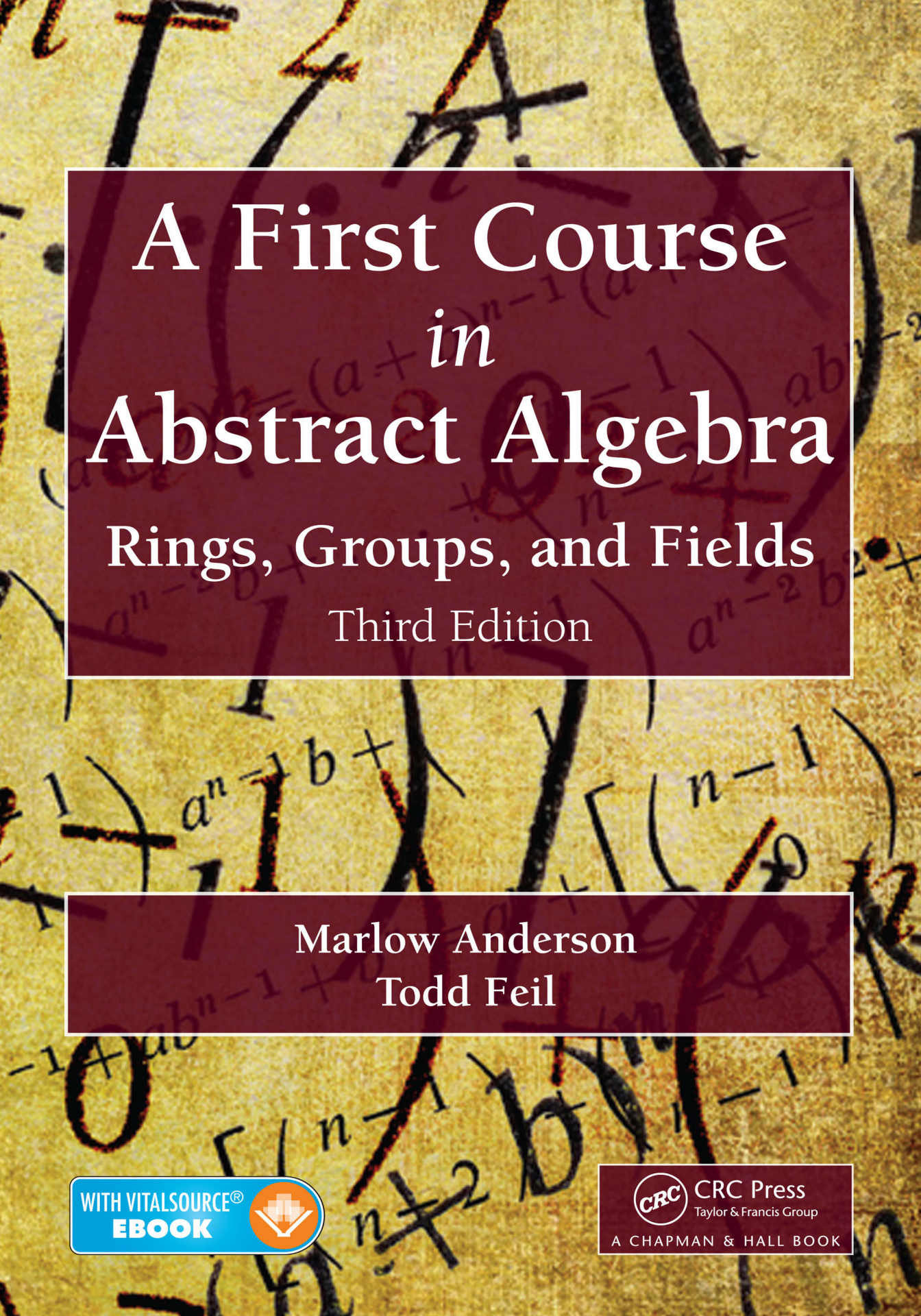

A First Course
in
Abstract Algebra
Rings, Groups, and Fields
Third Edition
Marlow Anderson
Todd Feil
WITH VITALSOURCE®
EBOOK
CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Accessing the E-book edition

Using the VitalSource® ebook

Access to the VitalBook™ ebook accompanying this book is via VitalSource® Bookshelf – an ebook reader which allows you to make and share notes and highlights on your ebooks and search across all of the ebooks that you hold on your VitalSource Bookshelf. You can access the ebook online or offline on your smartphone, tablet or PC/Mac and your notes and highlights will automatically stay in sync no matter where you make them.

