = 6$, G consists of all such permutations. We shall thus label the three roots $\sqrt[3]{2}$, $\sqrt[3]{2}\zeta$, $\sqrt[3]{2}\zeta^2$ by the three integers $1, 2, 3$ when thinking of the elements of the Galois group as permutations. It is actually easiest to determine the elements φ of G by choosing both $\varphi(\sqrt[3]{2})$ and $\varphi(\zeta)$, since φ must also permute the two roots ζ, ζ^2 of the polynomial $x^2 + x + 1 = \frac{x^3-1}{x-1}$.

$\varphi(\sqrt[3]{2})$	$\varphi(\zeta)$	perm
$\sqrt[3]{2}$	ζ	ι
$\sqrt[3]{2}$	ζ^2	(23)
$\sqrt[3]{2}\zeta$	ζ	(123)
$\sqrt[3]{2}\zeta$	ζ^2	(12)
$\sqrt[3]{2}\zeta^2$	ζ	(132)
$\sqrt[3]{2}\zeta^2$	ζ^2	(13)

▷ **Quick Exercise.** Check that these six permutations make sense. ◁

Example 47.8

The argument in Example 47.7 can be made more generally. Suppose that $f \in \mathbb{Q}[x]$ is any irreducible cubic polynomial with one real root and two complex roots. Let K be the splitting field for this polynomial over the rational field. Then K contains a degree 3 field extension of $\mathbb{Q}$, and so 3 divides $[K : \mathbb{Q}]$. But complex conjugation is a non-trivial element of $\mathrm{Gal}(K|\mathbb{Q})$. Since this has order two as a group element, 2 divides $|\mathrm{Gal}(K|\mathbb{Q})| = [K : \mathbb{Q}]$. But then the Galois group consists of all 6 permutations of the three roots of f in K and so is isomorphic to S_3.

Example 47.9

In Exercise 45.5 you factored the polynomial $x^4 - 2 \in \mathbb{Q}$ over the complex numbers as follows:

$$\begin{aligned} x^4 - 2 &= \left(x^2 - \sqrt{2}\right)\left(x^2 + \sqrt{2}\right) \\ &= \left(x - \sqrt[4]{2}\right)\left(x + \sqrt[4]{2}\right)\left(x - \sqrt[4]{2}i\right)\left(x + \sqrt[4]{2}i\right). \end{aligned}$$

You then argued that the splitting field of this polynomial is $\mathbb{Q}(\sqrt[4]{2}, i)$. Let's compute the Galois group $G = \mathrm{Gal}(\mathbb{Q}(\sqrt[4]{2}, i)|\mathbb{Q})$ of this splitting field. Now

$$[\mathbb{Q}(\sqrt[4]{2}, i) : \mathbb{Q}] = [\mathbb{Q}(\sqrt[4]{2}, i) : \mathbb{Q}(\sqrt[4]{2})][\mathbb{Q}(\sqrt[4]{2}) : \mathbb{Q}] = 2 \cdot 4 = 8,$$

and so the Galois group has eight elements. We can view the elements of G as permutations of the four roots

$$\sqrt[4]{2}, \quad -\sqrt[4]{2}, \quad \sqrt[4]{2}i, \quad -\sqrt[4]{2}i;$$

we will label these roots by the integers 1 through 4. Notice that in this case we will not obtain the entire permutation group, which has 24 elements. As in the previous example, in practice the elements φ of this Galois group are determined by $\varphi(\sqrt[4]{2})$, which is equal to one of the four roots, and $\varphi(i)$, which is equal to one of the two roots of $x^2 + 1$, namely $\pm i$. We thus obtain the following elements of the Galois group G.

$\varphi(\sqrt[4]{2})$	$\varphi(i)$	perm
$\sqrt[4]{2}$	i	ι
$\sqrt[4]{2}$	$-i$	(34)
$-\sqrt[4]{2}$	i	(12)(34)
$-\sqrt[4]{2}$	$-i$	(12)
$\sqrt[4]{2}i$	i	(1324)
$\sqrt[4]{2}i$	$-i$	(13)(24)
$-\sqrt[4]{2}i$	i	(1423)
$-\sqrt[4]{2}i$	$-i$	(14)(23)

We recognize this group of order 8 as D_4, the group of symmetries of a square.

Example 47.10

We consider next the Galois group for the splitting field of $x^7 - 1$ over $\mathbb{Q}$. Now $x^7 - 1 = (x - 1)\Phi_7(x) = (x - 1)(x^6 + \cdots + x + 1)$, where $\Phi_7(x)$ is the cyclotomic polynomial, which is irreducible by Exercise 5.17. Furthermore, its six distinct roots are exactly $\zeta = e^{\frac{2\pi i}{7}}, \zeta^2, \cdots, \zeta^6$. So in this case the splitting field is exactly $\mathbb{Q}(\zeta)$, and $[\mathbb{Q}(\zeta) : \mathbb{Q}] = 6$. There are thus 6 elements in $\mathrm{Gal}(\mathbb{Q}(\zeta)|\mathbb{Q})$, and they are determined by which seventh root of unity ζ^k that ζ is sent to. We thus obtain the following elements of the Galois group:

$\varphi(\zeta)$	perm
ζ	ι
ζ^2	(124)(365)
ζ^3	(132645)
ζ^4	(142)(356)
ζ^5	(154623)
ζ^6	(16)(25)(34)

We recognize this group as a cyclic group of order 6.

The previous example admits a natural generalization that we will later find quite useful.

Theorem 47.3 *Let p be a prime integer. Suppose that F is a subfield of the complex numbers $\mathbb{C}$. Then the splitting field for the polynomial $\Phi_p(x) = x^{p-1} + x^{p-2} + \cdots + x + 1$ is the field $F(\zeta)$, where $\zeta = e^{\frac{2\pi i}{p}}$ is the primitive pth root of unity. Furthermore,* $\mathrm{Gal}(F(\zeta)|F)$ *is abelian.*

Proof: We know that the splitting field for Φ_p can be considered a subfield of $\mathbb{C}$, and its roots are exactly the powers ζ, ζ^2, ζ^3, $\cdots$ ζ^{p-1} of the primitive pth root of unity ζ. Consequently, given an element $\varphi \in \mathrm{Gal}(F(\zeta)|F)$, it is certainly determined by its value applied to ζ, and that value must be a root of Φ_p, namely, one of the powers ζ^k. So if $\varphi(\zeta) = \zeta^k$, we will denote φ by φ_k. Let's compute $\varphi_k\varphi_j$ and $\varphi_j\varphi_k$. We have

$$\varphi_k\varphi_j(\zeta) = \varphi_k(\zeta^j) = \varphi_k(\zeta)^j = \zeta^{kj}.$$

Since we clearly get the same result in the other order of composition, we must then have that the Galois group is abelian, as claimed. □

Let's summarize what we have learned from our previous examples. Suppose that K is a splitting field over the field F for an irreducible polynomial $f \in F[x]$. Then $[K : F] = |\mathrm{Gal}(K|F)|$ must at least be $n = \deg(f)$, because K contains an isomorphic copy of the field $F[x]/\langle f\rangle$ constructed by Kronecker's theorem, and this field has degree n over F. And since $\mathrm{Gal}(K|F)$ can and should be viewed as a group of permutations of the roots of f, it can never have more than $|S_n| = n!$ elements. In Example 47.10, the Galois group is as small as possible; it has n elements. In Example 47.7, the Galois group is as large as possible; it has $n!$ elements. And in Example 47.9, the Galois group has a number of elements intermediate between n and $n!$.

Example 47.11

The splitting field of the polynomial $x^7 - 2 \in \mathbb{Q}[x]$ is quite evidently $\mathbb{Q}(\sqrt[7]{2}, \zeta)$, where ζ is the seventh root of unity with smallest positive argument, which we considered in Example 47.10.

▷ **Quick Exercise.** What are the roots of this polynomial, in terms of ζ and $\sqrt[7]{2}$? ◁

Let's now compute $G = \mathrm{Gal}(\mathbb{Q}(\sqrt[7]{2}, \zeta)|\mathbb{Q}(\zeta))$. Any element $\varphi \in G$ clearly leaves ζ fixed and is determined by what it does to $\sqrt[7]{2}$. There are thus seven possible choices, and since our extension is normal, all lead to automorphisms. Thus G is a group of order seven, and so is necessarily isomorphic to a cyclic group of order seven.

We can again generalize this example:

Theorem 47.4 *Suppose that p is a positive prime integer, and F a subfield of $\mathbb{C}$. Suppose that $r \in F$, but r has no pth root in F while $\alpha \in \mathbb{C}$ and $\alpha^p = r$. Suppose that $\zeta = e^{\frac{2\pi i}{p}} \in F$. Then $F(\alpha)$ is the splitting field of $x^p - r \in F[x]$ over F, and* $\mathrm{Gal}(F(\alpha)|F)$ *is a cyclic group of order p.*

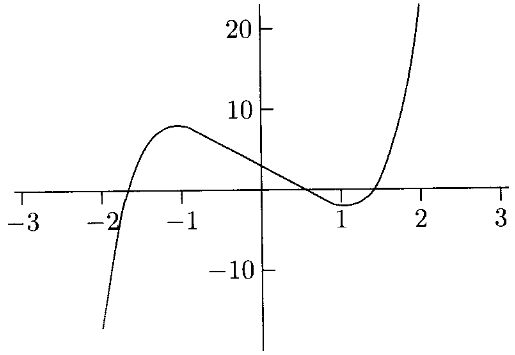
Proof: We may again suppose that the splitting field of $x^p - r$ is a subfield of $\mathbb{C}$. But the roots of $x^p - r$ in $\mathbb{C}$ are evidently the elements α, $\alpha\zeta$, $\alpha\zeta^2$, $\cdots$ $\alpha\zeta^{p-1}$. Because we are assuming that $\zeta \in F$, all these elements belong to the field $F(\alpha)$, and so the latter field is evidently the splitting field, as required.

Given an element $\varphi \in \mathrm{Gal}(F(\alpha)|F)$, it is clearly determined by what it does to α, and obviously $\varphi(\alpha) = \alpha\zeta^k$, for some integer k, with $0 \le k \le p - 1$. Because $F(\alpha)$ is normal over F, each such choice leads to a distinct automorphism. So the Galois group has exactly p elements. The only group with p elements is a cyclic group of that order. □

Example 47.12

Consider now the polynomial $x^5 - 6x + 3 \in \mathbb{Q}[x]$. This is clearly irreducible, by Eisenstein's Criterion 5.7. We will now calculate the Galois group of its splitting field over the rational field. We will do this rather less directly than in our previous examples.

Let's consider the polynomial function $f(x) = x^5 - 6x + 3$ and use some calculus. Now $f'(x) = 5x^4 - 6$, which has exactly two real roots $\pm\sqrt[4]{\frac{6}{5}}$. The negative root corresponds to a local maximum for f and the positive root corresponds to a local maximum. Furthermore, f takes on a positive value at the local maximum and a negative value at the local minimum. Since $\lim_{x\to\pm\infty} f(x) = \pm\infty$, f has exactly three real roots. It must consequently have two complex roots, and these form a complex conjugate pair. The graph of f in the picture below illustrates its properties.



▷ **Quick Exercise.** Verify the details regarding the graph of this function. ◁

Now let K be the splitting field for f over $\mathbb{Q}[x]$. We shall as usual view the Galois group $\mathrm{Gal}(K|\mathbb{Q})$ as a group of permutations of the 5 distinct roots $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$ of f in K. We may as well label our roots so that the first two are the complex roots. Since f is irreducible and of degree 5 and its Kronecker field is a subfield of the splitting field, 5 divides $[K:\mathbb{Q}] = |\mathrm{Gal}(K|\mathbb{Q})|$. Now Theorem 31.5 implies that the group $\mathrm{Gal}(K|\mathbb{Q})$ has an element of order 5, and as a permutation of a set of five elements, the element must be a 5-cycle. As usual, we will think of the elements of our Galois group as permutations on the root subscripts. So we may suppose that this 5-cycle is $\alpha = (1abcd)$, where a, b, c, d are the integers $2, 3, 4, 5$ in some order. But some power of α is then of the form $(12abc)$, and we may as well relabel the three real roots so that the permutation (12345) belongs to the Galois group.

But complex conjugation is clearly a field automorphism of K leaving $\mathbb{Q}$ fixed and so belongs to this Galois group. Complex conjugation leaves the three real roots fixed and interchanges the two conjugate roots, and so as a permutation of the roots it is simply (12).

But now Exercise 34.9 says that (12) and (12345) together generate the entire group S_5. Consequently, we have shown that the Galois group of the splitting field K of the polynomial $x^5 - 6x + 3 \in \mathbb{Q}[x]$ is the full permutation group S_5.

This example will be of considerable importance in Chapter 49, where we will show that the Galois group being S_5 will imply that it is impossible to solve for the roots of the polynomial $x^5 - 6x + 3$ using ordinary field arithmetic, and the extraction of roots!

Chapter Summary

For any finite field extension $K \supseteq F$ we defined the Galois group $\mathrm{Gal}(K|F)$ as the group of all automorphisms of the field K that leave F fixed. When K is the splitting field for some polynomial $f \in F[x]$, then the Galois group can be viewed as a group of permutations of the roots of f in K, whose size is equal to the degree of the field extension K over F. This size is always at least $n = \deg(f)$ and can in principle be as large as $n!$.

Warm-up Exercises

a. Answer the following true or false; if your answer is false, give a counterexample. Assume that $F \subseteq K$ are fields.

   (a) $\mathrm{Gal}(F|F)$ is the trivial group.

   (b) Let $\alpha \in K \backslash F$ and α is algebraic over F. Then $\mathrm{Gal}(F(\alpha)|F)$ is a finite group.

   (c) $\mathrm{Gal}(\mathbb{Q}(\sqrt{5})|\mathbb{Q})$ is a two element group.

   (d) An automorphism preserves addition but need not preserve multiplication.

   (e) The group $\mathrm{Gal}(\mathbb{Q}(\sqrt[5]{2})|\mathbb{Q})$ has at least five elements.

b. Suppose that ζ is the primitive 11th root of unity. What is the Galois group $\mathrm{Gal}(\mathbb{Q}(\sqrt[11]{2}, \zeta)|\mathbb{Q}(\sqrt[11]{2}))$?

c. Give examples of the following, or else argue that such an example does not exist:

(a) A Galois group with exactly 4 elements.

(b) A field F and an element α that is algebraic over F, where $\mathrm{Gal}(F(\alpha)|F)$ has infinitely many elements.

(c) A non-cyclic Galois group.

Exercises

1. Compute the Galois group $\mathrm{Gal}(\mathbb{Q}(\sqrt{2}, i)|\mathbb{Q})$. Compute the Galois group $\mathrm{Gal}(\mathbb{Q}(\sqrt{2}, i)|\mathbb{Q}(\sqrt{2}))$.

2. Consider the splitting field $\mathbb{Z}_2(\alpha)$ of the polynomial $x^3 + x + 1 \in \mathbb{Z}_2[x]$ that we looked at in Example 45.5. Explicitly describe the elements of the Galois group $\mathrm{Gal}(\mathbb{Z}_2(\alpha)|\mathbb{Z}_2)$ as permutations of the three roots $\alpha, \alpha^2, \alpha + \alpha^2$.

3. Compute the Galois group $\mathrm{Gal}(K|\mathbb{Q})$, where K is the splitting field for the polynomial $x^4 - 4x^2 + 2$. You computed this splitting field in Exercise 45.5. Now compute $\mathrm{Gal}(K|\mathbb{Q}(\sqrt{2}))$.

4. Suppose that ζ is the primitive cube root of unity. Let K be the splitting field of $x^9 - 1 \in \mathbb{Q}[x]$. Compute $\mathrm{Gal}(K|\mathbb{Q}(\zeta))$.

5. Consider the polynomial $f = x^3 - 3x - 1 \in \mathbb{Q}[x]$. This is the polynomial we considered in Section 39.2 (and Section 44.2) when we proved that it is impossible to trisect a 60° angle with ruler and compass. In this exercise we will compute the Galois group of the splitting field for f over the rational field.

   (a) Show that f is irreducible over $\mathbb{Q}$.

   (b) In Section 39.2 we used a trigonometric identity to show that $\alpha = 2\cos 20^\circ$ is a root of this polynomial. Use similar arguments to show that $\beta = -2\cos 40^\circ$ and $\gamma = -2\sin 10^\circ$ are the other two roots of f.

   (c) Use elementary trigonometry to show that $\mathbb{Q}(\beta)$ is the splitting field for f over $\mathbb{Q}$. (That is, show that $\alpha, \gamma \in \mathbb{Q}(\beta)$.)

   (d) We can think of $\mathrm{Gal}(\mathbb{Q}(\beta)|\mathbb{Q})$ as a group of permutations of the roots α, β, γ. Argue that in this case this group is a cyclic group of order three.

   (e) In this problem we solved for the roots of f by taking advantage of trigonometry. Instead, use the Cardano-Tartaglia formula to obtain a root of f. Show that the root you obtain is one of the roots given above.

6. Apply the same reasoning as in Example 47.12 to conclude that the Galois group of the splitting field for $x^5 - 14x^2 + 7$ over the rational numbers is the permutation group S_5.

7. Repeat the previous exercise for the polynomial $x^5 - 4x^4 + 2x + 2 \in \mathbb{Q}[x]$.

8. Generalize the reasoning from Example 47.12 to prove the following: Consider the irreducible $f \in \mathbb{Q}[x]$ with $\deg(f) = p$, where p is a prime integer. Suppose that f has exactly two non-real roots. Prove that the Galois group of the splitting field of f is the permutation group S_p.

9. Prove the following modest variation on Theorem 47.2: Let K be a simple algebraic extension of the field F (of any characteristic). Then $|\mathrm{Gal}(K|F)| = [K : F]$ if and only if K is a normal extension of F.

Chapter 48

The Fundamental Theorem of Galois Theory

We saw in the last chapter that the Galois group of a finite extension of a field provides a lot of information about the structure of the extension field. In fact, if the extension is normal, then the degree of the extension is equal to the number of automorphisms belonging to the Galois group. In this chapter we encounter the Fundamental Theorem of Galois Theory, which shows that this connection between field extensions and groups carries even more information than that. In Chapter 49 we will be able to exploit this connection between field theory and group theory to address our goal of better understanding the solution of polynomial equations by field arithmetic and the extraction of roots.

48.1 Subgroups and Subfields

Suppose that we have fields $F \subseteq E \subseteq K$. Then it is easy to see that $\mathrm{Gal}(K|E)$ is a subgroup of $\mathrm{Gal}(K|F)$, because automorphisms of K that fix E clearly also fix $F \subseteq E$.

▷ **Quick Exercise.** Check that $\mathrm{Gal}(K|E)$ is not only a subset of $\mathrm{Gal}(K|F)$ but also a subgroup. ◁

Example 48.1

Referring to Example 47.7, we have that

$$\mathrm{Gal}(\mathbb{Q}(\sqrt[3]{2},\zeta)|\mathbb{Q}) = \{\iota, (23), (123), (12), (132), (13)\},$$

viewed as a group of permutations of the roots of $x^3 - 2$. But $\mathrm{Gal}(\mathbb{Q}(\sqrt[3]{2},\zeta)|\mathbb{Q}(\zeta))$ is quite evidently the subgroup $\{\iota, (123), (132)\}$.

▷ **Quick Exercise.** What subgroup is $\mathrm{Gal}(\mathbb{Q}(\sqrt[3]{2},\zeta)|\mathbb{Q}(\sqrt[3]{2}))$? ◁

Example 48.2

Referring back to Example 47.5, we have that

$$\mathrm{Gal}(\mathbb{Q}(\sqrt{2},\sqrt{3})|\mathbb{Q}) = \{\iota, \phi_1, \phi_2, \phi_3\}.$$

Note that $\mathrm{Gal}(\mathbb{Q}(\sqrt{2},\sqrt{3})|\mathbb{Q}(\sqrt{3}))$ is quite evidently the subgroup $\{\iota, \phi_1\}$.

▷ **Quick Exercise.** What subgroup is $\mathrm{Gal}(\mathbb{Q}(\sqrt{2},\sqrt{3})|\mathbb{Q}(\sqrt{2}))$? How about $\mathrm{Gal}(\mathbb{Q}(\sqrt{2},\sqrt{3})|\mathbb{Q}(\sqrt{6}))$? ◁

So for a field E between the fields $F \subseteq K$ we always obtain a subgroup $\mathrm{Gal}(K|E)$ of the Galois group $\mathrm{Gal}(K|F)$. But we can also proceed in the reverse direction and obtain fields from subgroups. Suppose that H is a subgroup of the Galois group $\mathrm{Gal}(K|F)$. We define

$$\mathrm{Fix}(H) = \{k \in K : \eta(k) = k, \text{ for all } \eta \in H\}.$$

It is obvious that $F \subseteq \mathrm{Fix}(H) \subseteq K$, because all elements of $H \subseteq \mathrm{Gal}(K|F)$ fix elements from F, by definition. But more is true:

Theorem 48.1 *Suppose that $F \subseteq K$ are fields, and H is a subgroup of* $\mathrm{Gal}(K|F)$*. Then* $\mathrm{Fix}(H)$ *is a subfield of K containing F.*

Proof: We leave the straightforward details to Exercise 48.5. □

We thus call $\mathrm{Fix}(H)$ the **fixed field** of the subgroup H.

