A GENTLE INTRODUCTION TO ABSTRACT ALGEBRA

by

B.A. Sethuraman

California State University Northridge

Source files for this book are available at

`<http://www.csun.edu/~asethura/giaa/>`

Contents

Preface

This book is a gentle introduction to abstract algebra. It is ideal as a text for a one-semester course designed to provide a first exposure of the subject to students in mathematics, science, or engineering. Such a course would teach students the basic objects of algebra, providing plentiful examples and enough theory to allow interested students to transition easily to more advanced abstract algebra. At the same time, this course would allow future users of the subject, including students interested in various other subfields of mathematics and students of science and engineering, to gain enough familiarity with the objects of algebra to be able to study them further within the manifold contexts in which they are needed.

Thus, this book deals with groups, rings and fields, and vector spaces. The approach to these objects is elementary, with a focus on examples and on computation with these examples. The book starts with rings, reflecting my experience that students find rings easier to grasp as an abstraction since they are already familiar with the integers, the rationals, the reals, the complexes, 2×2 matrices with real entries, and polynomials with real coefficients. Vector spaces are treated next, followed by groups. It is expected that students have had some exposure to proof-based mathematics, such as can be obtained in basic "proofs" courses common in many American universities. Such students are likely to be familiar with the properties of the integers already, but for completeness, a preliminary chapter on divisibility in the integers has been included. Material on sets, functions, and relations, that belong more commonly to a "proofs" course, has also been provided as an appendix.

The style of the book is conversational (a style that mirrors my own approach to teaching), with a stress on exposition. I have attempted to show that there are some common themes to the study of the three objects: rings, vector spaces, and groups. For each, I introduce the object using a large number of examples. For each, I introduce their various subobjects (subrings, ideals, subspaces, subgroups, normal subgroups), again with numerous examples. I introduce quotient objects, and then for each object I introduce the appropriate notion of homomorphism and isomorphism. I end with the fundamental homomorphism theorem for each object. I find that when students see the same concept three different times in mildly different guises, such as the notion of a structure preserving map, the notion of a kernel, or the notion of an appropriate quotient object, they become quite comfortable with these concepts by the end of the semester. For example, I find that they have no trouble with quotient groups (a traditionally difficult idea to convey if abstract algebra is introduced first through groups) since they have already computed with quotient rings in more intuitive settings such as the integers mod n or the polynomials over a field mod a linear or quadratic polynomial.

The entire material in the book can be covered in a traditional sixteen week semester, judiciously speeding up here and there. Besides copious examples and exercises (most of a computational kind, based on the examples, and some that extend the theory developed in the text), each chapter comes with end notes: remarks about various aspects of the theory, occasional hints to some exercise, and several glimpses into material beyond the course. The book shares some material with an earlier text I wrote called Rings, Fields and Vector Spaces, but the focus and end goal of the two books are quite different.

I am grateful to the various faculty members at California State Univer-

sity Northridge who have taught the introductory abstract algebra course, Math 360, for several years now from this book. I am also grateful to the students in the course; together, both the faculty and students have provided valuable feedback. The National Science Foundation has supported me professionally through two research grants during much of the time when this book was being developed, and I am grateful to them.

I owe a special debt of gratitude to the most extraordinary student I have ever worked with, one whom I have never met. He will remain unnamed here. He is currently in prison, but rather than succumb to circumstances, he chose the positive route, and enrolled in mathematics courses at California State University Northridge as an extension student. Faculty would send him course material by U.S. mail (the only form of interaction he is allowed under incarceration), and he would complete his assignments under supervision and mail them back. He offered to read through this book and give suggestions, an offer I readily accepted. I was amazed when I received his edit suggestions! I have yet to see such meticulousness in any student, such attention to the right word, such alertness for the clumsy phrase. But more importantly, he proved to be a brilliant student, and made several powerful suggestions that went beyond the writing and into the mathematics. There are many explanations here and many additional remarks that owe their existence to him. (All errors that remain, of course, are to be blamed on me.) I was privileged that he learned abstract algebra from this book, and to him I would like to say: Thank you, my friend! I hope to meet you some day.