1. **Create a VitalSource Bookshelf account at** ***https://online.vitalsource.com/user/new*** or log into your existing account if you already have one.
2. **Redeem the code provided in the panel below to get online access to the ebook.** Log in to Bookshelf and click the **Account** menu at the top right of the screen. Select **Redeem** and enter the redemption code shown on the scratch-off panel below in the **Code To Redeem** box. Press **Redeem**. Once the code has been redeemed your ebook will download and appear in your library.

DOWNLOAD AND READ OFFLINE

To use your ebook offline, download BookShelf to your PC, Mac, iOS device, Android device or Kindle Fire, and log in to your Bookshelf account to access your ebook:

On your PC/Mac

Go to ***http://bookshelf.vitalsource.com/*** and follow the instructions to download the free **VitalSource Bookshelf** app to your PC or Mac and log into your Bookshelf account.

On your iPhone/iPod Touch/iPad

Download the free **VitalSource Bookshelf** App available via the iTunes App Store and log into your Bookshelf account. You can find more information at ***https://support.vitalsource.com/hc/en-us/categories/200134217-Bookshelf-for-iOS***

On your Android™ smartphone or tablet

Download the free **VitalSource Bookshelf** App available via Google Play and log into your Bookshelf account. You can find more information at ***https://support.vitalsource.com/hc/en-us/categories/200139976-Bookshelf-for-Android-and-Kindle-Fire***

On your Kindle Fire

Download the free **VitalSource Bookshelf** App available from Amazon and log into your Bookshelf account. You can find more information at ***https://support.vitalsource.com/hc/en-us/categories/200139976-Bookshelf-for-Android-and-Kindle-Fire***

N.B. The code in the scratch-off panel can only be used once. When you have created a Bookshelf account and redeemed the code you will be able to access the ebook online or offline on your smartphone, tablet or PC/Mac.

SUPPORT

If you have any questions about downloading Bookshelf, creating your account, or accessing and using your ebook edition, please visit ***http://support.vitalsource.com/***

A First Course
in
Abstract Algebra
Rings, Groups, and Fields
Third Edition

This page intentionally left blank

A First Course *in* Abstract Algebra

Rings, Groups, and Fields

Third Edition

Marlow Anderson
Colorado College
Colorado Springs, USA

Todd Feil
Denison University
Granville, Ohio, USA

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20141001

International Standard Book Number-13: 978-1-4822-4553-0 (eBook - PDF)

Contents

Preface

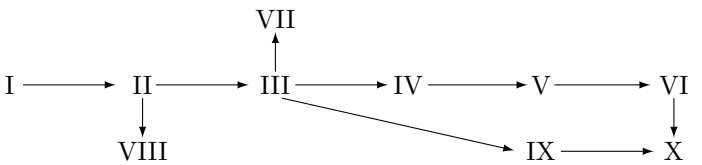
Traditionally, a first course in abstract algebra introduces groups, rings, and fields, in that order. In contrast, we have chosen to develop ring theory first, in order to draw upon the student's familiarity with integers and with polynomials, which we use as the motivating examples for studying rings.

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 and permutations. 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.

This is the third edition of this text and significant changes have been made to the second edition. Both comments from adopters and our own experiences have prompted this. The biggest change is in moving the section on Unique Factorization (now Section VII) further back in the book reflecting that for many teaching a first course, this topic is optional. Doing so necessitated reorganizing the introduction of ideals. Now Sections I, II, and III form the core material on rings, integral domains, and fields.

Sections IV and V contain the basic group theory material. We have compressed the motivating examples of symmetries of the plane and of space into one chapter, following it with a more detailed treatment of permutations, our other motivating example for abstract groups. Section VI introduces more topics in group theory including new chapters on the Sylow theorems. Sections VIII, IX, and X remain sections on Constructibility, Vector Spaces and Field Extensions, and Galois Theory; the latter contains much edited material and many new exercises.

The diagram below roughly indicates the dependency of the large sections.



Descriptions of 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. An introduction to complex numbers is given. We also introduce ideals in this section.

Section III (*Ring Homomorphisms and Ideals*) has as its main goal the proof of the Fundamental Isomorphism Theorem. Section III 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 IV (*Groups*) begins with two chapters on concrete examples motivating abstract groups: symmetries of geometric figures (in the plane and in space) and permutations. We then define abstract groups and group isomorphisms and consider subgroups and cyclic groups.