Example 48.3

Returning again to Example 47.7, we have that $\mathrm{Gal}(\mathbb{Q}(\sqrt[3]{2},\zeta)|\mathbb{Q})$ is (up to isomorphism) the permutation group S_3. This group has the following subgroups:

$$\{\iota\},\ G_1 = \{\iota,(23)\},\ G_2 = \{\iota,(13)\},\ G_3 = \{\iota,(12)\},\ A_3,\ S_3$$

It is easy to compute the fixed fields:

$$\mathrm{Fix}(\{\iota\}) = \mathbb{Q}$$
$$\mathrm{Fix}(G_1) = \mathbb{Q}(\sqrt[3]{2})$$
$$\mathrm{Fix}(G_2) = \mathbb{Q}(\sqrt[3]{2}\zeta)$$
$$\mathrm{Fix}(G_3) = \mathbb{Q}(\sqrt[3]{2}\zeta^2)$$
$$\mathrm{Fix}(A_3) = \mathbb{Q}(\zeta)$$
$$\mathrm{Fix}(S_3) = \mathbb{Q}(\sqrt[3]{2},\zeta)$$

▷ **Quick Exercise.** Check these fixed fields. ◁

Notice that in Example 48.3 we actually have a one-to-one correspondence between all subgroups of S_3 and all fields intermediate between $\mathbb{Q}$ and $\mathbb{Q}(\sqrt[3]{2},\zeta)$. It turns out that this is the case precisely because we have a normal extension. We shall discover that this (and more) is the content of the Fundamental Theorem of Galois Theory 48.3 that we prove below.

48.2 Symmetric Polynomials

Before we can tackle the problem of understanding the correspondence we have begun establishing between intermediate fields and subgroups of the Galois group, we have a technical matter we need to discuss, regarding the coefficients of a polynomial that splits in a given field.

Suppose that $f \in F[x]$ is a monic polynomial with roots $\alpha_1, \alpha_2, \cdots, \alpha_n$ that exist in the splitting field for f over F. We then have that

$$f = (x-\alpha_1)(x-\alpha_2)\cdots(x-\alpha_n).$$

If we multiply this out by using the distributive law, we will obtain $f = \sum_{k=0}^{n}(-1)^{n-k}a_{n-k}x^k$, where the coefficient a_k consists of the sum of all products of exactly k of the α_j with distinct subscripts. (We set $a_0 = 1$). For example, if $n = 4$, then

$$a_0 = 1$$
$$a_1 = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4$$
$$a_2 = \alpha_1\alpha_2 + \alpha_1\alpha_3 + \alpha_1\alpha_4 + \alpha_2\alpha_3 + \alpha_2\alpha_4 + \alpha_3\alpha_4$$
$$a_3 = \alpha_1\alpha_2\alpha_3 + \alpha_1\alpha_2\alpha_4 + \alpha_1\alpha_3\alpha_4 + +\alpha_2\alpha_3\alpha_4$$
$$a_4 = \alpha_1\alpha_2\alpha_3\alpha_4$$

Notice that for convenience of description we have reversed the usual subscript convention: a_k is (up to plus or minus) the coefficient on the x^{n-k} term.

We can justify these formulas with a combinatorial argument. After all multiplications have been distributed out, there will be 2^n terms obtained by making all possible choices, taking one of the two terms in each binomial factor $x - \alpha_i$. To obtain a contributor to the x^k term, we need to choose exactly k x-terms, and $n - k$ α_j terms, each with a distinct subscript j. When we add up each of these terms, we get exactly the a_{n-k} terms described above. Note also that each α_j term selected contributes a factor of -1 as well. It is also possible to construct a careful proof by induction; such a proof is tedious but straightforward, and we will leave it as Exercise 48.7.

We call these coefficients the **symmetric polynomials** in the constants $\alpha_1, \cdots, \alpha_n$. What is important for our purposes to observe is that these symmetric polynomials are symmetric in the α_i's. This means that if we have any field automorphism that permutes the roots of f, it will leave these coefficients fixed. This observation will be important in our proof of the Fundamental Theorem of Galois Theory 48.3 below.

48.3 The Fixed Field and Normal Extensions

Given fields $F \subseteq K$, we have a way to obtain a subgroup $\text{Gal}(K|E)$ of $\text{Gal}(K|F)$, for each intermediate field E. And for each subgroup H of $\text{Gal}(K|F)$, we have a way to obtain an intermediate field $\text{Fix}(H)$. When K is a finite *normal* extension of F, it turns out that these two processes are inverses of one another: group theory will perfectly mirror field theory, and vice versa.

If E is a field with field extension K, then it is clear from the definition of the fixed field and the Galois group that $\text{Fix}(\text{Gal}(K|E)) \supseteq E$.

▷ **Quick Exercise.** Check this. ◁

The following theorem asserts that the reverse inclusion holds only in case that K is normal over E:

Theorem 48.2 *Suppose that $E \subseteq K$ are fields of characteristic zero, and K is a finite extension of E. Then* $\text{Fix}(\text{Gal}(K|E)) = E$ *if and only if K is a normal extension of E.*

Proof: Suppose first that K is normal extension of E; we may as well assume that K is strictly larger than E. Since K is finite over E and of characteristic zero, this means by Theorem 45.5 that $K = E(\alpha)$, for some algebraic $\alpha \in K\backslash E$. We shall assume that the minimal polynomial for α over E is f, and $\deg(f) = n$. Then $1, \alpha, \alpha^2, \cdots, \alpha^{n-1}$ is a basis for K over E. Furthermore, because K is normal over E, K contains all n roots $\alpha = \alpha_1, \cdots, \alpha_n$ of f; these roots are distinct, by Theorem 45.4.

Let $\beta \in K$, and suppose that β is fixed by all elements of $\text{Gal}(K|E)$. We will complete the proof if we can conclude that $\beta \in E$. (As we observed above, the reverse inclusion is always true.)

Now we can express β in terms of our basis as $\beta = \sum_{k=0}^{n-1} a_k\alpha^k$. But β is left fixed by all n elements of $\text{Gal}(K|E)$, and for each i there is such a Galois group element φ_i that takes α to α_i. We thus have

$$\beta = \varphi_i(\beta) = \varphi_i\left(\sum_{k=0}^{n-1} a_k\alpha^k\right) = \sum_{k=0}^{n-1} a_k\alpha_i^k.$$

But now consider the polynomial $g \in K[x]$ defined by

$$g = a_{n-1}x^{n-1} + \cdots + a_1x + (a_0 - \beta).$$

By the previous equation, we see that g has n distinct roots α_i. But this is too many roots for a polynomial with degree no more than $n - 1$. Thus, g must be the identically zero polynomial. This means that $\beta = a_0 \in E$.

For the converse, suppose that $\text{Fix}(\text{Gal}(K|E)) = E$. We again may assume that K is strictly larger than E, and furthermore that it is a simple extension $K = E(\alpha)$ with minimal polynomial f, with $\deg(f) = n$. We will show that K contains all n of the roots of f and thus is the splitting field for f over E, and so is normal, by Theorem 45.3.

Since $\alpha \in K\backslash E$ and $\text{Fix}(\text{Gal}(K|E)) = E$, there must be some $\varphi \in \text{Gal}(K|E)$ so that $\varphi(\alpha) = \alpha_2 \neq \alpha$. Now α_2 is clearly a root of f, and $\alpha_2 \in K$. If $\deg(f) = n = 2$, then f splits in K as required.

If $n > 2$ we must continue by induction. So we will suppose that $\alpha = \alpha_1, \alpha_2, \cdots, \alpha_k$ are all roots of f belonging to K. Let

$$g = (x - \alpha_1)(x - \alpha_2)\cdots(x - \alpha_k).$$

This is clearly a factor of f in $K[x]$. If $k < n$ then g is a non-trivial factor of f. Since f is irreducible in $E[x]$, this means that $g \notin E[x]$. But the coefficients of g are precisely the symmetric polynomials in $\alpha_1, \alpha_2, \cdots, \alpha_k$. So if each element of $\text{Gal}(K|E)$ merely permutes these k roots, the coefficients must belong to the fixed field of the Galois group, which is by assumption E. Thus it must be the case that at least one of the α_i is moved by an element of $\mathrm{Gal}(K|E)$ to a root of f not on our current list. This provides us with another root of f that belongs to K. By induction, we then obtain all the roots of f in K. Thus K is normal extension of E. □

Example 48.4

> We return to Example 47.6. Since $\mathrm{Gal}(\mathbb{Q}(\sqrt[3]{2})|\mathbb{Q})$ is trivial, the fixed field is $\mathbb{Q}(\sqrt[3]{2})$. Because this is strictly larger than the rational field, it is not a normal extension. (Of course, it is easy to see that this extension is not normal, directly from the definition.)

48.4 The Fundamental Theorem

We are now ready to state and prove the Fundamental Theorem of Galois Theory.

Theorem 48.3 The Fundamental Theorem of Galois Theory, Part One *Suppose that K is a finite normal extension of the field F, which is of characteristic zero. There is a one-to-one order reversing correspondence between fields E with $F \subseteq E \subseteq K$ and the subgroups H of* $\mathrm{Gal}(K|F)$. *We can describe this correspondence by two maps that are inverses of one another; namely, we have*

$$E \longmapsto \mathrm{Gal}(K|E)$$

and

$$H \longmapsto \mathrm{Fix}(H).$$

Proof: It is clear from the definition of the Galois group and the fixed field that if we have fields with $E_1 \subseteq E_2$ then $\mathrm{Gal}(K|E_1) \supseteq \mathrm{Gal}(K|E_2)$. Also, if $H_1 \subseteq H_2$ are subgroups of the Galois group, then $\mathrm{Fix}(H_1) \supseteq \mathrm{Fix}(H_2)$. These two maps are thus order-reversing. We need only show that the maps are inverses of one another. To do this, we will compose them in both directions.

Since K is normal over F, it is clearly normal over E (see Exercise 45.a(i)). But then $\mathrm{Fix}(\mathrm{Gal}(K|E)) = E$, as required.

We'd now like to show that for any subgroup H of $\mathrm{Gal}(K|F)$, we have that $H = \mathrm{Gal}(K|\mathrm{Fix}(H))$. Because our fields are of characteristic zero, we have that $K = \mathrm{Fix}(H)(\alpha)$, where α is algebraic over $\mathrm{Fix}(H)$.

Suppose that the subgroup H has h elements, which we may specify as $\iota = \eta_1, \eta_2, \cdots, \eta_h$. We can now consider the polynomial

$$f = (x - \eta_1(\alpha))(x - \eta_2(\alpha)) \cdots (x - \eta_h(\alpha)) \in K[x].$$

The coefficients for the polynomial f are then the symmetric polynomials in the constants

$$\eta_1(\alpha),\ \eta_2(\alpha),\ \cdots,\ \eta_h(\alpha).$$

But if we apply any element of H to the elements of this set, we will just permute them. Consequently, the coefficients for the polynomial f belong to $\mathrm{Fix}(H)$. Thus α is a root of a polynomial with degree h in $\mathrm{Fix}(H)[x]$. This means that $[K : \mathrm{Fix}(H)] = [\mathrm{Fix}(H)(\alpha) : \mathrm{Fix}(H)] \leq h$.

But K is normal over $\mathrm{Fix}(H)$, and so

$$[K : \mathrm{Fix}(H)] = |\mathrm{Gal}(K|\mathrm{Fix}(H))| \geq |H| = h.$$

Putting these inequalities together, we have that

$$h \leq |\mathrm{Gal}(K|\mathrm{Fix}(H))| \leq h,$$

and so $|\mathrm{Gal}(K|\mathrm{Fix}(H))| = h$ and thus $\mathrm{Gal}(K|\mathrm{Fix}(H)) = H$, as required. □

We will at the end of this chapter give numerous examples of the Fundamental Theorem in action, but we will first provide some important additional information about the correspondence between subfields and subgroups.

We first make a few observations about counting. On the group side, we can count by making use of Lagrange's Theorem 31.2, while on the field side our primary counting tool is Theorem 44.2 (a finite extension of a finite extension is finite). So if K is a (characteristic zero) normal field extension of the field F, and E is an intermediate field, we have that $|\mathrm{Gal}(K|F)| = [K : F]$ and $|\mathrm{Gal}(K|E)| = [K : E]$, and

$$[K : F] = [K : E][E : F].$$

But on the group side we have that

$$|\mathrm{Gal}(K|F)| = |\mathrm{Gal}(K|E)|[\mathrm{Gal}(K|F) : \mathrm{Gal}(K|E)].$$

This means that the degree $[E : F]$ of the field extension E over F is precisely equal to the index $[\mathrm{Gal}(K|F) : \mathrm{Gal}(K|E)]$ of the subgroup $\mathrm{Gal}(K|E)$ in the Galois group $\mathrm{Gal}(K|F)$. In the examples in the next section this will be useful for us, because in practice it is often easier to count group elements than it is to calculate the degree of field extensions.

There is another more profound refinement we can add to the one-to-one correspondence we have between intermediate fields and subgroups of the Galois group. It turns out that normal subgroups correspond exactly to normal extensions. This is the reason why such extensions are called normal! This is important enough that we will call this result the second part of the Fundamental Theorem:

Theorem 48.4 Fundamental Theorem of Galois Theory, Part Two *Suppose that K is a finite normal extension of the field F, with characteristic zero. An intermediate field E is a normal extension of F if and only if the Galois group* $\mathrm{Gal}(K|E)$ *is a normal subgroup of* $\mathrm{Gal}(K|F)$. *Furthermore, the Galois group* $\mathrm{Gal}(E|F)$ *is isomorphic to*

$$\mathrm{Gal}(K|F)/\mathrm{Gal}(K|E).$$

Proof: Suppose that E is field intermediate between K and F. Then $\mathrm{Gal}(K|E)$ is normal in $\mathrm{Gal}(K|F)$ if and only if $\varphi^{-1}\psi\varphi \in \mathrm{Gal}(K|E)$, for all $\varphi \in \mathrm{Gal}(K|F)$ and $\psi \in \mathrm{Gal}(K|E)$. But because K is a normal extension of E, this is equivalent to asserting that $\varphi^{-1}\psi\varphi(e) = e$, for all $e \in E$. But this is true exactly if $\psi\varphi(e) = \varphi(e)$, for all $e \in E$. But this means precisely that $\varphi(e) \in \mathrm{Fix}(\mathrm{Gal}(K|E)) = E$. And this says that all roots of the minimal polynomial for e over F actually belong to E. This means exactly that E is a normal extension of F.

To show the group isomorphism, we shall define a group homomorphism Γ from $\mathrm{Gal}(K|F)$ onto $\mathrm{Gal}(E|F)$ with the appropriate kernel. Given $\varphi \in \mathrm{Gal}(K|F)$, we shall define $\Gamma(\varphi) = \varphi|_E$, the restriction of φ to the subfield E. Because E is a normal extension of F, we have that $\varphi(e) \in E$, for all $e \in E$. This means that this map is well defined.

▷ **Quick Exercise.** Why is the map well defined? ◁

It clearly preserves the group operation (functional composition).

Now suppose that $\Gamma(\varphi) = \iota$, the identity automorphism. This means exactly that φ leaves all elements of E fixed. In other words, $\varphi \in \mathrm{Gal}(K|E)$. So the Fundamental Isomorphism Theorem for Groups 33.4 asserts the isomorphism we require. □

48.5 Examples

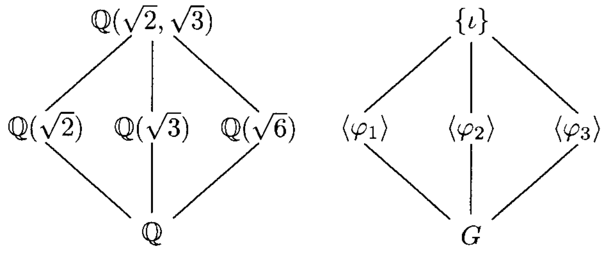
We shall conclude this chapter by looking at a number of examples illustrating the full strength of the Fundamental Theorem of Galois Theory.

Example 48.5

Let's examine the normal field extension $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ of $\mathbb{Q}$ that we considered in Examples 47.5 and 48.2. The Galois group $G = \{\iota, \varphi_1, \varphi_2, \varphi_3\}$ is (isomorphic to) the Klein Four Group, which clearly has three two element subgroups $\langle\varphi_1\rangle$, $\langle\varphi_2\rangle$ and $\langle\varphi_3\rangle$. These subgroups correspond exactly to the three intermediate fields

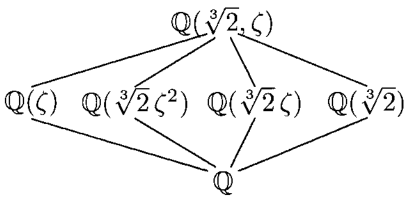
$$\mathbb{Q}(\sqrt{2}),\ \mathbb{Q}(\sqrt{3}),\ \mathbb{Q}(\sqrt{6}),$$

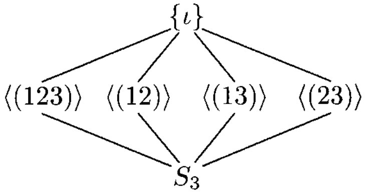
respectively. Since these are the only proper, nontrivial subgroups, the Fundamental Theorem guarantees that these are the only proper intermediate subfields. The index of each of these subgroups in G is two, which corresponds precisely to the fact that each of these extensions is of degree two over $\mathbb{Q}$. The group G is abelian, and so each of these subgroups is normal. This corresponds to the fact that the three quadratic extensions are of course normal. Pictured below is the order-reversing correspondence between subgroups and intermediate fields:



Example 48.6

In Example 48.3 we had already computed the fixed fields corresponding to all of the subgroups of $\mathrm{Gal}(\mathbb{Q}(\sqrt[3]{2}, \zeta)|\mathbb{Q})$ (see also Example 47.7, where we actually computed the Galois group). Only one of the subgroups is normal, namely A_3. Its fixed field is $\mathbb{Q}(\zeta)$ which is a quadratic normal extension; this is the only intermediate field that is normal over $\mathbb{Q}$. Here is the picture of the subgroups and corresponding intermediate fields:


$\mathbb{Q}(\sqrt[3]{2},\zeta)$
$\mathbb{Q}(\zeta)$
$\mathbb{Q}(\sqrt[3]{2}\zeta^2)$
$\mathbb{Q}(\sqrt[3]{2}\zeta)$
$\mathbb{Q}(\sqrt[3]{2})$
$\mathbb{Q}$



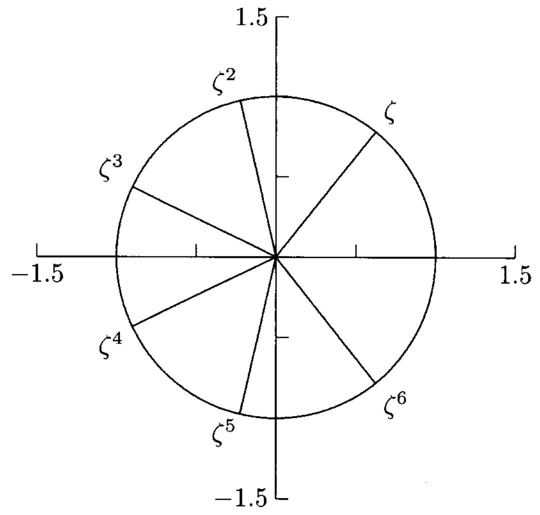
Example 48.7

Let's return to Example 47.9, where we considered the splitting extension of $x^7 - 1$ over $\mathbb{Q}$. We know that the splitting extension is $\mathbb{Q}(\zeta)$, where $\zeta = e^{\frac{2\pi i}{7}}$. The Galois group $G = \mathrm{Gal}(\mathbb{Q}(\zeta)|\mathbb{Q})$ is a cyclic group of order six represented as a permutation group as

$$\{\iota, (132645), (124)(365), (16)(25)(34), (142)(356), (154623)\}.$$

A cyclic group of order six has exactly two proper, nontrivial subgroups, which in this case are $G_1 = \{\iota, (16)(25)(34)\}$ and $G_2 = \{\iota, (124)(365), (142)(356)\}$. We thus have exactly two proper intermediate fields, $\mathrm{Fix}(G_1)$ and $\mathrm{Fix}(G_2)$. The first must be a cubic extension of $\mathbb{Q}$, while the second is a quadratic extension. It is evident that such field elements as $\zeta+\zeta^6$, $\zeta^2+\zeta^5$, $\zeta^3+\zeta^4$ are left fixed by G_1, while elements $\zeta+\zeta^2+\zeta^4$ and $\zeta^3+\zeta^5+\zeta^6$ are left fixed by G_2. These are consequently candidates for field elements that might produce the appropriate intermediate fields. However, we cannot immediately rule out the possibility that these elements might belong to $\mathrm{Fix}(G) = \mathbb{Q}$.

To rigorously determine the fixed fields for these subgroups requires us to inquire more carefully into the arithmetic in $\mathbb{Q}(\zeta)$. We first should remember that ζ (and its powers) are the roots of the irreducible cyclotomic polynomial $1 + x + x^2 + \cdots + x^6$. For more insight into this field, it is helpful to look at the picture of the seventh roots of unity as complex numbers.



It is now quite evident that $\bar{\zeta} = \zeta^6$, $\bar{\zeta^2} = \zeta^5$ and $\bar{\zeta^3} = \zeta^4$.

▷ **Quick Exercise.** Check the above example by examining these complex numbers in trigonometric form. ◁

Consequently, $\zeta+\zeta^6$, $\zeta^2+\zeta^5$, $\zeta^3+\zeta^4$ are real numbers, and we can in fact express them trigonometrically (we will freely use the double angle and sum formulas for the cosine function in the calculations below):

$$\zeta + \zeta^6 = 2\cos\frac{2\pi}{7}$$

$$\zeta^2 + \zeta^5 = 2\cos\frac{4\pi}{7} = 4\cos^2\frac{2\pi}{7} - 2$$

$$\zeta^3 + \zeta^4 = 2\cos\frac{6\pi}{7} = 8\cos^3\frac{2\pi}{7} - 6\cos\frac{2\pi}{7}$$

▷ **Quick Exercise.** Verify these trigonometric calculations. ◁

We can now show that $\zeta + \zeta^6$ actually satisfies an irreducible cubic polynomial in $\mathbb{Q}[x]$. We begin with the cyclotomic polynomial (with terms reordered):

$$0 = 1 + (\zeta+\zeta^6) + (\zeta^2+\zeta^5) + (\zeta^3+\zeta^4) =$$

$$1 + 2\cos\frac{2\pi}{7} + 4\cos^2\frac{2\pi}{7} - 2 + 8\cos^3\frac{2\pi}{7} - 6\cos\frac{2\pi}{7}.$$

This means that $\zeta + \zeta^6$ is a root of $x^3 + x^2 - 2x - 1 \in \mathbb{Q}[x]$, which is clearly irreducible by the Rational Root Theorem 5.6.