B.A. Sethuraman
California State University Northridge

To the Student: How to Read a Mathematics Book

How should you read a mathematics book? The answer, which applies to every book on mathematics, and in particular to this one, can be given in one word—*actively.* You may have heard this before, but it can never be overstressed—*you can only learn mathematics by doing mathematics.* This means much more than attempting all the problems assigned to you (although attempting every problem assigned to you is a *must*). What it means is that you should take time out to think through every sentence and confirm every assertion made. You should accept nothing on trust; instead, not only should you check every statement, you should also attempt to go beyond what is stated, searching for patterns, looking for connections with other material that you may have studied, and probing for possible generalizations.

Let us consider an example. On page 29 in Chapter 2, you will find the following sentence:

> Yet, even in this extremely familiar number system, multiplication is not commutative; for instance,

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \neq \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

(The "number system" referred to is the set of 2×2 matrices whose entries are real numbers.) When you read a sentence such as this, the first thing that you should do is *verify the computation yourselves.* Mathematical insight comes from mathematical experience, and you cannot expect to gain mathematical experience if you merely accept somebody else's word that the product on the left side of the equation does not equal the product on the right side.

The very process of multiplying out these matrices will make the set of 2×2 matrices a more familiar system of objects, but as you do the calculations, more things can happen if you keep your eyes and ears open. Some or all of the following may occur:

1. You may notice that not only are the two products not the same, but that the product on the right side gives you the *zero* matrix. This should make you realize that although it may seem impossible that two nonzero "numbers" can multiply out to zero, this is only because you are confining your thinking to the real or complex numbers. Already, the set of 2×2 matrices (with which you have at least some familiarity) contains nonzero elements whose product is zero.

2. Intrigued by this, you may want to discover other pairs of nonzero matrices that multiply out to zero. You will do this by taking arbitrary pairs of matrices and determining their product. It is quite probable that you will not find an appropriate pair. At this point you may be tempted to give up. However, you should not. You should try to be creative, and study how the entries in the various pairs of matrices you have selected affect the product. It may be possible for you to change one or two entries in such a way that the product comes out to be zero. For instance, suppose you consider the product
$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 4 & 0 \\ 2 & 0 \end{pmatrix} = \begin{pmatrix} 6 & 0 \\ 6 & 0 \end{pmatrix}$$
You should observe that no matter what the entries of the first matrix are, the product will always have zeros in the $(1, 2)$ and the $(2, 2)$ slots.

This gives you some freedom to try to adjust the entries of the first matrix so that the $(1,1)$ and the $(2,1)$ slots also come out to be zero. After some experimentation, you should be able to do this.

3. You may notice a pattern in the two matrices that appear in our inequality on page ix. Both matrices have only one nonzero entry, and that entry is a 1. Of course, the 1 occurs in different slots in the two matrices. You may wonder what sorts of products occur if you take similar pairs of matrices, but with the nonzero 1 occuring at other locations. To settle your curiosity, you will multiply out pairs of such matrices, such as
$$\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},$$
or
$$\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$
You will try to discern a pattern behind how such matrices multiply. To help you describe this pattern, you will let $e_{i,j}$ stand for the matrix with 1 in the (i,j)-th slot and zeros everywhere else, and you will try to discover a formula for the product of $e_{i,j}$ and $e_{k,l}$, where i, j, k, and l can each be any element of the set $\{1,2\}$.

4. You may wonder whether the fact that we considered only 2×2 matrices is significant when considering noncommutative multiplication or when considering the phenomenon of two nonzero elements that multiply out to zero. You will ask yourselves whether the same phenomena occur in the set of 3×3 matrices or 4×4 matrices. You will next ask yourselves whether they occur in the set of $n \times n$ matrices, where n is arbitrary. But you will caution yourselves about letting n be too arbitrary. Clearly n needs to be a positive integer, since "$n \times n$ matrices" is meaningless otherwise, but you will wonder whether n can be allowed to equal 1 if you want such phenomena to occur.