Section V (*Group Homomorphisms*) defines group homomorphisms with one of its goals the Fundamental Isomorphism Theorem for groups. Cayley's and Lagrange's theorems are presented in this section.

Section VI (*Topics from Group Theory*) explores three topics: the alternating group, the Sylow theorems, and solvable groups. The latter is needed in Section X. We use groups acting on sets to prove the Sylow theorems, which provides not only a very accessible method of proof but also experience with permutations from a slightly different perspective.

In Section VII (*Unique Factorization*) we explore more general contexts in which unique factorization is possible. Chapter 33 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 VIII (*Constructibility Problems*) 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 III, if the instructor wishes to delay the introduction of groups.

We revisit the impossibility proofs in Section IX (*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 X develops *Galois Theory* with the goal of showing the impossibility of solving the quintic with radicals. This section depends heavily on Section IX, as well as Chapters 27 and 30.

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 odd-numbered exercises that do not require proof are given in the Hints and Answers section. Hints, of varying depth, to odd-numbered proof 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.

Notes for the Instructor

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 suggested track through the text for a one-semester course is to tackle Sections I through V (except possibly Chapter 16). We consider these chapters as the core ideas in the text. If time permits, one could include topics from Section VI, VII, VIII, or a combination, as to your taste.

A second semester could pick up wherever the first semester left off with the goal of completing Section X on Galois Theory, or else finish the remaining topics in the text.

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 Chapters 8 and 10 should be quite adequate.

Instructors may contact the publisher for an Instructor's Manual which contains answers to all of the exercises, including in-depth outlines for proof problems.

Acknowledgments

Each of us has spent over thirty-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 Maxfields' 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. We're grateful to our respective institutions (The Colorado College and Denison University), which provided financial support to this project at various crucial stages. We're also grateful to the readers of the first two editions, who have made helpful suggestions, and pointed out errors both typographical and mathematical.

A big thank you goes to Sunil Nair from Chapman & Hall/CRC for encouraging us to undertake this third edition. His support of this project is very much appreciated.

Most of all, we thank Audrey and Robin, for their moral support that was, as always, indispensable.

We take the only reasonable position regarding errors that may remain in the text, both typographical and otherwise, and blame each other.

Marlow Anderson
Mathematics Department
The Colorado College
Colorado Springs, CO
manderson@coloradocollege.edu

Todd Feil
Department of Mathematics and Computer Science
Denison University
Granville, OH
feil@denison.edu

Part I

Numbers, Polynomials, and Factoring

This page intentionally left blank

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 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:

$$\begin{array}{c} \text{Is } 1 \in X? \\ \text{Is } 2 \in X? \\ \vdots \end{array}$$

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 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 positive 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,

$$\begin{aligned} 1 + 3 + 5 + \cdots + (2(n-1)-1) + (2n-1) &= \\ (n-1)^2 + (2n-1) &= n^2, \end{aligned}$$

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 the 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 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 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 $n \geq 4$.

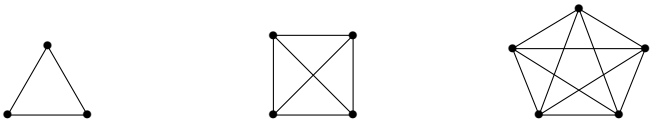
6. Prove that for all positive integers n,
$$1^3 + 2^3 + \cdots + n^3 = \left(\frac{n(n+1)}{2}\right)^2.$$

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) Use mathematical induction to 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.

This page intentionally left blank

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:

$$\begin{aligned} b_0 &= a_0q_0 + r_0 \\ b_1 &= a_1q_1 + r_1 \\ &\vdots \\ b_{n-1} &= a_{n-1}q_{n-1} + r_{n-1} \\ b_n &= a_nq_n + 0 \end{aligned}$$

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 $p|a_1a_2\cdots a_n$, then $p|a_i$ for some i.*

Proof: This is Exercise 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 $x = a_1 a_2 \cdots a_n = b_1 b_2 \cdots b_m$ 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 $i = 1, 2, \cdots, n$.