Consequently, $\mathbb{Q}\left(\zeta+\zeta^6\right)$ is a cubic extension of $\mathbb{Q}$ and is thus necessarily the fixed field of G_1. In Exercise 48.6 you will check that the other two roots of x^3+x^2-2x-1 are precisely $\zeta^2+\zeta^5$ and $\zeta^3+\zeta^4$.

Now $\zeta+\zeta^2+\zeta^4$ is clearly not a real number, and so

$$\mathbb{Q}\left(\zeta+\zeta^2+\zeta^4\right)$$

must be the other intermediate field. To actually show that it is a quadratic extension, we square it by brute force:

$$\begin{aligned}\left(\zeta+\zeta^2+\zeta^4\right)^2 &= \zeta^8+2\zeta^6+2\zeta^5+\zeta^4+2\zeta^3+\zeta^2\\ &= \left(\zeta+\zeta^2+\zeta^4\right)+2\left(\zeta^3+\zeta^5+\zeta^6\right).\end{aligned}$$

But $\zeta^3+\zeta^5+\zeta^6=-1-\left(\zeta+\zeta^2+\zeta^4\right)$, and so $\zeta+\zeta^2+\zeta^4$ is a root of the polynomial x^2+x+2, and its conjugate $\zeta^3+\zeta^5+\zeta^6$ is the other root. By the quadratic formula we obtain these roots as $-\frac{1}{2}\pm\frac{\sqrt{7}i}{2}$.

We thus have the following diagram of the fields and subgroups:

$$\begin{tikzcd}
& \mathbb{Q}(\zeta) \arrow[dl, dash] \arrow[dr, dash] & \\
\mathbb{Q}(\zeta+\zeta^6) \arrow[dr, dash] & & \mathbb{Q}(\zeta+\zeta^2+\zeta^4) \arrow[dl, dash] \\
& \mathbb{Q} &
\end{tikzcd}
\qquad
\begin{tikzcd}
& \{\iota\} \arrow[dl, dash] \arrow[dr, dash] & \\
G_1 \arrow[dr, dash] & & G_2 \arrow[dl, dash] \\
& G &
\end{tikzcd}$$

Example 48.8

In Example 47.12 we showed that the Galois group of the splitting field K of the irreducible quintic $x^5-6x+3\in\mathbb{Q}[x]$ consists of the full permutation group S_5. As before, we will denote the roots by $\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5$, where α_1 and α_2 are the two complex conjugate roots.

Now S_5 has a large number of subgroups, and we will be unable to analyze all of them! However, it will be useful to look at a couple of examples that are easy to handle.

Consider first the five stabilizer subgroups G_k (for k = 1, 2, 3, 4, 5). The subgroup G_k is precisely the set of elements of S_5 that leave the root α_k fixed (see Exercise 29.6 for information about these subgroups). Each of these groups is isomorphic to S_4, and so has 24 elements. It is evident that the fixed field of G_k is precisely $\mathbb{Q}(\alpha_k)$. These five fields are necessarily distinct, because the subgroups are. Furthermore, they are all isomorphic as field extensions of $\mathbb{Q}$, by Kronecker's Theorem. Kronecker's theorem says these are degree 5 field extensions; equivalently we can see this because $[S_5:G_k]=120/24=5$. These are of course not normal subgroups, and the corresponding fields $\mathbb{Q}(\alpha_k)$ are not normal extensions.

We can generalize to other stabilizer subgroups, as in Exercise 29.7. For example, if $K=\{3,4\}$, then the stabilizer subgroup G_K is isomorphic to S_3 and its fixed field is $\mathbb{Q}(\alpha_3,\alpha_4)$, which is necessarily a field extension of degreee $[S_5:G_K]=120/6=20$.

An obvious element of the Galois group to consider is complex conjugation. It leaves the three real roots $\alpha_3,\alpha_4,\alpha_5$ fixed, and interchanges the conjugate pair α_1 and $\bar{\alpha_1}=\alpha_2$. The 2 element subgroup $\{\iota,(12)\}$ has as its fixed field $\mathbb{R}\cap K$, which is necessarily a field extension of $\mathbb{Q}$ of degree 60.

We have left out an obvious subgroup of S_5 to consider. What about its unique non-trivial normal subgroup A_5? Its fixed field is necessarily a quadratic extension of $\mathbb{Q}$ (which will be a normal extension, of course). It turns out that this intermediate field can be found – our polynomial has an invariant integer d called its *discriminant*, and the appropriate field extension of $\mathbb{Q}$ inside K is $\mathbb{Q}(\sqrt{d})$. We will not pursue this topic here.

Chapter Summary

For a finite normal extension of a field of characteristic zero, we prove the Fundamental Theorem of Galois Theory. This theorem provides a one-to-one order-reversing correspondence between the subgroups of the Galois group and the intermediate fields. Normal subgroups correspond to normal extensions in this correspondence.

Warm-up Exercises

a. Suppose that $f=(x-\alpha_1)\cdots(x-\alpha_5)$. Calculate the coefficient on the x^3 term that results when we multiply f out.

b. Suppose that we have fields $F\subseteq E\subseteq K$. Which of the following groups is a subgroup of which?

$$\mathrm{Gal}(K|E)\quad \mathrm{Gal}(K|F)\quad \mathrm{Gal}(E|F)$$

c. What further relationship holds among the groups in Exercise b, if K is a finite normal extension of F, and E is a nomal extension of F?

d. Suppose that we have the following containments among groups: $H_1 \subseteq H_2 \subseteq \mathrm{Gal}(K|F)$. What containment relationship always holds for the fixed fields $\mathrm{Fix}(H_1)$ and $\mathrm{Fix}(H_2)$?

e. Suppose that K is a finite normal extension of the field F, and $|\mathrm{Gal}(K|F)| = p$, where p is a positive prime integer. What can you say about intermediate fields E, where $F \subset E \subset K$?

Exercises

1. Consider the splitting field K of $x^5 - 3$ over $\mathbb{Q}(\zeta)$, where ζ is the primitive fifth root of unity. Compute the Galois group $\mathrm{Gal}(K|\mathbb{Q}(\zeta))$. Then illustrate the correspondence between subfields and subgroups given by the Fundamental Theorem of Galois Theory, by drawing a diagram illustrating all of the subgroups of the Galois group of this splitting field and the corresponding order-reversed picture of all the fields between $\mathbb{Q}(\zeta)$ and the splitting field.

2. Consider the cyclotomic polynomial $\Phi_5 = x^4 + x^3 + x^2 + x + 1 \in \mathbb{Q}[x]$. Then we know that the splitting field for Φ_5 is just $\mathbb{Q}(\zeta)$, where ζ is the primitive fifth root of unity, and the Galois group of this splitting field is a cyclic group of order 4 (see Theorem 47.3 and Example 47.10). Draw a diagram illustrating all of the subgroups of the Galois group of this splitting field and the corresponding order-reversed picture of all the fields between $\mathbb{Q}$ and the splitting field.

3. Consider again the splitting field of $x^4 - 2 \in \mathbb{Q}[x]$, as in Exercise 45.6 and Example 47.9. As in the previous exercise, draw a diagram illustrating all of the subgroups of the Galois group of this splitting field and the corresponding order-reversed picture of all the fields between $\mathbb{Q}$ and the splitting field. You should explicitly verify that each intermediate field is the fixed field of the appropriate subgroup.

4. Which of the subgroups of the Galois group in the previous exercise are normal in the Galois group of the splitting field? What property do the corresponding fields have?

5. Prove Theorem 48.1.

6. In this exercise you check a computation from Example 48.7. In that example we showed that $\zeta + \zeta^6$ is a root of the polynomial $x^3 + x^2 - 2x - 1$, where ζ is the primitive seventh root of unity. Show that $\zeta^2 + \zeta^5$ and $\zeta^3 + \zeta^4$ are the other two roots.

7. In this exercise you provide an inductive proof for the formula for the symmetric polynomials, as discussed in Section 48.2. Suppose that $f = (x - \alpha_1)(x - \alpha_2) \cdots (x - \alpha_n) \in F[x]$, where F is a field. Let $(-1)^{n-k} a_{n-k}$ be the coefficient on x^k when we multiply f out using the distributive law. Use induction on n to prove that
$$a_{n-k} = \Sigma\{\alpha_{j_1}\alpha_{j_2} \cdots \alpha_{j_k}\},$$
where the sum is over all possible choices of k distinct α_i.

8. Suppose that $f \in F[x]$ is irreducible polynomial of degree n over the field F, where F is a field of characteristic zero. Let K be the splitting field of f over F. Then we can think of $G = \mathrm{Gal}(K|F)$ as a subgroup of the permutation group S_n. Prove that G is a transitive subgroup of this permutation group (see Exercise 34.13 for a definition).

Chapter 49

Solving Polynomials by Radicals

We are now ready to focus our attention on the problem of whether it is possible to solve all polynomial equations over a subfield of the complex numbers, by ordinary field arithmetic, together with the extraction of roots. We are able to do this in the quadratic case (using the quadratic formula Exercise 9.1), in the cubic case (using the Cardano-Tartaglia approach, Exercise 9.12), and in the quartic case (using the Ferrari approach, Exercise 9.20).

In this chapter we will recast this problem in terms of field extensions, just as we did for the notion of constructible numbers, in Chapter 38. In that context we simplified matters considerably, by focusing our attention on the sequence of ever larger fields necessary to obtain the constructible number. In that case the larger fields were built as quadratic extensions. In our present case, we will need to build larger and larger fields, but allow extensions by higher power roots instead.

49.1 Field Extensions by Radicals

Let's consider a single step in the process of building larger fields by extraction of roots. Given a field F and an integer $n \geq 2$, a **radical extension** of F is a simple algebraic field extension of the form $F(\beta)$, where $\beta^n \in F$. Thus every element of the radical extension can be expressed in terms of field arithmetic on elements of F and β. But we can view β as the result of an extraction of an nth root of $\beta^n \in F$. That is, every element of $F(\beta)$ can be obtained by a formula involving only ordinary field arithmetic, together with a root extraction.

We then generalize this inductively, supposing that we have a finite sequence of radical extensions

$$F = F_0 \subset F_1 = F_0(\beta_1) \subset F_2 = F_1(\beta_2) \subset \cdots \subset F_N = F_{N-1}(\beta_N),$$

where at each stage $\beta_i^{n_i} \in F_{i-1}$. We call the field F_N an **extension of F by radicals**; for every element in F_N there is a (perhaps very complicated!) formula, involving only field operations and extractions of roots, applied iteratively to the elements of the field F. We'll call the finite sequence of fields a **root tower over** F.

Now, if $f \in F[x]$ is a polynomial, and its splitting field K is contained in an extension of F by radicals, we say that f is **solvable by radicals.**

This description of solvability by radicals can and should be compared to our description of constructible numbers in Theorem 38.3.

Let's look at some examples.

Example 49.1

Any quadratic polynomial $f = ax^2 + bx + c \in \mathbb{Q}[x]$ is solvable by radicals over $\mathbb{Q}$, because the splitting field for f is contained in (and is in this case equal to) the field extension $\mathbb{Q}\left(\sqrt{b^2 - 4ac}\right)$. Here the root tower consists of just two fields: $\mathbb{Q} \subseteq \mathbb{Q}\left(\sqrt{b^2 - 4ac}\right)$. Obviously the quadratic formula expression is our explicit solution by radicals.

Example 49.2

Consider the polynomial $f = x^6 - 2x^3 - 1 \in \mathbb{Q}[x]$. Since this is a quadratic equation in x^3, it is evident by the quadratic formula that it factors as

$$f = (x^3 - (1 + \sqrt{2}))(x^3 - (1 - \sqrt{2})).$$

But the roots in $\mathbb{C}$ of these two factors are in turn just cube roots. If as usual we let $\zeta = -\frac{1}{2} + \frac{\sqrt{3}}{2}i$ be the primitive cube root of unity, we then have f completely factored as

$$f = \left(x - \sqrt[3]{1+\sqrt{2}}\right)\left(x - \sqrt[3]{1+\sqrt{2}}\zeta\right)\left(x - \sqrt[3]{1+\sqrt{2}}\zeta^2\right)$$
$$\left(x - \sqrt[3]{1-\sqrt{2}}\right)\left(x - \sqrt[3]{1-\sqrt{2}}\zeta\right)\left(x - \sqrt[3]{1-\sqrt{2}}\zeta^2\right).$$

So each of these six roots can be expressed in terms of a formula involving only field operations, together with both cube and square roots. But we can also look at this in terms of the corresponding root tower:

$$\mathbb{Q} \subset \mathbb{Q}(\sqrt{2}) \subset \mathbb{Q}(\sqrt{2})\left(\sqrt[3]{1+\sqrt{2}}\right) \subset$$
$$\mathbb{Q}(\sqrt{2})\left(\sqrt[3]{1+\sqrt{2}}\right)(\zeta) = F_3.$$

It may not be immediately obvious that all six roots belong to the field F_3, but we can conclude this if we note that

$$\left(\sqrt[3]{1+\sqrt{2}}\right)\left(\sqrt[3]{1-\sqrt{2}}\right) = \sqrt[3]{1-2} = -1,$$

and so

$$\sqrt[3]{1-\sqrt{2}} = \frac{-1}{\sqrt[3]{1+\sqrt{2}}} \in \mathbb{Q}(\sqrt{2})\left(\sqrt[3]{1+\sqrt{2}}\right).$$

In this case F_3 is again equal to the splitting field for f over $\mathbb{Q}$, and $[F_3 : \mathbb{Q}] = 12$.

▷ **Quick Exercise.** Why are the assertions in the last statement true? Note that by Theorem 47.2 the Galois group $\mathrm{Gal}(F_3|\mathbb{Q})$ is a group of 12 elements — you will determine which group in Exercise 49.1. ◁

Example 49.3

Let's now consider the irreducible cubic polynomial $f = x^3 + 3x + 1 \in \mathbb{Q}[x]$. We can use the Cardano-Tartaglia formula to obtain a real root for this polynomial. In Exercise 9.16 you do this to obtain

$$\alpha_1 = \sqrt[3]{\frac{-1+\sqrt{5}}{2}} - \sqrt[3]{\frac{1+\sqrt{5}}{2}}.$$

Furthermore in that exercise you verify that if ζ is the primitive cube root of unity, then the other two roots for f in $\mathbb{C}$ are a conjugate pair of complex numbers:

$$\alpha_2 = \sqrt[3]{\frac{-1+\sqrt{5}}{2}}\zeta - \sqrt[3]{\frac{1+\sqrt{5}}{2}}\zeta^2$$

and

$$\alpha_3 = \sqrt[3]{\frac{-1+\sqrt{5}}{2}}\zeta^2 - \sqrt[3]{\frac{1+\sqrt{5}}{2}}\zeta.$$

It is now quite evident that the splitting field K for f can be reached by the following root tower:

$$\mathbb{Q} \subset \mathbb{Q}(\zeta) \subset \mathbb{Q}(\zeta, \sqrt{5}) \subset \mathbb{Q}\left(\zeta, \sqrt{5}, \sqrt[3]{\frac{-1+\sqrt{5}}{2}}\right) = F_3.$$

That is, we have that K is a subfield of F_3.

▷ **Quick Exercise.** Why does $\sqrt[3]{\frac{1+\sqrt{5}}{2}} \in F_3$? ◁

In Exercise 49.2 you will check that each of these field extensions is in fact proper, and so $[F_3 : \mathbb{Q}] = 12$. But from Example 47.8 we know that $[K : \mathbb{Q}] = |\mathrm{Gal}(K|\mathbb{Q})| = 6$. Consequently, in this example our root tower reaches a field extension strictly larger than the splitting field. We will have to account for this possibility in the general theory we develop later in this chapter.

49.2 Refining the Root Tower

Suppose again that we have a field F, a polynomial $f \in F[x]$, and a root tower over F that enables us to solve f by radicals. Our eventual goal is to show that the existence of such a root tower places restrictions on the sort of group we can get as the Galois group of the splitting field for f. To accomplish this, we will first modify our root tower in a couple of ways. We'd like each step along the tower to be small enough that it can be well understood, and we'd also like to have enough normal extensions along the way in the root tower, so that we can apply the Fundamental Theorem of Galois Theory 48.3 and 48.4 to them. We will accomplish these goals by inserting a number of extra radical extensions into the root tower; we will eventually be able to show that this new 'refined' root tower has the properties we desire.

Consider now a single step in our root tower. We suppose that we have the field extension $E \subseteq E(\beta)$, where $\beta^n \in E$, and E is one of the fields reached at some stage of our root tower for F. We now claim that we can assume that n is a prime integer. For if $n = p_1 p_2 \cdots p_m$ is the prime factorization of n into (not necessarily distinct) prime factors, we can replace the single radical extension from E to $E(\beta)$ by a sequence of extensions as follows:

$$E \subseteq E(\beta^{p_2 p_3 \cdots p_m}) \subseteq E(\beta^{p_3 \cdots p_m}) \subseteq \cdots \subseteq E(\beta).$$

We obviously arrive at the same spot in our root tower, at the expense of taking finitely many extra steps in the construction. When we pass to the group side later, it will be more convenient for us to assume we need only extract roots of prime order.

At this point we will impose a serious restriction on our arguments that follow. We will assume not only that our field F is of characteristic zero, but that it is actually a subfield of the complex numbers $\mathbb{C}$. The examples of fields of characteristic zero that we have explored in this and earlier chapters have all satisfied this criterion, although there certainly do exist fields of characteristic zero that do not. Our desire in this chapter is to recover Abel's result that there exist fifth degree polynomials over $\mathbb{Q}$ that are not solvable by radicals, and so this restrictive assumption does not effect that goal. The advantage of this restriction is that we then have a very concrete understanding of the splitting field of the polynomial $x^p - 1 \in \mathbb{Q}[x]$ obtained as $\mathbb{Q}(\zeta)$, where ζ is the primitive pth root of unity (see Theorem 47.3).

With this restriction in place, we are now ready to return to our project of refining our root tower. Let's suppose that the root tower we now have looks like this:

$$F = F_0 \subset F_1 = F(\beta_1) \subset F_2 = F(\beta_2) \subset \cdots \subset F_m = F_{m-1}(\beta_m),$$

where for each i there is a prime integer p_i with $\beta_i^{p_i} \in F_{i-1}$. Now consider the set of all these primes p_i appearing as the degrees of the radical extensions in our root tower (of course, some of these primes may occur more than once in this list). We shall next refine our root tower, by first adjoining each of the p_ith primitive roots of unity ζ_i, one at a time. The first terms of our root tower will then look like this:

$$F \subseteq E_1 = F(\zeta_1) \subseteq E_2 = E_1(\zeta_2) \subseteq \cdots \subseteq E_m = F(\zeta_1, \zeta_2, \cdots, \zeta_m).$$

We note that by Theorem 47.3 it is evident that each of these fields E_i is normal over F, because evidently E_i is the splitting field over F for the polynomial

$$(x^{p_1} - 1)(x^{p_2} - 1) \cdots (x^{p_i} - 1).$$

Note that in case we have repetitions in our list of primes, some of these extensions may be trivial; this is harmless for what follows.

We record our progress so far in refining our root tower:

Lemma 49.1 *Suppose that F is a subfield of the complex numbers, and $f \in F[x]$ is solvable by radicals over F. Then there exists a root tower of the following form:*

$$F \subseteq E_1 = F(\zeta_1) \subseteq E_2 = E_1(\zeta_2) \subseteq \cdots \subseteq E_m = K_0 = F(\zeta_1, \zeta_2, \cdots, \zeta_m)$$

$$\subseteq F_1 = F_0(\beta_1) \subseteq F_2 = F_1(\beta_2) \subseteq \cdots \subseteq F_m = F_{m-1}(\beta_m),$$

where p_i are prime integers, ζ_i is the primitive p_ith root of unity, and $\beta_i^{p_i} \in F_{i-1}$.

We will now modify our root tower further. Our goal now is to include enough elements in addition to the elements β_i, so that we will obtain *normal* extensions of F. We can do this inductively, one step at a time. Our starting point for the induction is the normal field extension $F \subseteq K_0 = F(\zeta_1, \zeta_2, \cdots, \zeta_m)$.

So suppose that we have already constructed a new sequence of fields

$$F \subseteq K_0 \subseteq K_1 \subseteq K_2 \cdots \subseteq K_i,$$

so that for every $1 \leq j \leq i$, we have that $\beta_j \in K_j$, and each K_j is an extension of K_{j-1} by radicals, and is a normal extension of F; in fact, we will assume that K_j is the splitting field of $g_j \in F[x]$. Note that because $F_i = F(\beta_1, \beta_2, \cdots, \beta_i)$, it is evident that $F_j \subseteq K_j$, for all $1 \leq j \leq i$.

We shall now provide the inductive argument to construct K_{i+1}. We know that $\beta_{i+1}^{p_{i+1}} \in F_i \subseteq K_i$; for notational convenience, let's call this element e, and set $p = p_{i+1}$. We can enumerate the elements of the Galois group $\mathrm{Gal}(K_i|F)$ as $\{\iota = \varphi_1, \varphi_2, \cdots, \varphi_k\}$. Consider the polynomial

$$h = (x^p - \varphi_1(e))(x^p - \varphi_2(e)) \cdots (x^p - \varphi_k(e)) \in K_i[x].$$

We will now build the splitting field of h over K_i by radical extensions, one factor at a time. Notice that the coefficients of the polynomial h are symmetric polynomials in the coefficients

$$\varphi_1(e), \cdots, \varphi_h(e),$$

which are permuted by the elements of $\mathrm{Gal}(K_i|F)$. Because by the induction hypothesis K_i is normal over F, we know that these coefficients belong to F, the fixed field of $\mathrm{Gal}(K_i|F)$. That is, $h \in F[x]$.