5. You may combine 3 and 4 above, and try to define the matrices $e_{i,j}$ analogously in the general context of $n \times n$ matrices. You will study the product of such matrices in this general context and try to discover a formula for their product.

Notice that a single sentence can lead to an enormous amount of mathematical activity! Every step requires you to be alert and actively involved in what you are doing. You observe patterns for yourselves, you ask yourselves questions, and you try to answer these questions on your own. In the process, you discover most of the mathematics yourselves. This is really the only way to learn mathematics (and in particular, it is the way every professional mathematician has learned the subject). Mathematical concepts are developed precisely because mathematicians observe patterns in various mathematical objects (such as the 2×2 matrices), and to have a good understanding of these concepts you must try to notice these patterns for yourselves.

May you spend many many hours happily playing in the rich and beautiful world of mathematics!

Exercises

1. Carry out the program in steps (1) through (5) above.

Chapter 1

Divisibility in the Integers

We will begin our study with a very concrete set of objects, the integers, that is, the set $\{0, 1, -1, 2, -2, \dots\}$. This set is traditionally denoted $\mathbb{Z}$ and is very familiar to us—in fact, we were introduced to this set so early in our lives that we think of ourselves as having grown up with the integers. Moreover, we view ourselves as having completely absorbed the process of integer division; we unhesitatingly say that 3 divides 99 and equally unhesitatingly say that 5 does *not* divide 101.

As it turns out, this very familiar set of objects has an immense amount of structure to it. It turns out, for instance, that there are certain distinguished integers (the primes) that serve as building blocks for all other integers. These primes are rather beguiling objects; their existence has been known for over two thousand years, yet there are still several unanswered questions about them. They serve as building blocks in the following sense: every positive integer greater than 1 can be expressed uniquely as a product of primes. (Negative integers less than -1 also factor into a product of primes, except that they have a minus sign in front of the product.)

The fact that nearly every integer breaks up uniquely into building blocks is an amazing one; this is a property that holds in very few number systems, and our goal in this chapter is to establish this fact. (In the exercises to

Chapter 2 we will see an example of a number system whose elements *do not* factor uniquely into building blocks. Chapter 2 will also contain a discussion of what a "number system" is—see Remark 2.8.)

We will begin by examining the notion of divisibility and defining divisors and multiples. We will study the division algorithm and how it follows from the Well-Ordering Principle. We will explore greatest common divisors and the notion of relative primeness. We will then introduce primes and prove our factorization theorem. Finally, we will look at what is widely considered as the ultimate illustration of the elegance of pure mathematics—Euclid's proof that there are infinitely many primes.

Let us start with something that *seems* very innocuous, but is actually rather profound. Write $\mathbb{N}$ for the set of nonnegative integers that is, $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$. ($\mathbb{N}$ stands for "natural numbers," as the nonnegative integers are sometimes referred to.) Let S be any *nonempty* subset of $\mathbb{N}$. For example, S could be the set $\{0, 5, 10, 15, \ldots\}$, or the set $\{1, 4, 9, 16, \ldots\}$, or else the set $\{100, 1000\}$. The following is rather obvious: there is an element in S that is smaller than every other element in S, that is, S has a *smallest* or *least* element. This fact, namely that every nonempty subset of $\mathbb{N}$ has a least element, turns out to be a *crucial* reason why the integers possess all the other beautiful properties (such as a notion of divisibility, and the existence of prime factorizations) that make them so interesting.

Contrast the integers with another very familiar number system, the rationals, that is, the set $\{a/b \mid a \text{ and } b \text{ are integers, with } b \neq 0\}$. (This set is traditionally denoted by $\mathbb{Q}$.) Can you think of a nonempty subset of the positive rationals that fails to have a least element?

We will take this property of the integers as a fundamental *axiom*, that is, we will merely accept it as given and not try to prove it from more fundamental principles. Also, we will give it a name:

> *Well-Ordering Principle*: Every nonempty subset of the nonnegative integers has a least element.