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

$$a_2 a_3 \cdots a_n = \pm b_2 \cdots b_m.$$

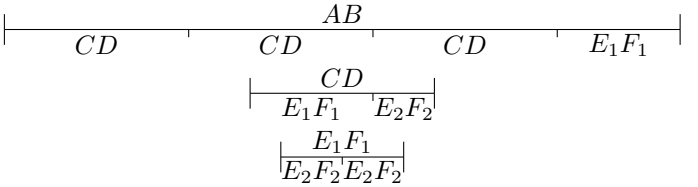
(Because b_2 is irreducible, so is $-b_2$, and we consider $\pm b_2$ as an irreducible factor.) We now have two factorizations into irreducibles, and the number of a_i factors is $n - 1$. So by the induction hypothesis $n - 1 = m - 1$, and by renumbering the b_j as necessary, $a_i = \pm b_i$ for $i = 1, 2, \cdots, n$. 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 14 and 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.
   (c) Do the same for 2007 and 203.
   (d) Do the same for 3604 and 4770.

2. (a) Prove that $\gcd(a, b)$ divides $a - b$. This sometimes provides a short cut in finding gcds.
   (b) Use this to find $\gcd(1962, 1965)$.
   (c) Now find $\gcd(1961, 1965)$.
   (d) Find the gcds in Exercise 1 using this short cut.

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

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

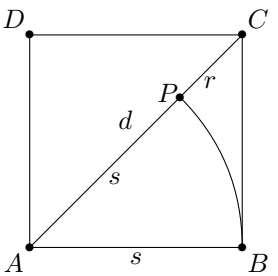
13. Prove that if m is a common multiple of both a and b, then $\text{lcm}(a, b)|m$.

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

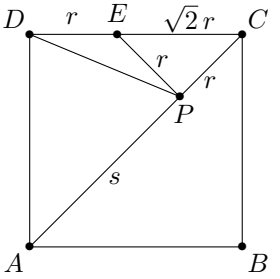
15. This problem outlines another proof that $\sqrt{2}$ is irrational. We show that Euclid's Algorithm never halts if applied to a diagonal d and side s of a square. The first step of the algorithm yields

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

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

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

Conclude that this algorithm never halts, and so there is no common measure for the diagonal and side of the square.

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 mod b is the remainder when dividing a by b.)

```
function gcd(a, b);
  if b = 0 then gcd ← a else gcd ← gcd(b, a mod 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 ← a; x ← 1; y ← 0;
  else
    extgcd(b, a mod b, g, x, y);
    temp ← y;
    y ← x − ⌊a/b⌋y;
    x ← 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 1.14 the definition of the binomial coefficient $\binom{n}{k}$. Suppose that p is a positive prime integer, and k is an integer with $1 \leq k \leq p-1$. 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 $a \equiv b \,(\mathrm{mod}\ m)$. 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 $x \equiv y \,(\text{mod } m)$. 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:

$$\begin{aligned} [a]_m + [b]_m &= [a+b]_m \\ [a]_m \cdot [b]_m &= [a \cdot b]_m. \end{aligned}$$

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 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.'

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

$\cdot$	$[0]$	$[1]$
$[0]$	$[0]$	$[0]$
$[1]$	$[0]$	$[1]$

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. (a) 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 $\mathbb{Z}_6$?
   (b) Repeat this in $\mathbb{Z}_{10}$.
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.

This page intentionally left blank

Chapter 4

Polynomials with Rational Coefficients

In Chapter 2 we proved that every integer ($\neq 0, \pm 1$) 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 ($\neq \pm 1$) 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 ($\neq \pm 1$) 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

$$f = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n + \cdots$$

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 $f = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$. 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

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

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 $f = a_n x^n + f_1$, where $a_n \neq 0$ and $\deg(f_1) < n$. Similarly, $g = b_m x^m + g_1$, where $b_m \neq 0$ and $\deg(g_1) < m$. By the definition of multiplication of polynomials, the coefficient on x^{n+m} is $a_n b_m$, 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 $g = x^4 + x^2 - x + 2$.

$$\begin{array}{r|l} & x^2 - 2x + 6 \\ \hline x^2 + 2x - 1 & x^4 + x^2 - x + 2 \\ & \underline{x^4 + 2x^3 - x^2} \\ & -2x^3 + 2x^2 - x + 2 \\ & \underline{-2x^3 - 4x^2 + 2x} \\ & 6x^2 - 3x + 2 \\ & \underline{6x^2 + 12x - 6} \\ & -15x + 8 \end{array}$$

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 $f = a_0 + a_1 x + \cdots + a_n x^n$ and $g = b_0 + b_1 x + \cdots + b_m x^m$, 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 nonzero 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 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 $f=(x-a)q+r$ 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.