Now we know that the primitive pth root of unity ζ is an element of the field K_i. Consequently, in order to assure that the factor $x^p - \varphi_i(e)$ splits, we need only adjoin a single root ρ_i to the previous field (Theorem 47.4); here ρ_i is a root of the polynomial $x^p - \varphi_i(e)$. We thus have

$$K_i \subset K_i(\rho_1) \subset K_i(\rho_1, \rho_2) \subset \cdots \subset K_i(\rho_1, \cdots, \rho_k).$$

Note that at each stage in this sequence we have a splitting field for a polynomial $(x^p - \varphi_1(e))(x^p - \varphi_2(e)) \cdots (x^p - \varphi_j(e)) \in K_i[x]$, and so each of these fields is normal over K_i. (They may not be normal over F, because these intermediate polynomials need not belong to $F[x]$.) We shall now let $K_{i+1} = K_i(\rho_1, \cdots, \rho_h)$; evidently K_{i+1} is the splitting field over F of the polynomial $g_{i+1} = g_i h \in F[x]$.

Let's summarize the situation we now have.

Lemma 49.2 *Suppose that $F \subset \mathbb{C}$ is a field, and $f \in F[x]$ is a polynomial that can be solved by radicals over F. Then there exists a sequence of field extensions*

$$F \subseteq E_1 = F(\zeta_1) \subseteq E_2 = E_1(\zeta_2) \subseteq \cdots \subseteq E_m = K_0 = F(\zeta_1, \zeta_2, \cdots, \zeta_m)$$

$$\subseteq K_1 \subseteq K_2 \subseteq \cdots \subseteq K_m,$$

so that the following conditions hold:

1. *For each i, ζ_i is the p_ith primitive root for unity, for the prime integer p_i.*

2. *Each K_i is normal over F.*

3. *Each K_{i+1} is obtained from K_i by finitely many radical extensions of degree p_i; these extensions are all normal over K_i.*

4. *K_m contains the splitting field of f over F.*

We have now put all the technical details in place and are now ready to apply Galois theory to this lemma, which we shall do in the next section.

49.3 Solvable Galois Groups

In the last section we showed that whenever a polynomial $f \in F[x] \subseteq \mathbb{C}[x]$ is solvable by radicals, we can construct a root tower of a particular type, as specified in Lemma 49.2. This gives us the information we need to prove the following:

Theorem 49.3 *Suppose that $f \in F[x] \subseteq \mathbb{C}[x]$ is solvable by radicals. Then the Galois group of the splitting field for f is solvable.*

Proof: We will now apply the Fundamental Theorem of Galois Theory to the field extensions guaranteed by Lemma 49.2. Let

$$G = G_0 = \mathrm{Gal}(K_m|F),\ G_1 = \mathrm{Gal}(K_m|E_1),\ \cdots,\ G_m = \mathrm{Gal}(K_m|E_m)$$

$$G_{m+1} = \mathrm{Gal}(K_m|K_1),\ \cdots,\ G_{2m} = \mathrm{Gal}(K_m|K_m).$$

It is evident that

$$G_{2m} = \{\iota\} \subseteq G_{2m-1} \subseteq \cdots \subseteq G_m \subseteq \cdots \subseteq G_1 \subseteq G$$

and each of these subgroups is normal in G, because each of the fields E_i and K_i are normal extensions of F.

We would like to conclude that G is a solvable group; we will use Theorem 36.4. For $0 < i \leq m$, we have that

$$G_{i-1}/G_i = \mathrm{Gal}(K_m|E_{i-1})/\mathrm{Gal}(K_m|E_i) \approx \mathrm{Gal}(E_{i-1}|E_i).$$

The isomorphism follows from the Fundamental Theorem of Galois Theory. But by Theorem 47.3 we know that this is an abelian group, and hence solvable, of course.

For $0 < i \leq m$, we have that

$$G_{m+i-1}/G_{m+i} = \mathrm{Gal}(K_m|K_{i-1})/\mathrm{Gal}(K_m|K_i) \approx \mathrm{Gal}(K_{i-1}|K_i).$$

We would like to argue that this last group is a solvable group.

But note that we obtain K_i from K_{i-1} by finitely many radical extensions of degree $p = p_i$, a prime integer. That is, we have that

$$K_{i-1} \subseteq L_1 \subseteq L_2 \cdots \subseteq L_h \subseteq K_i,$$

where each of the fields L_j is obtained from its predecessor by adjoining a pth root of an element from the previous field. Furthermore, we know by construction that the primitive pth root of unity $\zeta \in K_{i-1}$. Consequently, each of these extensions is normal over K_{i-1}, because it is the splitting field over K_{i-1} for a polynomial of the form $x^p - e$ (see Lemma 49.2 above).

We thus have the subnormal series

$$\{1\} \subseteq \mathrm{Gal}(K_i|L_h) \subseteq \mathrm{Gal}(K_i|L_{h-1}) \subseteq \cdots \subseteq \mathrm{Gal}(K_i|L_1) \subseteq \mathrm{Gal}(K_i|K_{i-1}).$$

When we compute the quotient group of two of these groups we have

$$\mathrm{Gal}(K_i|L_{j-1})/\mathrm{Gal}(K_i|L_j) \approx \mathrm{Gal}(L_{j-1}|L_j).$$

But this last group is cyclic by Theorem 47.4. This means that our subnormal series actually makes $\mathrm{Gal}(K_i|K_{i-1})$ a solvable group.

Putting this all together, we have now concluded by Theorem 36.4 that $G = \mathrm{Gal}(K_m|F)$ is also a solvable group.

This is almost the conclusion that we want. However, K_m is in practice probably much larger than the splitting field for f over F. But if we let K be that splitting field, then K is a normal extension of F, and so another application of the Fundamental Theorem of Galois Theory gives us that

$$\mathrm{Gal}(K|F) \approx \mathrm{Gal}(K_m|F)/\mathrm{Gal}(K_m|K).$$

But this means that $\mathrm{Gal}(K|F)$ is a homomorphic image of the solvable group $\mathrm{Gal}(K_m|F)$ and so is itself solvable, by Theorem 36.1. □

It is actually the case that the converse of Theorem 49.4 is true — that is, if the Galois group for the splitting field of a polynomial $f \in F[x] \subseteq \mathbb{Q}[x]$ is solvable, then f is actually solvable by radicals over F. To prove this, we would have to start with the subnormal series with abelian quotients for the Galois group, and build a sequence of radical field extensions that eventually contain the splitting field. We will not actually carry out this project here.

The next example is what we've been looking for! It is a polynomial of degree five that is not solvable by radicals.

Example 49.4

Consider again the polynomial $f = x^5 - 6x + 3 \in \mathbb{Q}[x]$. In Example 47.12 we showed that the Galois group of the splitting field for f over $\mathbb{Q}$ is S_5. In Example 36.4 we asserted that S_5 is not a solvable group. Theorem 49.3 now means that f is *not solvable by radicals!*

Example 49.5

In Example 49.2 we considered the polynomial $f = x^6 - 2x^3 - 1 \in \mathbb{Q}[x]$ and showed explicitly that it is solvable by radicals. In Exercise 49.1 you computed the Galois group for the splitting field of this polynomial and showed that it is a solvable group.

We have thus solved (in the negative) a problem that bedeviled European mathematicians for several centuries. It is in principle impossible to extend the progressively more complicated algebraic formulas we have for second, third and fourth degree equations to the fifth degree or higher. The mathematics involved is more complicated, but the situation is the same as for the classical constructibility problems: field theory (in this case with an important assist from the theory of groups) has solved an important mathematical problem that at first blush does not seem to require the abstract approach.

The abstract algebraic approach was successful in settling the construction problems of classical antiquity and in solving the solution by radicals problems of the Italian renaissance; this is only the beginning of the story. Powerful algebraic techniques have played an important role in bringing profound insights into many difficult and important problems, over the course of the nineteenth, twentieth and twenty-first centuries. You are invited to learn more about the successes and beauty of modern mathematics.

Chapter Summary

In this chapter we prove that if a polynomial over a subfield of the complex numbers can be solved by radicals, then the Galois group of its splitting field over the base field is necessarily a solvable group. We can then easily exhibit a fifth degree polynomial over the rational numbers that *cannot* be solved by radicals.

Warm-up Exercises

a. Give a root tower whose last field contains the splitting field for $x^4 - 2$ over the rational numbers. (This is Example 47.9.)

b. Give a root tower whose last field contains the splitting field for $x^3 - 2$ over the rational numbers. (This is Example 47.7.)

c. Give an example of a root tower whose last field is *not* normal over the base field.

d. Suppose that ζ is the primitive pth root of unity, F is a subfield of the complex numbers, $\zeta \in F$, but F contains no pth root of $a \in F$. Let α be a pth root of a in $\mathbb{C}$. Give as much information as you can about the field extension $F(\alpha)$ over F.

e. Is the sequence of fields given by the Constructible Number Theorem 38.3 a root tower?

f. Give an example of an irreducible polynomial $f \in \mathbb{Q}[x]$ that can be solved by radicals, none of whose roots are constructible numbers. Give such an example where all the roots are real numbers.

g. Suppose that K is the splitting field for $f \in \mathbb{Q}[x]$, and $\mathrm{Gal}(K|\mathbb{Q}) \approx S_7$. What can you say about this situation?

Exercises

1. Consider the polynomial $x^6 - 2x^3 - 1 \in \mathbb{Q}[x]$ from Example 49.2 and 49.5. Compute the Galois group for the splitting field of this polynomial over $\mathbb{Q}$, and show explicitly that this is a solvable group. Draw the containment diagrams for the subgroups of the Galois group, and the corresponding diagram of fields intermediate between the rational field and the splitting field.

2. Consider the root tower obtained in Example 49.3 to obtain a field containing the splitting field K for $x^3 + 3x + 1 \in \mathbb{Q}[x]$. Show that each each field extension in the root tower is proper, and so the field F_3 is a proper extension of K.

3. In this problem and its successor we will describe a method that helps determine the Galois group of the splitting field for a quartic polynomial. Let F be a subfield of the complex numbers, and suppose that $f \in F[x]$ is an irreducible polynomial of degree four. We may as well assume that f is monic, and so

$$f = (x - \alpha_1)(x - \alpha_2)(x - \alpha_3)(x - \alpha_4),$$

where the α_i are distinct elements of $\mathbb{C}$. Let $K = F(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ be the splitting field for f. We will consider $G = \mathrm{Gal}(K|F)$ as a subgroup of the group S_4 of permutations of these roots, in this order. By Section 48.2 we know that $f = x^4 - a_1x^3 + a_2x^2 - a_3x^1 + a_4$, where the a_i are the symmetric polynomials in the roots α_i.

(a) Let

$$\beta_1 = \alpha_1\alpha_2 + \alpha_3\alpha_4, \quad \beta_2 = \alpha_1\alpha_3 + \alpha_2\alpha_4,$$
$$\beta_3 = \alpha_1\alpha_4 + \alpha_2\alpha_3.$$

Form the polynomial

$$g = (x - \beta_1)(x - \beta_2)(x - \beta_3) = x^3 - b_1x^2 + b_2x - b_3.$$

By making use of what we know about symmetric polynomials, prove that

$$b_1 = a_2, \quad b_2 = a_1a_3 - 4a_4, \quad b_3 = a_1^2a_4 + a_3^2 - 4a_2a_4.$$

Thus, $g \in F[x]$. *Note:* The polynomial g is called the **resolvent polynomial** for f.

(b) Let $E = F(\beta_1, \beta_2, \beta_3)$ be the splitting field for g over F. Let $H = G \cap \{\iota, (12)(34), (13)(24), (14)(23)\}$. Prove that $E = \mathrm{Fix}(H)$.

(c) Why is H normal in G?

(d) Prove that $\mathrm{Gal}(E|F)$ is isomorphic to G/H.

4. In this exercise we use the same notation and terminology as in the previous exercise. By Exercise 48.8 the Galois group G is necessarily a transitive subgroup of S_4. In Exercise 34.13 you compiled a list of all the transitive subgroups of S_4. We will in what follows think of $H = \mathrm{Gal}(E|F)$ as a subgroup of the group of permutations S_3 of the roots of g. We now look at all possible cases for G:

(a) Suppose that $G = S_4$. Then show that $H = S_3$.

(b) Suppose that $G = A_4$. Then show that $H = A_3$.

(c) Suppose that $G = \{\iota, (12)(34), (13)(24), (14)(23)\}$. Then show that H is the trivial group.

(d) Suppose that G is a cyclic group of order 4. Then show that H is a cyclic group of order 2.

(e) Suppose that

$$G = \{\iota, (12)(34), (13)(24), (14)(23), (12), (34), (1423), (1324)\},$$

the subgroup of S_4 of order 8 (isomorphic to D_4). Then show that H is a cyclic group of order 2.

Note: Since H is the Galois group of the splitting field of a cubic polynomial, it is presumably easier to compute than is G. But the results above say that in most cases knowing H actually determines G. We will illustrate this principle in the exercises that follow. The ambiguous case that remains (when H is a cyclic group of order 2) can actually also be resolved by a refinement of the argument given here. For information about that case you can consult Nathan Jacobson's *Basic Algebra I* (W. H. Freeman and Company, 1974), from which these exercises have been adapted.

5. From Example 47.9 we know that the Galois group of the splitting field for $x^4 - 2$ is isomorphic to D_4. Thus if we form the resolvent polynomial g, we know from Exercise 4 that its Galois group is a cyclic group of order two. Compute g and the Galois group explicitly, to check this computation.

6. Consider the polynomial $f = x^4 + 2x - 2 \in \mathbb{Q}[x]$.

(a) Why is f irreducible over $\mathbb{Q}[x]$?

(b) Show that the resolvent polynomial g obtained as in Exercise 3 is $g = x^3 + 8x + 4$.

(c) Why is g irreducible over $\mathbb{Q}[x]$?

(d) Use Example 47.8 to argue that the Galois group for the splitting field for g over the rational field is S_3.

(e) What is the Galois group of the splitting field for f over $\mathbb{Q}$?

7. Consider the polynomial $f = x^4 + 2x + 2 \in \mathbb{Q}[x]$.

(a) Why is f irreducible over $\mathbb{Q}[x]$?

(b) Show that the resolvent polynomial g obtained as in Exercise 3 is $g = x^3 - 8x + 4$.

(c) Why is g irreducible over $\mathbb{Q}[x]$?

(d) Use calculus to argue that g has three real roots.

(e) Argue that the Galois group of g over the rational field must contain an element of order 3. What does this mean about the Galois group of the splitting field for f?

(f) Conclude from part e that at least one of the real roots of f is not constructible, even though it is of degree 4 over $\mathbb{Q}$. This is the promised example that shows that the converse of Corollary 44.5 is false.

Section IX in a Nutshell

This section applies field theory and group theory to prove that there is no general method to solve fifth degree equations using only field operations and extractions of roots.

We start by defining a *splitting field* for a polynomial f over a field F, which is a minimal field extension of F where f factors into linear polynomials. There is always a splitting field and a splitting field is a finite extension of F (Theorem 45.1). Furthermore, any two splitting fields for a given $f \in F[x]$ are isomorphic (Theorem 45.2). Splitting fields have the surprising property that if K is a splitting field over F for f, and $g \in F[x]$ is irreducible over F with a root in K, then g also factors into linear polynomials in K (Theorem 45.3).

An extension field K of F is a *normal extension* of F if whenever an irreducible $f \in F[x]$ has one root in K, then it splits in K. If F has characteristic zero, then K is a finite normal extension of F if and only if K is a splitting field for some irreducible polynomial in $F[x]$ (Theorem 45.6). Furthermore, the roots of an irreducible polynomial over F are all distinct in its splitting field (Theorem 45.4), as long as the field is of characteristic zero. Another important property of fields of characteristic zero is this: any finite extension is actually simple (Theorem 45.5).

The section digresses a bit with a chapter on finite fields where we use the idea of splitting field to prove that every finite field of characteristic p (prime) has p^n elements (Theorem 46.1), and, indeed, a unique such field exists of order p^n, for every p and n. We denote this *Galois field* by $GF(p^n)$. Also, $GF(p^m)$ is a subfield of $GF(p^n)$ if and only if m divides n.

If E is an extension field of F, then the set of automorphisms of E that fix elements in F forms a group, denoted by $\text{Gal}(E|F)$ and called the *Galois group of the field E over F*. We then explore the relationship between the order of these Galois groups and the degrees of the field extensions and discover that these counts are equal in the case of normal extensions:

- (Theorem 47.1) Let $F \subseteq K$ be fields, and $f \in F[x]$ an irreducible polynomial, and $\alpha \in K \backslash F$ a root of f. Suppose that $\varphi \in \text{Gal}(F(\alpha)|F)$. Then φ is entirely determined by $\varphi(\alpha)$. Furthermore, $\varphi(\alpha)$ must be a root of f in K, and so
$$|\text{Gal}(F(\alpha)|F)| \leq \deg(f) = [F(\alpha) : F].$$
- (Theorem 47.2) Let K be a finite extension of the field F, which is of characteristic zero. Then $|\text{Gal}(K|F)| = [K : F]$ if and only if K is a normal extension of F.

If H is a subgroup of $\text{Gal}(K|F)$, then the elements of K fixed by H (called the *fixed field* of H and denoted $\text{Fix}(H)$) is a subfield of K containing F (Theorem 48.1). Once again, in the case of normal extensions, the situation is particularly nice: $\text{Fix}(\text{Gal}(K|F)) = F$ if and only if K is a normal extension of F (Theorem 48.2).

The theory culminates in the Fundamental Theorem of Galois Theory, which we presented in two parts (Theorems 48.3 and 48.4) but consolidate here. It shows how group theory and field theory mirror one another:

Suppose that K is a finite normal extension of the field F, which is of characteristic zero. There is a one-to-one order reversing correspondence between fields E with $F \subseteq E \subseteq K$ and the subgroups H of $\text{Gal}(K|F)$. We can describe this correspondence by two maps that are inverses of one another; namely, we have
$$E \longmapsto \text{Gal}(K|E)$$
and
$$H \longmapsto \text{Fix}(H).$$
Furthermore, an intermediate field E is a normal extension of F if and only if the Galois group $\text{Gal}(K|E)$ is a normal subgroup of $\text{Gal}(K|F)$. Furthermore, the Galois group $\text{Gal}(E|F)$ is isomorphic to
$$\text{Gal}(K|F)/\text{Gal}(K|E).$$

We now have the theory in place to prove that the general fifth degree polynomial equation cannot be solved by radicals. A *radical extension* of a field F is a simple algebraic extension $F(\beta)$ where $\beta^n \in F$, for some integer $n \geq 2$. A sequence of radical extensions
$$F = F_0 \subset F_1 = F_0(\beta_1) \subset F_2 = F_1(\beta_2) \subset \cdots \subset F_N = F_{N-1}(\beta_N),$$

where at each stage $\beta_i^{n_i} \in F_{i-1}$ is called a *root tower* over F and the last field F_N is an *extension of F by radicals.* If $f \in F[x]$ has a splitting field contained in an extension of F by radicals, we say that f is *solvable by radicals.* Notice how this is similar to our description of constructible numbers in Section VII. After some delicate adjustments to this root tower, we obtain a careful refinement fully described in Lemma 49.2. We then translate this lemma into group theory using the Fundamental Theorem, thus obtaining the theorem we need (Theorem 49.3):

Suppose that $f \in F[x] \subseteq \mathbb{C}[x]$ is solvable by radicals. Then the Galois group of the splitting field for f is solvable.

Example 49.4 considers the polynomial $x^5 - 6x + 3 \in \mathbb{Q}[x]$, which has Galois group S_5 (as we showed in Example 47.12). But S_5 is not solvable (Example 36.4) and so this polynomial is *not* solvable by radicals.

Hints and Solutions

We provide here various short answers asked for in the exercises and also provide hints for many of the longer problems and proofs. Please note that if an exercise asks for an example we provide one, although there are usually many other possibilities.

Chapter 1 — The Natural Numbers

b. $\mathbb{Z}$ and $\mathbb{Q}$ don't have a least element. The subset $\{x : 0 < x \leq 1\}$ does not have a least element.

d. Any finite subset of an ordered set has a least element.

3. Use mathematical induction and the triangle inequality.

5. Note that here your base case for induction is when $n = 4$. You can easily check that this statement is false for $n = 1, 2$ and 3.

9. 4 (see Theorem 2.4).

16. Suppose that S is a subset of $\mathbb{N}$ that does not have a least element. Prove that S is empty, by using induction on the set $\mathbb{N}\backslash S$.

17. It's probably easier to prove that the Strong Principle is equivalent to the Well-ordering Principle. This works because of Theorem 1.1 and Exercise 16.

Chapter 2 — The Integers

a.

$$-120 = 13(-10) + 10;$$

$$120 = (-13)(-9) + 3;$$

$$-120 = (-13)(10) + 10.$$

b. $0, 1, 2$. $\{\cdots, -5, -2, 1, 4, 7, \cdots\}$.

c. 1 and $|p|$.

d. The multiples of m.

e.

$$92 = 2^2 \cdot 23; 100 = 2^2 5^2; 101; 102 = 2 \cdot 3 \cdot 17;$$

$$502 = 2 \cdot 251; 1002 = 2 \cdot 3 \cdot 167.$$

f. $1/30$.