Now let us look at divisibility. Why do we say that 2 divides 6? It is because there is another integer, namely 3, such that the product 2 times 3 *exactly* gives us 6. On the other hand, why do we say that 2 does not divide 7? This is because no matter how hard we search, we will not be able to find an integer b such that 2 times b equals 7. This idea will be the basis of our definition:

Definition 1.1. A (nonzero) integer d is said to *divide* an integer a (denoted $d|a$) if there exists an integer b such that $a = db$. If d divides a, then d is referred to as a *divisor* of a or a *factor* of a, and a is referred to as a *multiple* of d.

Observe that this is a slightly more general definition than most of us are used to—according to this definition, -2 divides 6 as well, since there exists an integer, namely -3, such that -2 times -3 equals 6. Similarly, 2 divides -6, since 2 times -3 equals -6. More generally, if d divides a, then all of the following are also true: $d| - a$, $-d|a$, $-d| - a$. (Take a minute to prove this formally!) It is quite reasonable to include negative integers in our concept of divisibility, but for convenience, we will often focus on the case where the divisor is positive.

The following easy result will be very useful:

Lemma 1.2. *If d is a nonzero integer such that $d|a$ and $d|b$ for two integers a and b, then for any integers x and y, $d|(xa+yb)$. (In particular, $d|(a+b)$ and $d|(a-b)$.)*

Proof. Since $d|a$, $a = dm$ for some integer m. Similarly, $b = dn$ for some integer n. Hence $xa + yb = xdm + ydn = d(xm + yn)$. Since we have succeeded in writing $xa + yb$ as d times the integer $xm + yn$, we find that $d|(xa+yb)$. As for the statement in the parentheses, taking $x = 1$ and $y = 1$, we find that $d|a + b$, and taking $x = 1$ and $y = -1$, we find that $d|a - b$. □

> **Question 1.3.** If a nonzero integer d divides both a and $a + b$, must d divide b as well?

The following lemma holds the key to the division process. Its statement is often referred to as the division algorithm. The Well-Ordering Principle plays a central role in its proof.

Lemma 1.4. *(Division Algorithm) Given integers a and b with $b > 0$, there exist* unique *integers q and r, with $0 \leq r < b$ such that $a = bq + r$.*

Remark 1.5. First, observe the range that r lies in. It is constrained to lie between 0 and $b - 1$ (with both 0 and $b - 1$ included as possible values for r). Next, observe that the lemma does not just state that integers q and r exist with $0 \leq r < b$ and $a = bq + r$, it goes further—it states that these integers q and r are *unique.* This means that if somehow one were to have $a = bq_1 + r_1$ and $a = bq_2 + r_2$ for integers q_1, r_1, q_2, and r_2 with $0 \leq r_1 < b$ and $0 \leq r_2 < b$, then q_1 must equal q_2 and r_1 must equal r_2. The integer q is referred to as the *quotient* and the integer r is referred to as the *remainder.*

Proof of Lemma 1.4. Let S be the set $\{a - bn \mid n \in \mathbb{Z}\}$. Thus, S contains the following integers: a $(= a - b \cdot 0)$, $a - b$, $a + b$, $a - 2b$, $a + 2b$, $a - 3b$, $a + 3b$, etc. Let S^* be the set of all those elements in S that are nonnegative, that is, $S^* = \{a - bn \mid n \in \mathbb{Z},\ and\ a - bn \geq 0\}$. It is not immediate that S^* is nonempty, but if we think a bit harder about this, it will be clear that S^* indeed has elements in it. For if a is nonnegative, then $a \in S^*$. If a is negative, then $a - ba$ is nonnegative (check! remember that b itself is positive, by hypothesis), so $a - ba \in S^*$. By the Well-Ordering Principle, since S^* is a *nonempty* subset of $\mathbb{N}$, S^* has a least element; call it r. (The notation r is meant to be suggestive; this element will be the "r" guaranteed by the lemma.)