1. $(13)(21) + (-8)(34) = 1, (157)(772) + (-50)(2424) = 4.$

3. There are two things to prove here: each linear combination of a and b is a multiple of $\gcd(a, b)$ and each multiple of $\gcd(a, b)$ can be written as a linear combination of a and b. The latter part is easy to show. For the former, note that a common divisor of a and b also divides $ax + by$.

4. Use Exercise 3.

7. Consider $(n + 1)! + 2, (n + 1)! + 3, \cdots, (n + 1)! + (n + 1)$.

8. Use induction.

19. We know from Exercise 1.14 that $\binom{p}{k}$ is an integer.

Chapter 3 — Modular Arithmetic

a. $[0]_3 = \{\cdots, -6, -3, 0, 3, \cdots\}, [1]_3 = \{\cdots, -5, -2, 1, 4, \cdots\},$
$[2]_3 = \{\cdots, -4, -1, 2, 5, \cdots\}$.

b. No. Yes. Yes.

c. $[15]$.

d. Arithmetic modulo 12.

e. 7 o'clock. Wednesday.

f. $[1]$. No solution. $[8]$. $[3]$.

2. $[1], [2], [4], [7], [8], [11], [13]$; $[3]X = [2]$.

3. You need to have $13x = 1 + 28y$; what has this to do with the GCD identity?

4. $[2][3] = [0][3]$.

6. Recall that $\gcd(a, m)$ is the smallest positive integer that can be expressed as a linear combination of a and m.

7. $[1], [5], [7], [11], [13], [17], [19], [23]$. Each is its own inverse.

8. $[1], [3], [7], [9]$.

11. To show that two sets are equal, show that each is a subset of the other.

12. Let $m = 12$. $12 \in [4][6]$, but $12 \notin \{xy : x \in [4], y \in [6]\}$.

13. $[0], [1], [4], [2]$. $[0], [1], [4]$. $[0], [1], [4], [7]$.

Chapter 4 — Polynomials with Rational Coefficients

a. The sum is $3 - 2x + 2x^2 - (1/2)x^3 - (2/3)x^4$. The difference is $-1 - 2x - 2x^2 + (5/2)x^3 - (2/3)x^4$. The product is $2 - 4x + 2x^2 - (7/2)x^3 + (5/3)x^4 + 2x^5 - (17/6)x^6 + x^7$.

b. The quotient is $(3/2)x^3 - (5/4)x^2 + (5/8)x + (27/16)$; the remainder is $5/16$.

c. $1 + x, 1 - x$.

d. No. Yes. Yes. No.

1. $28(x^4 - x) = (5 - x)(3x^6 + 4x^5 - 3x^3 - 4x^2) + (3x - 14)(x^6 + x^5 - 2x^4 - x^3 - x^2 + 2x)$.

2. $x^2 - 3x + 2 = (2x + 1)((1/2)x - (7/4)) + (15/4)$. That there is no Division Theorem for $\mathbb{Z}[x]$ follows easily from the uniqueness statement in the Division Theorem for $\mathbb{Q}[x]$.

3. $x^3 - 2$, $(x - 1)^3$, $(x - 1)^2 x$, $(x - 1)(x - 2)(x - 3)$.

4. Use the Root Theorem 4.3.

10. Consider the degree of pq.

11. Argue that the degree of g must be zero.

12. For example, $2(x^3 + x)$ and x.

13. For example, $x^4 + 2x^3 + 3x^2 + 2x$.

Chapter 5 — Factorization of Polynomials

a. In any factorization, one factor must be of degree 0.

b. No roots.

c. $(x - 2)(x - 1)(x + 1)(x + 2)$.

d. For example, $(2x - 4)(x - 1)((1/2)x + (1/2))(x + 2)$.

e. No.

f. kf, for all $0 \neq k \in \mathbb{Q}$.

g. $(2x - 1)(x^2 + 4x + 1)$.

4. Yes. No. Yes.

5. First show that this polynomial can have no linear factors, and hence no cubic factors. Then consider what a quadratic factor could look like.

6. Write out $f(p/q) = 0$, and cross-multiply to eliminate denominators.

7. $2x^3 - 17x^2 - 10x + 9 = (x - 9)(x + 1)(2x - 1)$.

10. Choose a polynomial m in the set that does have minimal degree, and use the Division Theorem 4.2 to divide p by m.

11. Use Exercise 10 and the Division Theorem.

12. $x^{2n} + 1$.

15a. Use $p = 2$ and $p = 5$, respectively.

16. Suppose that a polynomial satisfies the criterion but has a nontrivial factorization in $\mathbb{Z}[x]$. Look at the divisibility by p of the coefficients of these factors, starting with the highest and lowest degree terms.

Chapter 6 — Rings

a. It is defined to be a function on all ordered pairs.

b. Yes. No. No.

c. Matrix multiplication; subtraction.

d. $\mathbb{Z}_6, M_2(\mathbb{Z})$.

e. Yes. No. No. Yes. Yes. Yes.

f. $2, 1/16; 0, 0; 8, 0; 2, 1; 4 + 12x^2, 81x^8 + 108x^6 + 54x^4 + 12x^2 + 1;$
$\begin{pmatrix} 4 & 8 \\ -4 & 12 \end{pmatrix}, \begin{pmatrix} -31 & 48 \\ -24 & 17 \end{pmatrix}$

3. Show that you get 0 when $(-a)b$ or $a(-b)$ is added to ab.

4. Use Exercise 3.

16. $(a+b)^2 = a^2 + ab + ba + b^2$ in an arbitrary ring. $(a+b)^2 = a^2 + 2ab + b^2$ in a commutative ring.

17. The same proof you used in Exercise 1.14 will work.

18a. See Exercise 6.4.

18b. Apply the hypothesis to the element $a + b$, where a, b are arbitrary elements of R.

23. See Exercises 21 and 22.

Chapter 7 — Subrings and Unity

a. $\pi\mathbb{Z}$ in $\mathbb{R}$; $\mathbb{Q}^*$ in $\mathbb{Q}$.

b. Yes. No. No. No. No. No.

c. $\{0\}, \mathbb{Z}_5; \{0\}, \{0, 2, 4\}, \{0, 3\}, \mathbb{Z}_6$; $\{0\}, \mathbb{Z}_7$;
$\{0\}, \{0, 2, 4, 6, 8, 10\}, \{0, 3, 6, 9\}, \{0, 4, 8\}, \{0, 6\}, \mathbb{Z}_{12}$.

d. $D_2(\mathbb{Z})$ in $M_2(\mathbb{Z})$. Does not exist. $2\mathbb{Z}$ in $\mathbb{Z}$. $\mathbb{Z}_5$.

e. X.

f. $(1, 1)$.

7. The relationship is this: let $f = x$.

10. $2\mathbb{Z} \cup 3\mathbb{Z}$.

11a. $1/2 \notin \mathbb{Z}_{(2)}$.

11c. Not closed under subtraction.

13. $Z(M_2(\mathbb{Z})) = \left\{ \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix} : a \in \mathbb{Z} \right\}$.

15b. $\{0\}$.

15c. $\{0, 2, 4, 6\}$.

20. Pick two arbitrary elements a and b, and use the distributive law from different directions on the product

$$(a+1) \circ (b+1).$$

25a. In $\mathbb{Z}_6$, $4^2 = 4$.

Chapter 8 — Integral Domains and Fields

a. $\mathbb{Z} \times \mathbb{Z}$: units: $\{(1,1), (1,-1), (-1,1), (-1,-1)\}$; zero divisors: $\{(a,0) : a \neq 0\} \cup \{(0,b) : b \neq 0\}$. $\mathbb{Z}_{20}$: units: $\{1, 3, 7, 9, 11, 13, 17, 19\}$; zero divisors:$\{2, 4, 5, 6, 8, 10, 12, 14, 15, 16, 18\}$.
$\mathbb{Z}_4 \times \mathbb{Z}_2$: units: $\{(1,1), (3,1)\}$;
zero divisors: $\{(0,1), (1,0), (2,0), (3,0), (2,1)\}$. $\mathbb{Z}_{11}$: units: $(\mathbb{Z}_{11})^*$; no zero divisors. $\mathbb{Z}[x]$: units $\{1, -1\}$; no zero divisors.

f. argument $-\tan^{-1}(4/3) \sim -.927$; modulus 5; inverse $(3/25) + (4/25)i$

g. 0 if $r > 0$; π if $r < 0$.

i. $|\alpha\beta| = |\alpha||\beta| = 1$.

j. $1 + \sqrt{3}i$: argument $-\pi/3$, modulus 2. $2 + 2i$: argument $\pi/4$, modulus $2\sqrt{2}$.

k. -1; $\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}i$; $.5403 + .8415i$; $-1 + \sqrt{3}i$.

l. $\mathbb{Z}_5$; does not exist; does not exist; $\mathbb{Q}$; $\mathbb{Z}$.

m. No. All finite domains are fields.

n. $9 = 4 \cdot 2 + 1 \to 1 = (-4)(2) + 9 \to [1] = [-4][2] \to [2]^{-1} = [-4] = [5]$.

o. $[2]^3 = [8] = [3] = [2]^{-1}$.

4. $\left\{ \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} : a \neq 0, b \neq 0 \right\}$.

5. The modulus of all powers is 1.

6. $[6]^{-1} = [6]^{17} = [6][36]^8 = [6][17]^8 = [6][289]^4 = [6][4]^4 = [6][256] = [6][9] = [54] = [16]$.

7. $101 = 2 \cdot 36 + 29, 36 = 1 \cdot 29 + 7, 29 = 4 \cdot +1$. Solving these backwards gives $1 = 5 \cdot 101 + (-14) \cdot 36$, and so $[36]^{-1} = [-14] = [87]$.

8. $S = \{[3], [6], [9], [12], [15], [18] = [1], [21] = [4], [24] = [7], [27] = [10], [30] = [13], [33] = [16], [36] = [2], [39] = [5], [42] = [8], [45] = [11], [48] = [14]\} = T$.

11b. 0,0,17.

13. $(1-a)^{-1} = 1 + a + a^2 + \cdots + a^{n-1}$.

15. In the product $[1][2] \cdots [p-1]$, consider multiplicative inverses.

16. In Exercise 14 we have rings with unity where every element is either 0, a zero divisor, or a unit. (This is false in $\mathbb{Z}$.)

17a. The functions that never take on value zero.

Chapter 9 — Polynomials over a Field

a. $x^2+3x+3; 0$. $3x^2+x+4; 2$. $-ix^2+(-2+2i)x+(2+2i); 7+2i$. $(1/\pi)x^3-(2/\pi+1/\pi^2)x^2+(2/\pi^2+1/\pi^3)x-(2/\pi^3+1/\pi^4); 1/3+2/\pi^3+1/\pi^4$.

b. $(x+1)(x+2)(x+3)(x+4)$.

c. $1+i$ is a unit.

d. $2x^2+3x+3, 4x^2+x+1, x^2+4x+4=(x+2)^2, 3x^2+2x+2$.

e. No roots.

f. In $\mathbb{Q}[x]$: x^3-2. In $\mathbb{R}[x]$: $(x-\sqrt[3]{2})(x^2+\sqrt[3]{2}+\sqrt[3]{4})$. In $\mathbb{C}[x]$: $(x-\sqrt[3]{2})(x-(-\sqrt[3]{2}/2+\sqrt{8\sqrt[3]{2}-\sqrt[3]{4}}/2i))(x-(-\sqrt[3]{2}/2-\sqrt{8\sqrt[3]{2}-\sqrt[3]{4}}/2i))$.

g. The modulus of this number is $\sqrt{12}$ and its argument is $5\pi/6$. Thus its square roots are $\pm\sqrt[4]{12}\,(\cos(5\pi/12)+i\sin(5\pi/12))$.

h. Never. x can't have a multiplicative inverse.

2a. If α is a square root, so is $-\alpha$.

2c. $1+i; 2+i$. $1+i; 2i$.

3. $x^3+2x+1, 1$. $x^5+4x+1, 1$.

4. Roots are 0,1,2,3,4,5. $\mathbb{Z}_6$ is not a field.

5. 1; 1; 1.

6. x^2+x+1. x^2+x+1. x^2+x+1.

7. $1=(1/54)(x^2-x-9)(x^2+x-2)+(1/54)(-x+4)(x^3+4x^2+4x+9)$.

9. $2x^5-9x^4+16x^3-14x^2+14x-5$.

10. $(x+1)(x^3+2x+1)$.

11. $x, x+1, x^2+x+1, x^3+x^2+1, x^3+x+1, x^4+x^3+x^2+x+1, x^4+x^3+1, x^4+x+1$. Note that $(x^2+x+1)^2=x^4+x^2+1$.

14. $v=-1$ or -2. Root is 3. The other two roots are $-3/2\pm\sqrt{3}/2i$.

16. $v^3=\frac{-1+\sqrt{5}}{2}=\rho$; ρ is a famous number called the *Golden Section*, which we discuss in Exercise 37.5. Then $u^3=1/\rho$. Thus the root given by Exercise 12 is $\sqrt[3]{\rho}-\sqrt[3]{1/\rho}$. By Exercise 15 the other two roots are then

$$\sqrt[3]{\rho}\zeta-\sqrt[3]{1/\rho}\zeta^2$$

and

$$\sqrt[3]{\rho}\zeta^2-\sqrt[3]{1/\rho}\zeta.$$

18.a. x^3+x^2-8x-6. $y^3-(25/3)y-(88/27)$. $v=\sqrt[3]{-44+117i}/3$, $u=\sqrt[3]{44+117i}/3$. $(4+3i)^3=-44+117i$.

19. Use the fact that if $(c+di)^3=a+bi$, then $(-c+di)^3=-a+bi$.

21. The cubic you get is $2b^3-b^2+6b-7=0$; this has an integer root you can find by inspection.

22. $3.55411(\cos(.5819)+i\sin(.5819))$, $3.55411(\cos(-2.5597)+i\sin(-2.5597))$.

24. Use Exercise 23:

$$\cos(2\pi k/5)+i\sin(2\pi k/5), k=0,1,2,3,4.$$

$$\sqrt[10]{2}(\cos(\pi k/20)+i\sin(\pi k/20)), k=0,1,2,3,4.$$

26. For $p=2$ consider the polynomial x^2+x+1; for $p>2$ consider polynomials of the form x^2-a for some $a\in\mathbb{Z}_p$.

28. Consider the roots of the polynomial $f-g$.

Chapter 10 — Associates and Irreducibles

a. Yes, No, Yes, No, Yes, Yes, No, Yes.

b. 0; 1,2,4,7,8,11,13,14; 3,6,9,12; 5,10.

c. $x^2+3x+4, 2x^2+x+3, 3x^2+4x+2, 4x^2+2x+1$.

d. $\{(a,b)\in\mathbb{Q}\times\mathbb{Q}: a\neq 0, b\neq 0\}$. All the units. $\{(a,0): a\neq 0\}$.

e. $(3+2\sqrt{2})(1+\sqrt{2})^k, (3+2\sqrt{2})(1-\sqrt{2})^k$, any integer k.

f. $5+i, -5-i, -1+5i, 1-5i$.

g. $n=\sqrt{n}\cdot\sqrt{n}$.

h. $3+\sqrt{2}, 5+\sqrt{2}, 5+2\sqrt{2}, 1+3\sqrt{2}$, etc.

i. No (3 in $\mathbb{Z}[\sqrt{2}]$).

j. Yes, No, No, Yes, Yes, Yes, No, No, No.

3. $2\sqrt{3}=\sqrt{12}\in\mathbb{Z}[\sqrt{3}]\cap\mathbb{Z}[\sqrt{12}]$.

5. $(8+3\sqrt{7})^n$. $\sqrt{7}(8+2\sqrt{7})^n$.

6. Think about the norm of these units.

7. $2=4\cdot 2$, $3=3\cdot 3$, $4=4\cdot 4$.

8. 2 and 6 are irreducible and $4=2\cdot 2$.

9. $\{(p,\pm 1): p \text{ is irreducible in } \mathbb{Z}\}\cup\{(\pm 1,q): q \text{ is irreducible in } \mathbb{Z}\}$.

12. $2=(3+\sqrt{7})(3-\sqrt{7}), 2=(4+\sqrt{14})(4-\sqrt{14})$.

16. Suppose by way of contradiction that $\sqrt{n}=\frac{a}{b}$. Square both sides, clear the denominator, and then use the Fundamental Theorem of Arithmetic.

Chapter 11 — Factorization and Ideals

a. No. In the second factorization 5 is not irreducible.

b. No. 3 and $3+3\sqrt{2}$ are associates.

c. No, Yes, No, No, No, Yes, Yes, No.

d. No, Yes, No, Yes.

e. $I=R$.

f. $\{a + b\sqrt{7} \in \mathbb{Z}[\sqrt{7}] : 7 \text{ divides } a\}$.

g. No.

h. $\langle 2 \rangle$ in $\mathbb{Z}_8$. None. $\langle 1 + i \rangle$. $\langle x \rangle$.

1. No, Yes, No, No, Yes, Yes, Yes, Yes, No, No.

3. $x^2 + 1$.

4. $x - i$.

5. $(x - 3)\,(x^2 - 3)$.

6. $\langle 0 \rangle, \langle 1 \rangle, \langle 2 \rangle, \langle 3 \rangle, \langle 4 \rangle, \langle 6 \rangle$.

7. $8 = 2 \cdot 2 \cdot 2 = (1 + \sqrt{-7})(1 - \sqrt{-7})$.

8. $\langle \alpha \rangle = \{a + b\alpha + c\alpha^2 : 5 \text{ divides } a\}$. $\langle 2 \rangle = \{a + b\alpha + c\alpha^2 : 2 \text{ divides } a, b, c\}$.

9b. Consider (for example) multiplication on the right by $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

15d. It helps to think about part c and Exercise 11.

17. $I = \langle a \rangle$.

21. If $r \neq 0, s \neq 0$, $A((r, s)) = \langle (0, 0) \rangle$. If $r \neq 0$, $A((r, 0)) = \{0\} \times S$. If $s \neq 0$, $A((0, s)) = R \times \{0\}$. $A((0, 0)) = R \times S$.

Chapter 12 — Principal Ideal Domains

a. $\langle 35, 15 \rangle = \langle 5 \rangle$. $\langle 12, 20, 15 \rangle = \mathbb{Z}$.

b. $\langle 4, x \rangle = \{a_0 + a_1 x + \cdots a_n x^n : 4 \text{ divides } a_0\}$. $\langle 4, x^2 \rangle = \{a_0 + a_1 x + \cdots a_n x^n : 4 \text{ divides } a_0, a_1\}$.

c. $\langle x - 1 \rangle$.

d. Both are $\mathbb{Z} \times \mathbb{Z}$.

e. The only ideals of a field are $\langle 0 \rangle, \langle 1 \rangle$.

f. $\mathbb{Z}[x]$. $x^2, -x^2$. $g = x^2 + x, f = x$.

1. To say that it is the smallest such ideal merely means that $\langle a, b \rangle$ should be a subset of *every* ideal containing a and b.

6. Check that if x and y are integers and $x + y$ is even, then $x - y$ is even too. Also, note that $2 = (1 + \sqrt{3})(-1 + \sqrt{3})$.

10b. If $n = 2$ and $I = \langle \sqrt{2} \rangle$, then $I = \bar{I}$. Example 12.3 gives an ideal where $I \neq \bar{I}$.

12a. The ideal consists of all polynomials with zero constant term.

12b. Consider the degrees in x and y for any possible generator.

Chapter 13 — Primes and Unique Factorization

a. None. $2 \in \mathbb{Z}[\sqrt{-5}]$. $6 \in \mathbb{Z}$.

b. $\langle 3 \rangle \subseteq \mathbb{Z}$. $\langle 3 \rangle \subseteq \mathbb{Z}[\sqrt{-5}]$. $\langle 3, 1 + \sqrt{-5} \rangle \subseteq \mathbb{Z}\sqrt{-5}]$. $\langle 4, x \rangle \subseteq \mathbb{Z}[x]$.

c. $\langle 12 \rangle \subsetneq \langle 3 \rangle$. $\langle 2 + \sqrt{3} \rangle = \mathbb{Z}[\sqrt{3}]$. $\langle 2 + i \rangle = \langle -1 + 2i \rangle$. $\langle 2x + 1 \rangle$ is a maximal ideal. $\langle 2x + 1 \rangle$ is maximal among principal ideals.

d. Maximal ideals must be proper.

e. We care about unique factorization!

f. PID implies UFD.

g. $1 = \gcd(9, 50) = 9x + 50y \in \langle 9, 50 \rangle$.

1. The proof for $\mathbb{Z}$ will work.

2. The proof for $\mathbb{Q}[x]$ will work.

3. $4^n = 4$. $(1, 0)^n = (1, 0)$.

7. 2 is not prime in $\mathbb{Z}[\sqrt{-5}]$ or $\mathbb{Z}[\sqrt{-7}]$.

8. $\langle 3, x \rangle = \{a_0 + a_1 x + \cdots a_n x^n : 3 \text{ divides } a_0\}$.

11b. $I = \langle \sqrt{2} \rangle$.

14a. What are its factors in $\mathbb{C}[x]$?

15. Show that 2 is an irreducible in this domain, using a norm argument. Then show that 2 is not prime.

Chapter 14 — Polynomials with Integer Coefficients

a. GCD identity, PID.

b. 6,7,42.

d. $2 \cdot 3 \cdot (x - 1)(x^2 + x + 1)$, no. $3x(x^3 - 2)$, no. $(x + 1)^2(5x - 1)(x - 2)$, yes.

e. None. $\mathbb{Z}[x]$.

f. $x^4 + 1$, none, 3, -1, none, $x^2 - 2$, $2x + 2$.

1b. $3x + 1$.

2c. $3x + 1$.

Chapter 15 — Euclidean Domains

a. $q = 16$, $r = 4$. $q = (1/2)x^2 - 2x - (1/4)$, $r = 2x - (13/4)$. $q = 4 - i$, $r = -1$. $q = 7/\pi$, $r = 0$.

b. Find a non-unit in the ideal with smallest valuation.