Since r is in S (actually in S^* as well), r must be expressible as $a - bq$ for some integer q, since *every* element of S is expressible as $a - bn$ for some integer n. (The notation q is also meant to be suggestive, this integer will be the "q" guaranteed by the lemma.) Since $r = a - bq$, we find $a = bq + r$. What we need to do now is to show that $0 \leq r < b$, and that q and r are unique.

Observe that since r is in S^* and since all elements of S^* are nonnegative, r must be nonnegative, that is $0 \leq r$. Now suppose $r \geq b$. We will arrive at a contradiction: Write $r = b + x$, where $x \geq 0$ (why is $x \geq 0$?). Writing $b + x$ for r in $a = bq + r$, we find $a = bq + b + x$, or $a = b(q + 1) + x$, or $x = a - b(q + 1)$. This form of x shows that x belongs to the set S (why?). Since we have already seen that $x \geq 0$, we find further that $x \in S^*$. But more is true: since $x = r - b$ and $b > 0$, x must be less than r (why?). Thus, x is an element of S^* that is smaller that r—a contradiction to the fact that r is the least element of S^*! Hence, our assumption that $r \geq b$ must have been false, so $r < b$. Putting this together with the fact that $0 \leq r$, we find that $0 \leq r < b$, as desired.

Now for the uniqueness of q and r. Suppose $a = bq + r$ and as well, $a = bq' + r'$, for integers q, r, q', and r' with $0 \leq r < b$ and $0 \leq r' < b$. Then $b(q - q') = r' - r$. Thus, $r' - r$ is a multiple of b. Now the fact that $0 \leq r < b$ and $0 \leq r' < b$ shows that $-b < r' - r < b$. (Convince yourselves of this!) The only multiple of b in the range $(-b, b)$ (both endpoints of the range excluded) is 0. Hence, $r' - r$ must equal 0, that is, $r' = r$. It follows that $b(q - q') = 0$, and since $b \neq 0$, we find that $q = q'$.

□

Observe that to test whether a given (positive) integer d divides a given integer a, it is enough to write a as $dq + r$ $(0 \leq r < d)$ as in Lemma 1.4 and examine whether the remainder r is zero or not. For $d|a$ if and only there exists an integer x such that $a = dx$. View this as $a = dx + 0$. By the uniqueness part of Lemma 1.4, we find that $a = dx + 0$ if and only if $b = x$ and $r = 0$.

Now, given two nonzero integers a and b, it is natural to wonder whether they have any divisors in common. Notice that 1 is automatically a common divisor of a and b, no matter what a and b are. Recall that $|a|$ denotes the absolute value of a, and notice that every divisor d of a is less than or equal to $|a|$. (Why? Notice, too, that $|a|$ is a divisor of a.) Also, for every divisor

d of a, we must have $d \geq -|a|$. (Why? Notice, too, that $-|a|$ is a divisor of a.) Similarly, every divisor d of b must be less than or equal to $|b|$ and greater than or equal to $-|b|$ (and both $|b|$ and $|-b|$ are divisors of b). It follows that every common divisor of a and b must be less than or equal to the lesser of $|a|$ and $|b|$, and must be greater than or equal to the greater of $-|a|$ and $-|b|$. Thus, there are only finitely many common divisors of a and b, and they all lie in the range $\max(-|a|, -|b|)$ to $\min(|a|, |b|)$.

We will now focus on a very special common divisor of a and b.

Definition 1.6. Given two (nonzero) integers a and b, the *greatest common divisor* of a and b (written as $\gcd(a, b)$) is the *largest* of the common divisors of a and b.

Note that since there are only finitely many common divisors of a and b, it makes sense to talk about the largest of the common divisors.

> **Question 1.7.** By contrast, must an infinite set of integers necessarily have a largest element? Must an infinite set of integers necessarily *fail* to have a largest element? What would your answers to these two questions be if we restricted our attention to an infinite set of positive integers? How about if we restricted our attention to an infinite set of negative integers?

Notice that since 1 is already a common divisor, the greatest common divisor of a and b must be at least as large as 1. We can conclude from this that the greatest common divisor of two nonzero integers a and b must be *positive.*

> **Question 1.8.** If p and q are two positive integers and if q divides p, what must $\gcd(p, q)$ be?

See the notes on Page 20 for a discussion on the restriction that both a and b be nonzero in Definition 1.6 above.

Let us derive an alternative formulation for the greatest common divisor that will be very useful. Given two nonzero integers a and b, any integer that can be expressed in the form $xa + yb$ for some integers x and y is called a *linear combination* of a and b. (For example, $a = 1 \cdot a + 0 \cdot b$ is a linear combination of a and b; so are $3a - 5b$, $-6a + 10b$, $-b = 0 \cdot a + (-1) \cdot b$, etc.) Write P for the set of linear combinations of a and b that are *positive.* (For instance, if $a = 2$ and $b = 3$, then $-2 = (-1) \cdot 2 + (0) \cdot 3$ would not be in P as -2 is negative, but $7 = 2 \cdot 2 + 3$ would be in P as 7 is positive.) Now here is something remarkable: the smallest element in P turns out to be the greatest common divisor of a and b! We will prove this below.

Theorem 1.9. *Given two nonzero integers a and b, let P be the set $\{xa + yb | x, y \in \mathbb{Z},\ xa + yb > 0\}$. Let d be the least element in P. Then $d = gcd(a, b)$. Moreover, every element of P is divisible by d.*

Proof. First observe that P is not empty. For if $a > 0$, then $a \in P$, and if $a < 0$, then $-a \in P$. Thus, since P is a nonempty subset of $\mathbb{N}$ (actually, of the positive integers as well), the Well-Ordering Principle guarantees that there is a least element d in P, as claimed in the statement of the theorem.

To show that $d = \gcd(a, b)$, we need to show that d is a common divisor of a and b, and that d is the largest of all the common divisors of a and b.

First, since $d \in P$, and since every element in P is a linear combination of a and b, d itself can be written as a linear combination of a and b. Thus, there exist integers x and y such that $d = xa + yb$. (Note: These integers x and y need not be unique. For instance, if $a = 4$ and $b = 6$, we can express 2 as both $(-1) \cdot 4 + 1 \cdot 6$ and $(-4) \cdot 4 + 3 \cdot 6$. However, this will not be a problem; we will simply pick one pair x, y for which $d = xa + yb$ and stick to it.)

Let us show that d is a common divisor of a and b. Write $a = dq + r$ for integers d and r with $0 \le r < d$ (division algorithm). We need to show that $r = 0$. Suppose to the contrary that $r > 0$. Write $r = a - dq$. Substituting $xa + yb$ for d, we find that $r = (1 - xq)a + (-yq)b$. Thus, r is a positive linear combination of a and b that is less than d—a contradiction, since d is the smallest positive linear combination of a and b. Hence r must be zero, that is, d must divide a. Similarly, one can prove that d divides b as well, so that d is indeed a common divisor of a and b.

Now let us show that d is the largest of the common divisors of a and b. This is the same as showing that if c is any common divisor of a and b, then c must be no larger than d. So let c be any common divisor of a and b. Then, by Lemma 1.2 and the fact that $d = xa + yb$, we find that $c|d$. Thus, $c \leq |d|$ (why?). But since d is positive, $|d|$ is the same as d. Thus, $c \leq d$, as desired.

To prove the last statement of the theorem, note that we have already proved that $d|a$ and $d|b$. By Lemma 1.2, d must divide all linear combinations of a and b, and must hence divide every element of P.