2. $q = -14 - 7\sqrt{2}$, $r = 3 - 3\sqrt{2}$.

3. Let $\beta = 2 + \sqrt{2}$, $\alpha = 2$.

4. The greatest common divisor will be greatest in what sense?

6b.

$$5 + 133i = (3 + i)(17 + 34i) + (-12 + 14i)$$

$$17 + 34i = (1 - 2i)(-12 + 14i) + (1 - 4i)$$

$$-12 + 14i = (1 - 4i)(-4 - 2i) + 0.$$

$$\gcd = 1 - 4i = (6-5i)(17+34i) + (-1+2i)(5+133i).$$

8. Find another element in $\mathbb{Z}[\sqrt{2}]$ with the same norm as $3+\sqrt{2}$, and show that it is not an associate by doing direct division (in $\mathbb{R}$).

9. For the Gaussian integers, we chose the quotient by looking at ordinary distance in the complex plane. For $\mathbb{Z}[\sqrt{2}]$, the 'distance' will involve hyperbolas!

Chapter 16 — Ring Homomorphisms

a. $f(x)$ is neither, $g(x)$ is both, $h(x)$ is one-to-one, but not onto, $k(x)$ is onto but not one-to-one.

b. Y has at least n elements.

c. Y has no more than n elements.

d. Example 16.1, 16.10, 16.4, 16.6, 16.10, none, 16.3 ($\pi(1,0)=1$), 16.10, 16.1 ($\varphi(2)$), 16.3 ($\pi(1,0)$).

e. No.

f. Does not preserve addition.

g. No, Yes.

1. Not unless $m=1$.

2. No.

3. No.

4. Yes.

5. Yes.

6. No, Yes.

7. Yes.

8. No.

9. Yes.

10. Yes.

11. No.

12. No.

13. Yes.

14. Yes.

15. No.

16. Yes.

17. No.

18. Yes.

19. Yes.

20. No.

21. No.

22. Yes.

23. No.

24. Yes.

25b. Consider $\varphi : \mathbb{C} \to \mathbb{C}$ given by $\varphi(a+bi) = a-bi$, as in Example 16.4.

30b. You will need some theorems from calculus.

31b. Let $R = \mathbb{Z}$ and $S = \mathbb{Z}[x]$, and consider the inclusion map.

Chapter 17 — The Kernel

a. $\{0,2,4\}$.

b. $\{0,2,4\}, \{1,3,5\}$.

c. 1,2. All preimages for a homomorphism have the same number of elements.

d. Identity map. $\varphi : \mathbb{Z} \to \mathbb{Z}_4, \varphi(n) = [n]$. *Any* one-to-one ring homomorphism. Impossible.

e. $\mathbb{Z} \subseteq \mathbb{Q}$.

f. None.

1. $\{0\}$.

2. $\left\{\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}\right\}$.

3. $\{b \in P(X) : x \notin b\} = \langle X \backslash b \rangle$.

4. $\{\{s_1, 0, 0, \cdots\}\} \subseteq S$.

6. $\{([0],[0])\}$.

7. $\{f \in C[0,1] : f(0) = f(1) = 0\}$.

10. No. Example 16.10.

11. $\{(0,s) : s \in S\}$. When they have the same first component. R.

13. $\{a_0 + a_1x + \cdots + a_nx^n : a_0 = a_1 = 0\} = \langle x^2 \rangle$.

14. Consider $f = 1 - e$.

16. Use Theorem 17.3.

18a. To do this, you must use Exercises 2.19 and 6.17.

Chapter 18 — Rings of Cosets

a. $4, \langle 4 \rangle + 0$, Yes, Yes.

c. $\langle x-2 \rangle + (-x^2+2)$.

d. $\langle x-2 \rangle + (1/2)$.

f. I.

g. Yes.

h. No.

i. Yes.

2. Consult Chapter 11 for a complete description of the ideals of $\mathbb{Z}$.

5. $(I + (3,1))(I + (1,4)) = I + (0,0)$.

7a. Treat these elements as polynomials in α, and divide.

8. Note that $\mathbb{Z} + (3/2) = \mathbb{Z} + (1/2)$. But $(\mathbb{Z} + (1/2))^2 = \mathbb{Z} + (1/4)$, while $(\mathbb{Z} + (1/2))(\mathbb{Z} + (3/2)) = \mathbb{Z} + (3/4) \neq \mathbb{Z} + (1/4)$.

9a. $\Sigma + (1,1,1,\ldots)$ consists of those sequences all but finitely many of whose terms are 1.

10c. $\mathbb{Z}_2$.

Chapter 19 — The Isomorphism Theorem for Rings

a. Different number of elements.

b. Yes.

c. Yes.

d. What ideals does every ring have?

e. Only if φ is onto.

g. No.

h. Does not preserve multiplication.

3. $\varphi(a + bi) = [a + b]_2$.

4. $\varphi(f) = f(1/2)$.

5. $\varphi(a, b) = [4a + 9b]_{12}$.

6. $\varphi(f, n) = (f(0), [n])$.

7. $\varphi(a_0 + a_1x + \cdots a_nx^n) = (a_0, a_1)$.

9. Build an isomorphism between these rings (with domain $P(X)$) by making use of the homomorphisms considered in Exercise 16.10.

10. A function of the form $\varphi(a + b\alpha\, c\alpha^2) = [ax + by\, cz]_6$ will work, where x, y, z are fixed elements of $\mathbb{Z}$.

11. $\mathbb{R}$ and C/I, where I is the ideal consisting of all sequences that converge to 0.

13. $\varphi(1) = 1$; $\varphi(-1) = -1$; $-1 = \varphi(i^2) = (\varphi(i))^2$.

14. $(\varphi(\sqrt{2}))^2 = 2$.

21b. Use the Isomorphism Theorem. Define a ring homomorphism from R/I onto R/A, and show that it has the appropriate kernel.

23. Define a homomorphism from I to $(I+J)/J$ with the appropriate kernel.

Chapter 20 — Maximal and Prime Ideals

a. $\langle x\rangle$ in $\mathbb{Z}[x]$. Impossible. $\{0\} \times \mathbb{Z}$ in $\mathbb{Z} \times \mathbb{Z}$.

b. $\langle x^2 - 2\rangle \subset \langle x - \sqrt{2}\rangle$.

c. $\mathbb{Z}[x]/\langle x - 2\rangle$ is isomorphic to the domain $\mathbb{Z}$.

d. Every field is a domain.

e. $\{(0,n) : n \neq 0\} \cup \{(m,0) : m \neq 0\}$.

$$\{(0,n) : n \neq 0\} \cup \{(2,n) : n \neq 0\} \cup \{(m,0) : m \neq 0\}.$$

f. R a domain; R a field.

3. Use the GCD identity.

4. $\{0\}$ and $\langle p\rangle$, p a prime integer. Prime $\equiv$ maximal.

5. $\langle 2\rangle, \langle 5\rangle$. Prime $\equiv$ maximal.

6. Prime, not maximal: $\{0\} \times \mathbb{Z}, \mathbb{Z} \times \{0\}$. Maximal: $\mathbb{Z} \times \langle p\rangle, \langle p\rangle \times \mathbb{Z}$, where p is prime.

7. Prime $\equiv$ maximal: $\mathbb{Z}_2 \times \{0\}, \{0\} \times \mathbb{Z}_3$. Prime $\equiv$ maximal: $\mathbb{Z}_2 \times \{0,2\}, \{0\} \times \mathbb{Z}_4$.

8. $\mathbb{Q}$: $\{0\}$, prime $\equiv$ maximal. $\mathbb{Q} \times \mathbb{Q}$: $\mathbb{Q} \times \{0\}, \{0\} \times \mathbb{Q}$, prime $\equiv$ maximal.

9b. $\ker(\rho) = \langle y\rangle \subset \langle x, y\rangle$.

9c. $\ker(\varphi) = \langle x - y\rangle \subset \langle x, y\rangle$.

10. The homomorphism is not onto.

11c. $\langle x^2 - 2\rangle$.

12. $\ker(\varphi) = \langle 3\rangle$.

13. An argument very similar to that we used in the text for $\langle x^2+1\rangle$ in $\mathbb{R}[x]$ will work.

14a. You need to use Gauss's Lemma.

17. Consider $R = \mathbb{Q}$ and its maximal ideal $I = \{0\}$.

23. Given $r \in R$, consider the product $r(1 - r)$.

Chapter 21 — The Chinese Remainder Theorem

a. $\mathbb{Z}_8 \times \mathbb{Z}_3, \mathbb{Z}_4 \times \mathbb{Z}_3 \times \mathbb{Z}_5, \mathbb{Z}_{11}, \mathbb{Z}_9$.

c. n divides m.

d. $x \equiv 5 \pmod 6$.

e.

$$\begin{aligned} x &\equiv 2 \pmod 3)\\ x &\equiv 3 \pmod 5)\\ x &\equiv 2 \pmod 7). \end{aligned}$$

1. $\mu([a]) = ([a]_3, [a]_4)$.

3. Prime $\equiv$ maximal; $\langle 4\rangle, \langle 2\rangle$.

5. $x \equiv 201(\text{mod } 252)$.

6. $x \equiv 83 (\text{mod } 4158)$.

7. $a = 2, b = 3$.

11b. $(0,1) \notin R$.

11d. Use the ideals P_1 and P_2.

11e. Note that $(3,0)(0,3) = (0,0) \in P$.

Chapter 22 — Symmetry of Figures in the Plane

c. $\begin{pmatrix} \sqrt{2}/2 & -\sqrt{2}/2 \\ \sqrt{2}/2 & \sqrt{2}/2 \end{pmatrix} \begin{pmatrix} \sqrt{3}/2 & -1/2 \\ 1/2 & \sqrt{3}/2 \end{pmatrix}$

e. Two.

f. One.

3. You should have 4 elements in your group.

4. You should have 4 elements in your group.

5a. $(x_2 - x_1)^2 + (y_2 - y_1)^2$.

7. There are an infinite number of rotations, one for each angle rotated, and an infinite number of flips, one about each line through the center.

8. Specify one original vertex position as the 'home' position. How many different vertices can be assigned to the home position? Now for each one of these assignments, in how many different ways can we visit the vertices as we travel around the polygon in, say, a clockwise direction?

Chapter 23 — Symmetry of Figures in Space

a. This works as long as faces are not on opposite sides of the cube.

b. $\rho_2^2, \varphi_1, \rho_2^2$.

d. $4\mu_4^2, \mu_1^2, \rho_3^3$.

1. There are four, all rotations of the base.

3. Flatlanders cannot understand flips.

7. $\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix}$.

Chapter 24 — Abstract Groups

a. $\mathbb{Z}, \mathbb{Z}_n, M_2(\mathbb{R})$, symmetries of the square.

b. $[2], [2], \begin{pmatrix} -5 & -4 \\ 4 & 3 \end{pmatrix}, \varphi\rho, (-2,4), (1/2, -1/4), 3/5 - 4/5i$.

c. $[4]$ in $\mathbb{Z}_5^*$; let $a = \varphi, b = \rho$ in D_3; any answer to c.a will do, but also $[1]$ in $\mathbb{Z}_4$.

d. R with operation addition and $U(R)$ with operation multiplication.

f. No, No, Yes, No, Yes, No, No, No, No, Yes.

g. No identity. Also, operation is not associative.

1. n^3.

2. $\{1,2,4,7,8,11,13,14\}$, $\mathbb{Q}^*$, $\{-1,1\}$, $\{-1,1\} \times \mathbb{Q}^*$.

4. You must first verify that this set is closed under the operation.

5a. You must first check that this is a binary operation on this set.

8. Apply the hypothesis to the element $x \circ y$.

12. Argue first that if φ is an automorphism of $\mathbb{Q}$, then $\varphi(n) = n$, for all integers $n \in \mathbb{Z}$. (We say that φ is **fixed** on the integers.)

14. Argue that for any automorphism ψ, we must have that $\psi(i)^2 = -1$ and $\psi(r) = r$, if $r \in \mathbb{R}$.

Chapter 25 — Subgroups

a. Yes, No, No, Yes, No, No.

b. Yes.

c. No.

d. One identity, one inverse per element.

e. ρ_1^3, α_4, 8, 14.

11.

$$U(\mathbb{Z}_8), \{1\}, \{1,3\}, \{1,5\}, \{1,7\}.$$

$$\mathbb{Z}_7^*, \{1\}, \{1,2,4\}, \{1,6\}.$$

$$U(\mathbb{Z}_{15}), \{1\}, \{1,4\}, \{1,11\}, \{1,14\}, \{1,2,4,8\}, \{1,7,4,13\}.$$

13. Only the trivial subgroup is finite. There are infinitely many — consider elements of the form $e^{2\pi/n}$.

14. You may wish to refer back to Exercise 22.8.

15a. $C(\rho) = \{\iota, \rho, \rho^2\}$.

15b. $C(4) = \mathbb{Z}_7$.

16a. The center contains the two elements $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$.

16b. $\mathbb{Z}_5$.

16d. The group is abelian.

Chapter 26 — Cyclic Groups

a. 3, 4, 2, ∞, ∞, 4, 4, 8.

b. Yes, No, No, Yes, No, No, No, Yes, No, No, Yes.

c. Yes, $\mathbb{Z} \times \mathbb{Z}_2$.

d. Yes.

e. Consider $\langle g \rangle$.

f. 2 (1, -1), 4 (1, 3, 7, 9), 6 (1, 2, 3, 4, 5, 6), 2 (2, 5).

h. 14, 7, 2, or 1.

i. It divides 8.

1. $\left\{\begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix} : n \in \mathbb{Z}\right\}, \left\{\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right\}.$

2. An element of the subgroup with 'smallest exponent' will serve as the generator, you'll need to use the Division Theorem 2.1 for $\mathbb{Z}$.

3. Any of several small finite groups we have considered will do the job. Provide more than one example if you can.

5. It must divide the lcm of m and n.

6. The intersection includes only the identity.

7. Two. If a is a generator, then so is $-a$.

9b. Consider the quaternions.

12. Consider the rotation in the subgroup of smallest angle.

13. The order is prime.

Chapter 27 — Group Homomorphisms

a. No.

b. Yes.

c. No.

e. No.

g. No.

h. $\{2x + a : a \in \mathbb{R}\}$.

i. $[0]_2\varphi = [0]_4$, $[1]_2\varphi = [2]_4$.

k. Yes.

2. There isn't one.

3. There are 4 homomorphisms, altogether.

5. Try a 'small' non-abelian group.

6. Model your proof on the corresponding ring theory theorem, Theorem 16.2.

8. You will need an appropriate theorem from Chapter 16.

10. Consider a function that has a derivative, but no second derivative.

12. A homomorphism from a cyclic group with generator g is entirely determined by the image of g.

13d. Define a homomorphism from End$(\mathbb{Z}_n)$ to $\mathbb{Z}_n$ with trivial kernel; by all means look at your work in Exercise 12b.

Chapter 28 — Group Isomorphisms

a. $\varphi : \mathbb{Z}_2 \to \mathbb{Z}_4$ where $\varphi(0) = 0, \varphi(1) = 2$; $\varphi : \mathbb{Z}_4 \to \mathbb{Z}_2$ where $\varphi(0) = \varphi(2) = 0, \varphi(1) = \varphi(3) = 1$; $\varphi : \mathbb{Z}_4 \to \mathbb{Z}_4$ where $\varphi(0) = \varphi(2) = 0, \varphi(1) = \varphi(3) = 1$.

b. Order of $\varphi(1)$ can be 1, 2, 4, or 8. Order of $\varphi(4)$ can be 1 or 2.

c. No, Yes.

d. D_4 is not abelian but the other two groups are. $\mathbb{Z}_8$ has an element of order 8 while $\mathbb{Z}_4 \times \mathbb{Z}_2$ does not.

e. Consider the elements 1 and i.

f. Yes, No.

g. No.

1. Let $\varphi(i) = 1$.

13. Not all criteria are satisfied; D_3 is not isomorphic to a direct product of these groups.

Chapter 29 — Permutations and Cayley's Theorem

a. $\alpha(5) = 1$, $\alpha(4) = 4$, $\alpha^{-1} = \begin{pmatrix} 1\,2\,3\,4\,5\,6\,7 \\ 5\,3\,1\,4\,7\,6\,2 \end{pmatrix}$.

b. $\beta^2 = \begin{pmatrix} 1\,2\,3\,4\,5\,6\,7 \\ 1\,7\,3\,4\,5\,2\,6 \end{pmatrix}$, $\beta\alpha = \begin{pmatrix} 1\,2\,3\,4\,5\,6\,7 \\ 3\,2\,6\,1\,4\,7\,5 \end{pmatrix}$, $\alpha\beta = \begin{pmatrix} 1\,2\,3\,4\,5\,6\,7 \\ 4\,6\,2\,3\,1\,5\,7 \end{pmatrix}$, $\beta\alpha^{-1} = \begin{pmatrix} 1\,2\,3\,4\,5\,6\,7 \\ 5\,3\,4\,1\,2\,7\,6 \end{pmatrix}$, $\alpha\beta\alpha = \begin{pmatrix} 1\,2\,3\,4\,5\,6\,7 \\ 2\,7\,6\,3\,4\,5\,1 \end{pmatrix}$.

c. $(\alpha\beta)^{-1} = \begin{pmatrix} 1\,2\,3\,4\,5\,6\,7 \\ 5\,3\,4\,1\,6\,2\,7 \end{pmatrix}$.

d. 24, 120, 720.

e. Every finite group is isomorphic to a subgroup of S_n for some n.

f. Isomorphic to a subgroup of S_{12} if considered as permuting the edges, to a subgroup of S_8 if considered as permuting the vertices, to a subgroup of S_6 if permuting the sides, and to a subgroup of S_4 if permuting the diagonals.

g. Can always embed in S_n for sufficiently large n.

1. 5, 6, 3, 6, 6, 10, 12.

3. 1 corresponds to $\begin{pmatrix}1\,2\,3\,4\\1\,2\,3\,4\end{pmatrix}$, -1 corresponds to $\begin{pmatrix}1\,2\,3\,4\\2\,1\,4\,3\end{pmatrix}$, i corresponds to $\begin{pmatrix}1\,2\,3\,4\\3\,4\,2\,1\end{pmatrix}$, -i corresponds to $\begin{pmatrix}1\,2\,3\,4\\4\,3\,1\,2\end{pmatrix}$.

6. G_k has $(n-1)!$ elements.

7. G_K has $(n-|K|)!$ elements.

Chapter 30 — More about Permutations

a. (134572), (1432), (123765).

b. (153)(29)(678) of order 6.

c. $(123)(4567)(8\,9\,10\,11)$ of order 12.

d. If both permutations are the same.

e. No, No, No.

1. $\{(357),(375),\iota\}$, $\{(14)(256),(265),(14),(256),(14)(265),\iota\}$, $\{(123)(456),(132)(465),\iota\}$

2. Do an induction on the number of cyclic factors.

3. $(123)^2$ is a cycle, but $(1234)^2$ is not.

4. 3, 12, 6.

5. 3, 5, 2. To obtain this, carefully consider all possible cases for the way their supports overlap.

8a. One 2-cycle, one 3-cycle, one 4-cycle, product of 2-cycles.

Chapter 31 — Cosets and Lagrange's Theorem

a. $\langle 6\rangle = \{6,12,18,4,10,16,2,8,14,0\}$, $\langle 6\rangle + 1 = \{7,13,19,5,11,17,3,9,15,1\}$.

b. $U(\mathbb{Z}_{20}) = \{1,3,7,9,11,13,17,19\}$, $\langle 7\rangle = \{7,9,3,1\}$, $\langle 7\rangle\cdot 11 = \{17,19,13,11\}$, $[U(\mathbb{Z}_{20}):\langle 7\rangle] = 2$, $|\langle 7\rangle| = 4$.

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

d. Those two cosets are in fact identical.

e. Yes, if $Ha = H$.

f. $Ha = Hb$.

g. Infinite.

h. Infinite.

1. $\langle(124)\rangle = \{(124),(142),\iota\}$, $\langle(124)\rangle(123) = \{(14)(23),(234),(123)\}$, $\langle(124)\rangle(132) = \{(134),(13)(24),(132)\}$, $\langle(124)\rangle(143) = \{(243),(12)(34),(143)\}$.

3. $H = \{\iota,(12)\}$, $(13)H = \{(13),(123)\}$, $(23)H = \{(23),(132)\}$. These are not the same as the right cosets.

4. $K = \{\iota,(123),(132)\}$, $(12)K = \{(12),(23),(13)\}$. These are the same as the right cosets of K.

12a. Corollary 31.3 says that the order of a divides $|G|$.

12b. Apply part a to the group $\mathbb{Z}_p^*$ of units of $\mathbb{Z}_p$.

12c. Use the same idea as in part b, by applying part a to the group $U(Z_n)$ of units of $\mathbb{Z}_n$.

13a. 6, 8, 40.

13b. 4, 9, 49.

Chapter 32 — Groups of Cosets

a. $H\iota = \iota H$.

b. Because $ab = ba$ for all $a, b \in G$, then $aH = Ha$ for all $a \in G$.

c. $\mathbb{Z}_7$.

d. $\mathbb{Z}$ does not have the absorption property and so is not an ideal of $\mathbb{R}$. But the *group* $\mathbb{Z}$ (under addition) is a normal subgroup in the abelian group $\mathbb{R}$.

e. $\{1\}$ is certainly a subgroup. Also, $a\{1\} = \{a\} = \{1\}a$ for all $a \in G$.

f. No, K is not closed under the group operation.

3. No, No.

4. Yes.

7. Define a function

$$\psi : \mathbb{Z}_n/\langle d\rangle \to \mathbb{Z}_d,$$

and prove that it works.