We have thus proved our theorem. □

In the course of proving Theorem 1.9 above, we have actually proved something else as well, which we will state as a separate result:

Proposition 1.10. *Every common divisor of two nonzero integers a and b divides their greatest common divisor.*

Proof. As remarked above, the ideas behind the proof of this corollary are already contained in the proof of Theorem 1.9 above. We saw there that if c is any common divisor of a and b, then c must divide d, where d is the minimum of the set P defined in the statement of the theorem. But this, along with the other arguments in the proof of the theorem, showed that d must be the greatest common divisor of a and b. Thus, to say that c divides d is really to say that c divides the greatest common divisor of a and b, thus proving the proposition. □

Exercise 1.37 will yield yet another description of the greatest common divisor.

Question 1.11. Given two nonzero integers a and b for which one can find integers x and y such that $xa + yb = 2$, can you conclude from Theorem 1.9 that $\gcd(a, b) = 2$? If not, why not? What, then, are the possible values of $\gcd(a, b)$? Now suppose there exist integers x' and y' such that $x'a + y'b = 1$. Can you conclude that $\gcd(a, b) = 1$? (See the notes on Page 20 *after* you have thought about these questions for at least a little bit yourselves!)

Given two nonzero integers a and b, we noted that 1 is a common divisor of a and b. In general, a and b could have other common divisors greater than 1, but in certain cases, it may turn out that the greatest common divisor of a and b is precisely 1. We give a special name to this:

Definition 1.12. Two nonzero integers a and b are said to be *relatively prime* if $\gcd(a, b) = 1$.

We immediately have the following:

Corollary 1.13. *Given two nonzero integers a and b, $gcd(a, b) = 1$ if and only if there exist integers x and y such that $xa + yb = 1$.*

Proof. You should be able to prove this yourselves! (See Question 1.11 above.) □

The following lemma will be useful:

Lemma 1.14. *Let a and b be positive integers, and let c be a third integer. If $a|bc$ and $gcd(a, b) = 1$, then $a|c$.*

Proof. Since $\gcd(a, b) = 1$, Theorem 1.9 shows that there exist integers x and y such that $1 = xa + yb$. Multiplying by c, we find that $c = xac + ybc$. Since $a|a$ and $a|bc$, a must divide c by Lemma 1.2. □

We are now ready to introduce the notion of a prime!

Definition 1.15. An integer p greater than 1 is said to be *prime* if its only divisors are ± 1 and $\pm p$. (An integer greater than 1 that is not prime is said to be *composite*.)

The first ten primes are 2, 3, 5, 7, 11, 13, 17, 19, 23, and 29. The hundredth prime is 541.

Primes are intriguing things to study. On the one hand, they should be thought of as being *simple,* in the sense that their only positive divisors are 1 and themselves. (This is sometimes described by the statement "primes have no nontrivial divisors.") On the other hand, there is an immense number of questions about them that are still unanswered, or at best, only partially answered. For instance: is every even integer greater than 4 expressible as a sum of two primes? (This is known as "Goldbach's conjecture." The answer is unknown.) Are there infinitely many twin primes? (The answer to this is also unknown.) Is there any pattern to the occurence of the primes among the integers? Here, some partial answers are known. The following is just a sample: There are arbitrarily large gaps between consecutive primes, that is, given any n, it is possible to find two consecutive primes that differ by at least n. (See Exercise 1.31.) It is known that for any $n > 1$, there is always a prime between n and $2n$. (It is unknown whether there is a prime between n^2 and $(n+1)^2$, however!) It is known that as n becomes very large, the number of primes less than n is approximately $n/\ln(n)$, in the sense that the ratio between the number of primes less than n and $n/\ln(n)$ approaches 1 as n becomes large. (This is the celebrated *Prime Number Theorem.*) Also, it is known that given any arithmetic sequence a, $a+d$, $a+2d$, $a+3d$, $\ldots$, where a and d are nonzero integers with $\gcd(a,d) = 1$, infinitely many of the integers that appear in this sequence are primes!

Those of you who find this fascinating should delve deeper into number theory, which is the branch of mathematics that deals with such questions. It is a wonderful subject with hordes of problems that will seriously challenge your creative abilities! For now, we will content ourselves with proving the unique prime factorization property and the infinitude of primes already referred to at the beginning of this chapter.