10. Use the Index 2 Theorem 32.3.

Chapter 33 — The Isomorphism Theorem for Groups

c. No.

d. Yes.

1. The kernel is $\mathbb{Z}$. The groups $\mathbb{R}/\mathbb{Z}$ and $\mathbb{S}$ are isomorphic.

4. First define a homomorphism from $\mathbb{Z}_n$ to $\mathbb{Z}_d$ and use the Fundamental Isomorphism Theorem.

9. Define a homomorphism from H to HK/K with the appropriate kernel.

10. The kernel is a plane through the origin. Geometrically, when we mod a plane out of three dimensional space, we get a line.

13. Note that H/K is actually a normal subgroup of G/K, by Exercise 13. Find a homomorphism from G onto $(G/K)/(H/K)$ with the correct kernel.

Chapter 34 — Alternating Groups

b. 60, 360.

c. The identity permutation is not odd.

d. The subgroup would then be normal.

e. The subgroup $\{\iota, (12)\}$ in S_3.

f. Even.

1. Yes.

3a. The identity; 20 3-cycles; 24 5-cycles; 15 products of two disjoint 2-cycles.

3b. The trivial subgroup; 10 distinct order 3 subgroups; 15 distinct order 2 subgroups; 6 distinct order 5 subgroups.

3c. Use Theorem 34.2.

5. $(1456)(29) = [(712)(3956)(48)](7895)(13)[(217)(6593)(84)]$.

6. You must use Theorem 34.4.

8. Use Exercise 7.

9. Look at repeated conjugation of the transposition by the n-cycle; use Exercise 8.

11. Let $K = H \cap A_n$. What is $[H : K]$?

12. We claim that you need only show that any product of two transpositions is a product of 3-cycles (justify this).

Chapter 35 — Fundamental Theorem for Finite Abelian Groups

a. 1, 2, 2, 1.

b. $\mathbb{Z}_4 \times \mathbb{Z}_2 \times \mathbb{Z}_3$, $\mathbb{Z}_2 \times \mathbb{Z}_3 \times \mathbb{Z}_{25}$, $\mathbb{Z}_2 \times \mathbb{Z}_4$, $\mathbb{Z}_2 \times \mathbb{Z}_9$, $\mathbb{Z}_3 \times \mathbb{Z}_4$.

c. $p = 2$, No, $p = 27$, No, $p = 3$, $p = 2$.

d. 50 is not a power of a prime.

e. $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$ has no element of order 4.

f. 1.

g. 2.

1. Show that the element $(1, 1)$ does the job.

4. kj.

5. Use the Fundamental Theorem for Finite Abelian Groups 35.1.

6. Use the group S of infinite sequences of real numbers, under addition (see Exercise 6.19).

Chapter 36 — Solvable Groups

a. Yes.

b. No. S_3.

c. Pick any proper subgroup H of Q_8. Then H and Q_8/H are abelian. (Check these assertions.)

2. If H is a subgroup of a solvable group G with the usual subnormal series, let $H_i = H \cap G_i$.

3. An easy proof uses the Fundamental Theorem for Finite Abelian Groups 35.1.

4. Consider Exercise 3.

7. Proceed by induction on the order of the group. Use Exercise 6.

Chapter 37 — Constructions with Compass and Straightedge

a. Yes; construct a 30° angle (how?) and bisect it.

b. Yes.

c. No.

d. No; it is imaginary.

e. Because we could the construct a square with area π.

f. Because $\sqrt[3]{2}$ is the length of the edge of a cube with volume 2.

g. An arbitrarily long straightedge.

h. An arbitrarily large compass.

2b. Construct a right triangle with legs of lengths 1 and 2.

3. To construct $\sqrt[4]{3} + 1$, construct 3, $\sqrt{3}$, $\sqrt[4]{3}$, then $\sqrt[4]{3} + 1$.

4. Show that both numbers are the positive square root of the same number.

Chapter 38 — Constructible Numbers and Quadratic Field Extensions

a.

$$\sqrt{6} \in \mathbb{Q}(\sqrt{6}).$$

$$\sqrt[8]{6} \in \mathbb{Q}(\sqrt{6}, \sqrt[4]{6}, \sqrt[8]{6}) \supseteq \mathbb{Q}(\sqrt{6}, \sqrt[4]{6}) \supseteq \mathbb{Q}(\sqrt{6}) \supseteq \mathbb{Q}.$$

$$\sqrt{2 + \sqrt[4]{5}} \in \mathbb{Q}\left(\sqrt{2 + \sqrt[4]{5}}\right) \supseteq \mathbb{Q}(\sqrt[4]{5}) \supseteq \mathbb{Q}(\sqrt{5}) \supseteq \mathbb{Q}.$$

b. $\mathbb{C}$.

c. No.

d. Yes: $\mathbb{K}(i)$; No.

e. No.

f. If (a, b) is the point of intersection, then both a and b are elements of the field F.

g. Yes.

4. 2, 3, 4, $\sqrt{3}$, $4\sqrt{3}$, $2+4\sqrt{3}$, $\sqrt{2+4\sqrt{3}}$, $\sqrt[4]{2+4\sqrt{3}}$; $\mathbb{Q}$, $\mathbb{Q}$, $\mathbb{Q}$, $\mathbb{Q}(\sqrt{3})$, $\mathbb{Q}(\sqrt{3})$, $\mathbb{Q}(\sqrt{3})$, $\mathbb{Q}(\sqrt{3})(\sqrt{2+4\sqrt{3}})$, $\mathbb{Q}(\sqrt{3})(\sqrt{2+4\sqrt{3}})(\sqrt[4]{2+4\sqrt{3}})$.

9. The line in question has equation $y = mx + b$ and passes through the point $(3, 1)$. Perform this intersection algebraically. When does a quadratic equation have a unique root?

Chapter 39 — The Impossibility of Certain Constructions

a. $\sqrt[3]{2}$, π.

b. We could construct a right triangle with hypotenuse 1 and one side of length $\cos\theta$. The adjacent angle would be θ.

c. Some irrational numbers are constructible, such as $\sqrt{2}$.

d. Double the length of a side of the original square, then construct the square root of this. Use this length for the sides of a new square.

e. Simply construct an edge of twice the original length.

1a. Consider the proof of the appropriate lemma in this chapter.

4a. Use Exercise 3.

5. This is a slight generalization of Lemma 39.3; the same proof will work.

7. This is a slight generalization of Lemma 39.1; the same proof will work.

10. What does the derivative of this function tell you?

Chapter 40 — Vector Spaces I

a. $(4, 2)$, $(0, -4)$, $(4, 6)$, $(1, -1/2)$, $(-2, 1)$.

b. 0 is a scalar, $\mathbf{0}$ is the zero vector.

c. Scalars are 0, 1, 2. Vectors are the polynomials $0, 1, 2, x, x+1, x+2, 2x, 2x+1, 2x+2, x^2, x^2+1, x^2+2, 2x^2, 2x^2+1, 2x^2+2, x^2+x, x^2+x+1, x^2+x+2, x^2+2x, x^2+2x+1, x^2+2x+2, 2x^2+x, 2x^2+x+1, 2x^2+x+2, 2x^2+2x, 2x^2+2x+1, 2x^2+2x+2$.

d. Yes, No.

e. $-\mathbf{v}$ is the vector added to $\mathbf{v}$ to get $\mathbf{0}$. $(-1)\mathbf{v}$ is the vector $\mathbf{v}$ multiplied by the scalar -1.

f. Yes, No.

g. No.

12. In $\mathbb{Z}_p$, we know that $[0] = [1] + [1] + \cdots + [1]$.

13. Consider the scalar product $(1/2)1$.

Chapter 41 — Vector Spaces II

a. $\{(1,0),(0,1),(1,1)\}$ in $\mathbb{R}^2$; can't happen; $\{(1,0),(0,1)\}$ and $\{(1,0),(1,1)\}$ in $\mathbb{R}^2$; can't happen; $\{(1,0),(0,1),(1,1)\}$ in $\mathbb{R}^2$; $\{(1,0,0),(0,0,1),(0,1,0),(1,1,1)\}$ in $\mathbb{R}^3$.

b. The set of vectors would not be linearly independent.

c. 2, 1.

7. If $\mathcal{S}$ is a finite subset of $\mathbb{Q}[x]$, give a limit on the degree of a polynomial expressible as a linear combination of vectors from $\mathcal{S}$.

10. No, No, Yes, No, Yes, Yes.

11. Yes, Yes, Yes, Yes.

Chapter 42 — Field Extensions and Kronecker's Theorem

a. $x^3 - 2$; $x^3 - 2$ or $x - \sqrt[3]{2}$; $x^2 - 2x - 1$; $x^4 - 2x^2 - 1$; $x^2 - 2x + 5$; $x^2 - 2\pi x + \pi^2 + 1$.

b. Yes, No.

c. The original field.

d. There are none.

e. None, π, $\sqrt[3]{2}$.

f. 0, $\mathbb{Q}$; 0, $\mathbb{Q}$; 0, $\mathbb{Q}$; 11, $\mathbb{Z}_{11}$; 0, $\mathbb{Q}$; 3, $\mathbb{Z}_3$.

g. See Kronecker's Theorem.

h. Yes, $\cos 20°$ is a root of $8x^3 - 6x - 1$.

i. $\sqrt[3]{2}$, $\cos 20°$.

1. Use Kronecker's Theorem 42.1, together with induction on the degree of f.

2. If p is linear, let $f \in F[x]$ and use the Division Theorem to divide f by p. Now consider the degree of the remainder. Conversely, what property does $F[x]/\langle p\rangle$ have that F does not?

3b. Think Gauss's Lemma 5.5.

4. You know α is a root of a polynomial in $\mathbb{Q}[x]$; find the corresponding polynomials for the given elements.

5. Think trig identities.

7a. What elements must belong to any subfield of F?

Chapter 43 — Algebraic Field Extensions

a. 1, $\sqrt[3]{2}$, $\sqrt[3]{4}$; 1; 1, $1+\sqrt{2}$; 1, $\sqrt{1+\sqrt{2}}$, $1+\sqrt{2}$, $(1+\sqrt{2})^{3/2}$; 1, $1+2i$; 1, $\pi+i$.

b. $3/2\cdot 1 - 1/2(1+\sqrt{2}), 8/17\cdot 1 + 1/17\sqrt[3]{2} - 2/17\sqrt[3]{4}$.

c. No, $\mathbb{Q}(\pi)$ over $\mathbb{Q}$; Yes; No, π in $\mathbb{Q}(\pi)$ over $\mathbb{Q}$; Yes; Yes.

d. It divides every polynomial with that root, and it is monic.

2b. 64.

4. $x^4 - 10x^2 + 1$, $x^2 - 2\sqrt{2}x - 1$, $x - \sqrt{2} - \sqrt{3}$.

5. You will need to solve for the six rational number unknowns that will serve as coefficients on the powers of α. Cross-multiply to eliminate the denominator, and then use the minimal polynomial to eliminate powers of α higher than five, as we did in the proof of Theorem 43.3.

8c. Consider the element $1/\alpha$.

Chapter 44 — Finite Extensions and Constructibility Revisited

a. They are equal.

b. E is either F or K.

c. They are equal.

d. See Theorem 44.1; $\mathbb{Q}\subset\mathbb{Q}(\pi)$ is not algebraic; $\mathbb{Q}\subset\mathbb{Q}(\pi)$ is not finite; Consider the field of all constructible numbers as an extension field of $\mathbb{Q}$; See the previous field.

e. 7^6.

f. $\mathbb{Q}(\sqrt{2})(\sqrt{3})$; There are none; $\mathbb{Z}_2(\alpha)$ where α is a root of x^3+x+1.

1. First show that $\sqrt{2}+\sqrt{3}$ is a root of $x^4-10x^2+1 \in \mathbb{Q}[x]$, and that this polynomial is irreducible in $\mathbb{Q}[x]$. Therefore, $[\mathbb{Q}(\sqrt{2}+\sqrt{3}):\mathbb{Q}] = 4$, and consequently $\sqrt{2}+\sqrt{3}\notin\mathbb{Q}(\sqrt{3})$. Conclude that $\sqrt{2}\notin\mathbb{Q}(\sqrt{3})$.

3. For a computational proof, compute $(\sqrt{2}+\sqrt{3})^3$ and then subtract an appropriate multiple of $\sqrt{2}+\sqrt{3}$ from your result to get a multiple of $\sqrt{2}$. From this, argue that

$$\sqrt{2}\in\mathbb{Q}(\sqrt{2}+\sqrt{3}).$$

Then conclude that $\sqrt{3}\in\mathbb{Q}(\sqrt{2}+\sqrt{3})$. Thus, $\mathbb{Q}(\sqrt{2},\sqrt{3})\subseteq\mathbb{Q}(\sqrt{2}+\sqrt{3})$. To show containment the other way is easy. A proof along the lines of Exercise 1 using Theorem 43.2 is also possible.

4b. Use an argument similar to the one found in Example 44.3.

5. Suppose by way of contradiction that $[F:\mathbb{Q}] = n$.

7. Use Theorem 44.2, but be careful.

8. Use DeMoivre's Theorem 8.4 to find the minimal polynomial of ζ over $\mathbb{Q}$. (Or refer to Exercise 5.17.)

9b. In this finite case you can actually write down all possible irreducible polynomials in $\mathbb{Z}_2[x]$ of degree 2 or 3. Show by direct calculation that $\alpha+\beta$ is not a root of any of these. Conclude that $\mathbb{Z}_2(\alpha+\beta)=\mathbb{Z}_2(\alpha,\beta)$.

10. Show that if $\alpha,\beta\in\mathbb{A}$, then $\mathbb{Q}(\alpha+\beta)$ and $\mathbb{Q}(\alpha\beta)$ are finite extensions. This exercise generalizes Exercise 42.4.

Chapter 45 — The Splitting Field

a. True; True; False ($\mathbb{C}$ is a normal extension of $\mathbb{Q}$); True; True; False (let $F=\mathbb{Q}$ and $f=x^3-2$); False (let $F=\mathbb{Q}$ and $f=x^2-2$); True; True; False (consider $\mathbb{Q}\subset\mathbb{Q}(\sqrt[3]{2})$).

b. Let ζ be the primitive cube root of unity. Then the answers are: $\mathbb{Q}(\sqrt[3]{2},\zeta)$; $\mathbb{R}(\zeta)=\mathbb{R}(i)=\mathbb{C}$; $\mathbb{Q}(i)(\sqrt[3]{2},\sqrt{3})$; $\mathbb{C}$.

c. F.

3c. Use induction on the degree of g.

3d. Use induction on n.

5b. Let $\alpha=\sqrt{2+\sqrt{2}}$ and $\beta=\sqrt{2-\sqrt{2}}$. Then $f=(x-\alpha)(x+\alpha)(x-\beta)(x+\beta)$.

5c.$\mathbb{Q}(\alpha,\beta)$.

5d. Show that β can be expressed in terms of α, and so the splitting field is $\mathbb{Q}(\alpha)$.

6b. Let $\alpha=\sqrt[4]{2}$. Then $f=(x-\alpha)(x+\alpha)(x-\alpha i)(x+\alpha i)$.

6c. $\mathbb{Q}(\alpha,i)$.

6e. Since $\mathbb{Q}(\alpha,i)$ is a finite algebraic extension of $\mathbb{Q}$.

Chapter 46 — Finite Fields

a. $p=x^2+2$ would work, among others.

b. It has prime order.

c. $\mathbb{Z}_9$; $\mathbb{Z}_3\times\mathbb{Z}_3$; $GF(9)$.

1c. There are two cases, depending on whether p is even or odd.

3. Use Exercise 46.1a.

5. You will need to show that the polynomial $x^{m-1}-1$ divides $x^{n-1}-1$ if and only if m divides n.

6. Any element of such a field is a root of a polynomial of the form $x^{p^n} - x$.

7. Look at the polynomial term by term. Use Exercise 46.1a and Exercise 46.6.

8. Use the proof of Theorem 45.4, together with Exercise 46.7.

10b. Eliminate all reducible polynomials. First eliminate those with a linear factor. What fourth degree polynomials are a product of two degree two irreducibles?

10c. Use a similar analysis to part b.

Chapter 47 — Galois Groups

a. True; True; True; False; False (it is the trivial group).

b. Cyclic group of order eleven.

c. Example 47.5; None; Examples 47.7, 47.8, 47.9.

1. Klein Four group; cyclic group of order two.

2. Cyclic group of order 3.

3. Cyclic group of order four; cyclic group of order two.

4. Use Theorem 47.4.

8. Exercise 34.9 still applies.

Chapter 48 — The Fundamental Theorem of Galois Theory

a.

$$\alpha_1\alpha_2+\alpha_1\alpha_3+\alpha_1\alpha_4+\alpha_1\alpha_5+\alpha_2\alpha_3+\alpha_2\alpha_4+\alpha_2\alpha_5+\alpha_3\alpha_4+\alpha_3\alpha_5+\alpha_4\alpha_5.$$

b. $\mathrm{Gal}(K|E)$ is a subgroup of $\mathrm{Gal}(K|F)$.

c. $\mathrm{Gal}(E|F)$ is isomorphic to $\mathrm{Gal}(K|F)/\mathrm{Gal}(K|E)$.

d. $\mathrm{Fix}(H_2) \subseteq \mathrm{Fix}(H_1)$.

e. There are none.

1. For the Galois group, see Theorem 47.4.

2. This will involve an easier version of the arguments in Example 48.7.

3. It is not easy to find all of these fields. Keep in mind that the splitting field includes such elements as $\sqrt{2}, i, \sqrt{2}i$, etc., and that every distinct subgroup must correspond to a distinct intermediate field.

8. Assume by way of contradiction that G is not transitive. Then there is a proper subset K of the set of roots of f in K, for which $\varphi(\alpha) \in K$, whenever $\varphi \in G$ and $\alpha \in K$. Form the product g of the linear terms $x - \alpha$, for each $\alpha \in K$. Argue that $g \in F[x]$, using a symmetric polynomial argument.

Chapter 49 — Solving Polynomials by Radicals

a. $\mathbb{Q} \subset \mathbb{Q}(\sqrt[4]{2}) \subset \mathbb{Q}(\sqrt[4]{2}, i)$.

b. $\mathbb{Q} \subset \mathbb{Q}(\sqrt[3]{2}) \subset \mathbb{Q}(\sqrt[3]{2}, \zeta)$.

c. $\mathbb{Q} \subset \mathbb{Q}(\sqrt[3]{2})$.

d. See Theorem 47.4.

e. Yes.

f. Find an irreducible cubic with three real roots.

g. f is not solvable by radicals.

1. Make a chart as in Example 47.9, considering that $\alpha = \sqrt[3]{1+\sqrt{2}}$ can be mapped only to $\alpha, \alpha\zeta, \alpha\zeta^2$, ζ can be mapped only to ζ, ζ^2, and $\sqrt{2}$ can be mapped only to $\pm\sqrt{2}$.

3b. You know that the α_i are distinct (why?). Use this to argue that the β_i are distinct. Then show that the permutations in H are the only elements of S_4 that leave the β_i fixed.

5. $g = x^3+8x = x(x^2+8)$. Clearly the splitting field for g is $\mathbb{Q}(2\sqrt{2}i)$, which as a quadratic extension has Galois group over $\mathbb{Q}$ isomorphic to $\mathbb{Z}_2$.

Guide to Notation

In this appendix we provide a guide to the mathematical notation we use in this book. In many cases we provide a reference in the text where the given notation first occurs. Rather than attempting an alphabetical listing, we have grouped this list of notations conceptually.

Set Theory

Modern mathematics is expressed in terms of set theory, and we use standard notation for these concepts. In what follows, assume that A and B are sets:

If a is an element of A, then we write $a \in A$. If it isn't, we write $a \notin A$.

The *intersection* of A and B is

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

The *union* of A and B is

$$A \cup B = \{x : x \in A \text{ or } x \in B\}.$$

The *set difference* is

$$A \backslash B = \{x : x \in A \text{ and } x \notin B\}.$$

If A is a subset of B, we write $A \subseteq B$. If it is a *proper* subset we write $A \subset B$. We can change emphasis and write these two as $B \supseteq A$ and $B \supset A$, respectively.

If the set A is finite we denote by $|A|$ the number of elements in A.

The *set product* is the set of all ordered pairs, with first entry from A and second entry from B:

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

If A and B are rings or groups, we can place a ring (Example 6.10) or group (Section 27.3) structure on $A \times B$.

Numbers

$\mathbb{N}$ is the set of *positive integers* (or *counting numbers*), $\mathbb{Z}$ is the set of all *integers*, and $\mathbb{Q}$ is the set of *rational numbers* (Chapter 1).

For integers a, b and positive integer n, $a \equiv b \pmod{n}$ means that a and b are *congruent modulo n* (Chapter 3).

$\mathbb{R}$ is the set of *real numbers* (Chapter 6), and $\mathbb{C}$ is the set of *complex numbers* (Exercise 6.11).

$\mathbb{K}$ is the set of *constructible numbers* (Section 37.2), and $\mathbb{A}$ is the set of *algebraic numbers* (Exercise 42.4).

Functions

$\det(A)$ is the *determinant*, which is a real-valued function defined on matrices (Example 27.4).

$\deg(f)$ is the *degree* of the polynomial f (Section 4.1). cont (f) is the *content* of a polynomial $f \in \mathbb{Z}[x]$ (Section 14.3).

$\phi(n)$ is *Euler's phi function* (Exercise 31.12).

$\gcd(a, b)$ is the *greatest common divisor*, for integers, polynomials or Euclidean domains (Section 2.2, Section 4.6, Exercise 15.4). Likewise, $\operatorname{lcm}(a, b)$ is the *least common multiple* (Exercise 2.11).

$|\alpha|$ is the ordinary absolute value, if α is a real number; it is the *modulus* if α is complex (Section 8.4). $\arg(\alpha)$ is the *argument* of the complex number (Section 8.4).

We use several important functions from calculus: the *natural logarithm* $\log(x)$, the *exponential function* $\exp(x) = e^x$ and the *trigonometric functions* $\sin(x), \cos(x)$, etc.

Rings

In what follows, assume that R is a ring and F is a field.

$\mathbb{Z}_n$ is the ring of integers, modulo n (Chapter 3 and Example 6.4). Of course, we also can consider it as a cyclic group (Example 24.3 and Example 26.17). $R[x]$ is the ring of polynomials in indeterminate x and coefficients from the ring R (see Exercise 6.23); particularly important are $\mathbb{Q}[x]$ and $\mathbb{Z}[x]$ (Chapters 4 and 5), and $F[x]$, where F is a field (Chapter 9).

$\mathbb{Z}[i]$ is the ring of *Gaussian integers* (Exercise 6.12); more generally, $\mathbb{Z}[\sqrt{n}]$ is a *quadratic extension of the integers* (Section 10.3). These rings are equipped with a *norm function* $N(\alpha)$ (Section 10.4).

$M_2(R)$ is the ring of 2×2 *matrices*, with entries from the ring R; see Example 6.13 and Exercise 6.8.

For any set X, $P(X)$ is the set of all subsets of X; it is called the *power set* and can be made into a ring (Example 6.20).

$C[0, 1]$ is the ring of all real-valued continuous functions on the unit interval (Example 6.20).

If I is an ideal of the ring R, then R/I is the *ring of cosets, modulo I* (see Section 18.1).

$U(R)$ is the *group of units* of the ring R (Section 8.2 and Section 24.3). For a field F, we denote $U(F)$ by $F^* = F \backslash \{0\}$ (Section 24.3).

If φ is a (ring or group) homomorphism, then $\ker(\varphi)$ is its kernel (Section 17.2 and Section 33.1).

Given $a \in R$, $\langle a \rangle$ is the *principal ideal* generated by a (Section 11.4 or Section 17.3). More generally, $\langle a_1, \cdots, a_n \rangle$ is the *ideal generated by the a_i* (Exercise 12.2).

$N(R)$ is the *nilradical* of R (Exercise 7.15).

$Z(R)$ is the *center* of R (Exercise 7.12).

Groups

In what follows, assume that G is a group.

D_n is the nth *dihedral group*, the group of symmetries of the regular n-gon (Section 22.4). See Exercise 28.12 for our usual notation.

S_n is the nth *symmetric group*, the group of permutations of a set of n elements (Section 29.2). A_n is the *alternating group*, the subgroup of S_n consisting of the even permutations (Section 34.3).

Q_8 is the group of *quaternions* (Example 24.15). See Exercise 28.14 for our usual notation.

Ha and aH are right and left cosets of the subgroup H of the multiplicative group G (Section 31.1 and Section 32.1). Additive cosets look like $H + a$ or $a + H$.

G/H is the *group of cosets* for G with normal subgroup H (Section 32.2).

$[G : H]$ is the *index* of the subgroup H in G; that is, it is the number of distinct cosets H has (Section 31.2).

$\langle a \rangle$ is the *cyclic subgroup of G generated by g* (Section 26.3).

For $g \in G$, $o(g)$ is the *order* of the element g (Section 26.2).

$\operatorname{End}(G)$ is the endomorphism ring for the group G (Exercise 27.13).

$Z(G)$ is the *center* of G (Exercise 25.16).

Fields

If F is a field, $F(\alpha)$ is a *simple extension* — the smallest field containing F and α (Sections 43.2 and 43.3).

For fields $F \subseteq E$, $[E : F]$ is the *degree* of the field extension: the dimension of the F as an E-vector space (Section 44.1).

$\mathrm{Aut}(R)$ is the *automorphism group* of the ring R (Example 24.18). $\mathrm{Gal}(E|F)$ is the *Galois group of E over F* (Section 47.1).

$GF(p^n)$ is the *Galois field* with p^n elements (Section 46.1).

Index

In this index, boldfaced page numbers (e.g., **123**) indicate a definition and italicized page numbers (e.g., *123*) indicate references in an exercise. We have not included references to the Nutshells at the end of each section, or the Chapter Summaries at the end of each chapter.

Abel, Niels, 127, 316, 494, 619
abelian, 316
absolute value, 107
addition, 3, 34, 43, 74
  closed under, 3
  coordinate-wise, 504
  in a ring of cosets, 238
  matrix, 79, **85**
additive translation, 225
algebraic
  over a field, **527**
  over a ring, **535**
algebraic geometry, 83
algebraic numbers, *535*, *560*
alternating group, *see* group
annihilator, **167**
Archimedes, *497*
arithmetic, 34
  clock, 39
  in a ring, 80
ascending chain condition, **173**
associate, 120
  in a ring, **141**, 162
  polynomials, **58**
associative, **75**, 314
automorphism
  field, 585
  group, **322**
Axiom of Arithmetic, 82
axiomatic method, 9, 82

bell ringing, 436
binary operation, **73**, *83*, 314
  closed, **74**
binomial coefficient, **13**, *30*
Binomial Theorem, **13**, *87*

cancellation, 329
  additive, 80
  multiplicative, 40, 102
Cardano-Tartaglia formula, *130*, *132*, *597*, 615
  irreducible case, *132*
Cauchy, Augustin-Louis, 386
Cayley, Arthur, 386
Cayley's Theorem, 384
center
  of a group, **337**
  of a ring, **97**, *222*
centralizer, **336**
characteristic, **117**
  of a field, *235*, **531**
Chinese Remainder Theorem, 275
Classification Theorem
  for Finite Simple Groups, 445, 459
closure of an operation, 3, 74

- coefficient, **42**
- common divisor, **49**
- commutative, **75**, 316
- commutator subgroup, *see* subgroup
- commutators, *423*
- compass, 465
  - collapsible, *473*
- complete graph, **12**
- completing the square, 122, *129*
- complex conjugate, **105**, 213, 227
- complex numbers, *86*, 105, 253, *257*, *298*, *327*, 506, 542
  - argument, **107**
  - exponential form, 109
  - imaginary part, **106**
  - length, **106**
  - modulus, **106**
  - real part, **106**
- complex plane, **106**, 145
- composite, **27**
- composition
  - functional, *see* functional composition
- composition series, *459*, *460*
- congruent, **31**, 275
- conjugate
  - of a group element, **422**, **439**, *446*
- Conjugation Theorem, 439
- Constructible Number Theorem, 484, 489, 556
- constructible numbers, 467, 615, *628*
- continuous, *86*, 218
- contradiction, 9
- coordinate-wise, *86*
- coset, **230**
  - left, *409*, **414**
  - right, **399**
- Coset Theorem
  - for groups, 401
  - for rings, 238, 401
- counting numbers, **3**
- cows, 14
- cube
  - symmetries of, 306
- cubic formula, 127, *130*, 565
- cycle, *see* permutation
- Cycle Factorization Theorem, 393
- cyclic group, *see* group
- cyclic subgroup, *see* subgroup
- cyclotomic polynomial, *see* polynomial

- Dedekind, Richard, 83, 184
- degree
  - of an algebraic element, **538**
  - of field extension, **551**
- DeMoivre's Theorem, 109, 320
- depressing
  - the conic, *130*
  - the quartic, *132*
- derivative, **54**, *131*, 220, 361, *365*, 428, 571, *575*
  - rules for, *576*
- Descartes, Rene, 64, 113, *488*
- determinant, **116**, *220*, *297*, 318, 359
- diagonal, *see* matrices
- differentiable, *86*
- dimension
  - of a vector space, 545
- Diophantine equation, **145**, *152*
- direct product
  - of groups, **362**, 372
  - of rings, **77**, *87*, 272
- discriminant
  - of the conic, *130*

  - of the quadratic, **122**, 142
  - of the quintic, 611
- *Disquisitiones Arithmeticae*, 38
- distributes, **75**
- divides, **17**, **47**
- division, 4, 105
- Division Theorem, 197
  - for $\mathbb{Q}[x]$, 45
  - for $F[x]$, 120
  - for Euclidean domains, **197**
  - for Integers, 15
- domain, *see* integral domain
- doubling the cube, 466, 489, 557

- Eisenstein's Criterion, 62, *66*
- *The Elements*, 10, 18, 22, 197, 465, *472*
- embedding, **362**, 428
- endomorphism, **365**
- equivalence relation, *235*
- Euclid, 10, 25
- Euclid's Algorithm, **18**, 19, 25, 50, 110, *117*, 277
  - for $F[x]$, 120
  - for Euclidean domains, *206*
  - recursive version, *30*
- Euclidean domain, **197**
- Euclidean valuation, **197**
- Euler, Leonhard, 26, 38
- Euler's phi function, **410**
- Euler's Theorem, *410*
- existence and uniqueness theorem, 15
- exponential function, 368
- exponents
  - negative, **342**
  - positive, 339
  - rule of, 342
- extension problem for groups, 445, 455, 459

- factor, **17**, **47**
- factorial, **12**
- Factorization Theorem for Permutations, 392
- Factorization Theorem for PIDs, 171
- Factorization Theorem for Quadratic Extensions, 155
- Feit, Walter, 459
- Feit-Thompson Theorem, 459
- Fermat, Pierre de, 38, 111, 113
- Fermat's Last Theorem, 113
- Fermat's Little Theorem, 111, *116*, *410*
- Ferrari formula, *132*, 615
- Fibonacci sequence, **7**, *13*, *28*
- field, **105**, 260
  - algebraically closed, **133**, *576*
  - base, **527**
  - finite, 110, *117*, *235*, 530, 542, 578
  - fixed, **600**
  - Galois, **579**
  - of algebraic numbers, **535**
  - of constructible numbers, 468
  - splitting, **566**, 573, 579
- field extension, **475**, **527**
  - algebraic, **551**
  - algebraic simple, **539**, 554, 572
  - by radicals, **616**
  - degree of, **551**
  - finite, **551**
  - normal, **570**, 573, 589, 602, 606
  - proper, **475**
  - quadratic, **475**
  - radical, **615**
  - simple, **539**

  splitting, **535**, 566
  transcendental simple, 544, 557
finite subdirect product, **279**
fix, **391**
fixed, *647*
flatlanders, *309*
flip, 288, 427
floor function, 30
Frobenius isomorphism, *235*, *547*, *582*
function
  domain of, 211
  identity, 380
  inverse, *see* inverse function
  one-to-one, **214**, 231
  onto, **214**, 231
  range of, 211
functional composition, 288, 314, 331, 380
Fundamental Isomorphism Theorem
  for Groups, 419, 429
  for Rings, 250, 399
Fundamental Theorem for Finite Abelian Groups, 449
Fundamental Theorem of Algebra, 123, 127, *534*, 569
Fundamental Theorem of Arithmetic, 22, 23, 25, 185, 392
Fundamental Theorem of Galois Theory, 601, 604, 606

Gallian, Joseph, 405
Galois field, *see* field
Galois group, *see* group
Galois, Evariste, 386, 495, 579
Gauss's Lemma, 60, 143, *655*
Gauss, Karl, 26, 38, 64, 123, 127
Gaussian integers, **86**, 90, 144, *152*, *195*, 199, 202, 315
gcd, *see* greatest common divisor
GCD identity, 184
  for $F[x]$, 120
  for Euclidean domains, *205*
  for integers, 20, 182
  for polynomials, 52
geometric progression, *12*
Golden Section, *472*, *473*
greatest common divisor, **17**, 21, **49**, 50, *205*
group, **314**
  p-group, **452**
  abelian, **316**
  additive, 331, 358
  alternating, **438**, 442
  cyclic, **347**, 369
  dihedral, 291, 327, *336*, *378*, 420, *434*
  Galois, **586**
  generators of, *309*
  Klein Four, 407
  multiplicative, 331
  of automorphisms, 322
  of cosets, **416**
  of symmetries, 291, 300, 371, 438
  of units, 317
  permutation, 381, 588
  quotient, **416**
  simple, **443**, 459, *460*
  solvable, 445, **455**, 622
  sporadic, 445
  symmetric, **381**, 442
  table, 291

Hölder, Otto, 459
Hadlock, Charles, 493
Hermite, Charles, 493, 528
Hilbert, David, 83
homomorphism
  group, **357**
  identity, **216**, 247
  inclusion, **216**, 248
  natural, **241**, **429**
  projection, **213**, 227
  ring, **212**, 358
  zero, **216**, 227
Huygens, Christiaan, 113

ideal, **158**, 226
  as kernel, 228
  finitely generated, **174**
  generators of, **160**, **174**
  improper, **159**
  maximal, **180**, 259, 260, *268*
  not principal, 169, *174*
  prime, **262**, 263
  principal, **160**, 181, 228, 346
  proper, **159**
  trivial, **159**
  zero, **159**
ideals
  ascending chain, 172, *176*
  intersection of, *166*
  product of, **167**, 269
  sum of, **167**
idempotent, **87**, *99*, *168*, *235*, *244*
identity
  additive, **36**, **75**, 78
  group, 314
  multiplicative, **37**, **93**
  uniqueness, 80, 329
imaginary numbers, *118*, 127, *335*
implication, 7
inclusive or, 22
Index 2 Theorem, 419, 439
index of a subgroup, *see* subgroup
induction, *see* mathematical induction
induction hypothesis, **7**
inductive definition, 8
inductive reasoning, 7
infinite
  countably, *257*
  uncountably, *257*
injective, **214**
integers, **3**, 75
  Gaussian, *see* Gaussian integers
integral domain, **102**, 105, 112, 162, 263
Internal Characterization Theorem, 373
invariant, 518
inverse
  additive, **36**, **75**
  group, 314
  matrix, *116*
  multiplicative, **37**, **103**
    left, **117**
    right, **117**
  uniqueness, 80, 329, 508
inverse function, 249, 314, 322, 367, 381
irrational numbers, *28*, 56, 64, *66*
irreducible, 141, 149
  in a ring, **142**, 178, 184
  integer, **22**
  polynomial, **55**, 57, 120, 124, 193
isomorphism
  for groups, **367**
  for rings, **247**

- Jacobson, Nathan, *627*
- Jordan, Camille, 254, 459
- journals
  - mathematical, 113, 459

- kernel, **226**, 228, 241, **425**
  - trivial, 232, 429
- Klein Four group, *see* group
- Klein, Felix, 219
- Kronecker, Leopold, 528
- Kronecker's Theorem, 528
- Kummer, Ernst, 83, 183

- Lagrange, Joseph, 386, 403
- Lagrange's Theorem, *235*, 403, 451
- Lambert, Johann, 127, 493
- Lambert's Theorem, 342
- lcm, *see* least common multiple
- least common multiple, **28**
- Legendre, Adrien-Marie, 26, 127
- Lindemann, Ferdinand, 493, 528
- Lindemann's Theorem, 493, 557
- linear algebra, 79, 503
- linear combination, **20**, 169, *174*, 511
- logarithm, 109, 359, *365*, 368

- mathematical induction, **5**, 198
  - base case, 5
  - bootstrap, 5
- matrices, 79
  - diagonal, **92**
- matrix
  - addition, *see* addition
  - multiplication, *see* multiplication
- maximal independent set, 517, *525*
- McKay, James, 405
- measures, **25**
- minimal spanning set, 516
- mirror reflection, 310
- modulo, **31**, 237
- modulus, **31**
- Moore, E.H., 579
- multiplication, 3, 34, 43, 74
  - in a ring of cosets, 238
  - matrix, 79, **85**, 293
  - scalar, 505
  - table, 289, 291
- multiplicative absorption property, **159**, 226

- natural numbers, **3**, 78
- nilpotent, **98**, *115*, *118*
- nilradical, **98**, *167*, *222*, *244*
- Noether, Emmy, 172
- Noetherian ring, *see* ring
- norm, **146**
- numbers, *see* complex, constructible, counting, irrational, natural, rational, real, transcendental

- orbit, **394**
  - trivial, 394
- order
  - of a group, **402**, 404
  - of an element, 340, 404
  - torsion, **340**
  - torsion-free, **340**, *423*
- orientation
  - left-hand, 299
  - right-hand, 299

- painting, 304, 446
- parabola
  - segment of, *497*
- partition, **33**, 230, *234*, 406
- Pell's equations, 147
- pentagon, *296*, *472*
- permutation, 290, *325*, 372, **379**
  - cycle, **390**
  - disjoint, **392**
  - even, **437**
  - odd, **437**
  - order-preserving, **325**
  - parity of, 437
  - transposition, 435
- pid, *see* principal ideal domain
- plane of F, **480**
  - circle in, **480**
  - line in, **480**
- point-wise, 79
- point-wise operation, 77
- polygon, *297*, *496*
- polynomial
  - content of, **191**
  - cyclotomic, *54*, **67**, *133*, 591, *612*
  - degree of, **43**
  - equality of, **42**
  - function, **43**, *129*
  - in $\mathbb{Q}[x]$, **42**, 76
  - in $F[x]$, 119
  - irreducible, *see* irreducible
  - minimal, **538**
  - monic, **195**, 538
  - over an arbitrary ring, *88*
  - prime, **57**
  - primitive, **143**, 191
  - reducible, **56**
  - resolvent, *626*
  - symmetric, **602**
- power set, **88**, *95*, *167*, *174*, *221*, *234*, *256*, *268*
- powers
  - of a group element, 339
- pre-image, 225, 425
- preserves addition, 212
- preserves multiplication, 145, 212
- prime
  - in a ring, **177**, 184
  - integer, **22**, 23
  - polynomial, **57**, 120, 193
- prime subfield, **532**, 577
- Primitive Root Theorem, 580
- principal ideal domain, **171**, *176*, 184, 202
- Principle of Mathematical Induction, 5, 9
  - Strong, *14*
- projection, **220**, *234*, 264, 362, 428
- Pythagorean Theorem, *472*

- Qin Jiushao, 277
- quadratic extension of the integers, **144**, 155, *175*
- quadratic formula, **121**, 125, *129*, 484, 565, 615
- quartic formula, 127, *132*, 565
- quaternions, **321**, *325*, *378*, *388*, *408*, *410*, *422*
- quotient, 15, 197

- rational numbers, 4, 25, 75, *257*, *326*
- rational plane, **479**
- Rational Root Theorem, 62, *65*
- real numbers, *257*, *326*
- recursive definition, 8
- reductio ad absurdum, 9
- reflection, 288
- reflexivity, 81
- relatively prime, **27**
- remainder, 15, 197

representation theory for groups, 382
residue, **31**
residue class, **31**
rhombus, *296*
rigid motion, **287**, 299
ring, **74**
  additive group of, 315
  Boolean, **269**
  commutative, **78**
  endomorphism, **366**
  Noetherian, **173**, *174*
  of cosets, **240**
  quotient, **240**
  zero, **76**, 94, 322
root, **48**
  repeated, **53**
root of unity, **133**, 619
  primitive, *131*, **133**
Root Theorem, 48
  for $F[x]$, 120
root tower, **616**, 618
Royal Society, 113

scalar, **43**
sequences, **87**, *99*, *175*, *221*, *223*, *244*
  bounded, **175**
  convergent, **223**
series
  MacLaurin, *118*
  Taylor, *118*
socks and shoes, 104, 329
solution of equations, 80
  in a group, 329
solvable by radicals, **616**, 622, 624
solvable group, *see* group
splits, **566**
splitting extension, *see* field extension
splitting field, *see* field
square, 294
  symmetries of, 295
square root, 125, 469
square-free integer, **144**, *454*
squaring the circle, 466, 493, 557
straightedge, 465
structure theorem, 445
subgroup, **331**
  commutator, *423*
  conjugate, **421**, *448*
  cyclic, **345**
  improper, 332
  index of, **402**
  normal, **415**, 425, 439
  proper, 332
  stabilizer, **388**, *397*, *410*
  torsion, **422**
  transitive, 447, *613*
  trivial, 332
Subgroup Theorem, 333
subgroups
  intersection of, *335*, *421*
  product of, *422*
subnormal series, **455**
subring, **89**, *see* ideal
  improper, **90**
  proper, **90**
  trivial, **90**
Subring Theorem, 90
subrings
  intersection of, *97*
Substitution Rule for Equality, 81
subtraction, 3, 43, 81, *85*, 331
Sun Tsu, 275, 277
support, **391**
surjective, **214**
Sylow theorems, 405
symmetric, 367
symmetric difference, **88**
symmetry, 81, **287**, 371

tangent line to a circle, 484, *488*
Taylor, Richard, 114
tetrahedron symmetries of, 303
Thompson, John, 459
torsion, *see* order
torsion-free, *see* order
transcendental numbers, 493, 544
transcendental over a field, **527**
transitivity, 81
transposition, *see* permutation
triangle
  equilateral, 287
  isosceles, *296*
  scalene, *296*
  symmetries of, 288
triangle inequality, *11*
trisecting the angle, 466, 490, 557

UFD, *see* unique factorization domain
unique factorization
  for integers, 24
  for polynomials, 58, 120
unique factorization domain, **179**, 185, 189, 193, 202
Unique Factorization Theorem
  for Integers, 24
Unique Factorization Theorem for Polynomials, 58, 185
unit, **103**, 141, 317
unit circle, 115, 319, 332, *411*, *432*
unit line segment, 467
unity, **93**, *98*

vector, 504
  addition, **504**
  additive inverse, 504
  linear combination, **511**
  linearly dependent set, **515**, 516
  linearly independent set, **514**
  spanning set, **511**
  spans, 512
  zero, **504**
vector space, **505**
  basis, **516**, 517
  dimension, **521**
  finite dimensional, **514**
  subspace, **524**
Viète, Francois, 64

Wantzel, Pierre, 494
well defined, 35, 237, 239, 251, 413, 430
well ordered, **4**
Well-ordering Principle, **4**, 9, 198
Wiles, Andrew, 114
Wilson's Theorem, *118*

zero, **36**, 75
zero divisor, **101**, 105
zero ring, *see* ring