The following lemmas will be needed:

Lemma 1.16. *Let p be a prime and a an arbitrary integer. Then either $p|a$ or else $gcd(p, a) = 1$.*

Proof. If p already divides a, we have nothing to prove, so let us assume that p does not divide a. We need to prove that $\gcd(p, a) = 1$. Write x for $\gcd(p, a)$. By definition x divides p. Since the only positive divisors of p are 1 and p, either $x = 1$ (which is want we want to show), or else $x = p$. Suppose $x = p$. Then, as x divides a as well, we find p divides a. But we have assumed that p does not divide a. Hence $x = 1$.

□

Lemma 1.17. *Let p be a prime. If $p|ab$ for two integers a and b, then either $p|a$ or else $p|b$.*

Proof. If p already divides a, we have nothing to prove, so let us assume that p does not divide a. Then by Lemma 1.16, $\gcd(p, a) = 1$. It now follows from Lemma 1.14 that $p|b$. □

The following generalization of Lemma 1.17 will be needed in the proof of Theorem 1.19 below:

> **Exercise 1.18.** Show using induction and Lemma 1.17 that if a prime p divides a product of integers $a_1 \cdot a_2 \cdots a_k$ $(k \geq 2)$, then p must divide one of the a_i's.

We are ready to prove our factorization theorem!

Theorem 1.19. *(Fundamental Theorem of Arithmetic) Every positive integer greater than 1 can be factored into a product of primes. The primes that occur in any two factorizations are the same, except perhaps for the order in which they occur in the factorization.*

Remark 1.20. The statement of this theorem has two parts to it. The first sentence is an *existence* statement—it asserts that for every positive integer greater than 1, a prime factorization *exists.* The second sentence is a *uniqueness* statement. It asserts that except for rearrangement, *there can only be* *one prime factorization.* To understand this second assertion a little better, consider the two factorizations of 12 as $12 = 3 \times 2 \times 2$, and $12 = 2 \times 3 \times 2$. The orders in which the 2's and the 3 appear are different, but in both factorizations, 2 appears twice, and 3 appears once. The uniqueness part of the theorem tells us that no matter how 12 is factored, we will at most be able to rearrange the order in which the two 2's and the 3 appear such as in the two factorizations above, but every factorization must consist of exactly two 2's and one 3.

Proof of Theorem 1.19. We will prove the existence part first. The proof is very simple. Assume to the contrary that there exists an integer greater than 1 that does not admit prime factorization. Then, the set of positive integers greater than 1 that do not admit prime factorization is nonempty, and hence, by the Well-Ordering Principle, there must be a *least* positive integer greater than 1, call it a, that does not admit prime factorization. Now a cannot itself be prime, or else, "$a = a$" would be its prime factorization, contradicting our assumption about a. Hence, $a = bc$ for suitable positive integers b and c, with $1 < b < a$ and $1 < c < a$. But then, b and c must both admit factorization into primes, since they are greater than 1 and less than a, and a was the least positive integer greater than 1 without a prime factorization. If $b = p_1 \cdot p_2 \cdots p_k$ and $c = q_1 \cdot q_2 \cdots q_l$ are prime factorizations of b and c respectively, then $a(= bc) = p_1 \cdot p_2 \cdots p_k \cdot q_1 \cdot q_2 \cdots q_l$ yields a prime factorization of a, contradicting our assumption about a. Hence, no such integer a can exist, that is, every positive integer must factor into a product of primes.

Let us move on to the uniqueness part of the theorem. The basic ideas behind the proof of this portion of the theorem are quite simple as well. The key is to recognize that if an integer a has two prime factorizations, then some prime in the first factorization must equal some prime in the second factorization. This will then allow us to cancel the two primes, one from each factorization, and arrive at two factorizations of a smaller integer. The rest is just induction.

So assume to the contrary that there exists a positive integer greater than 1 with two different (i.e., other than for rearrangement) prime factorizations. Then, exactly as in the proof of the existence part above, the Well-Ordering Principle applied to the (nonempty) set of positive integers greater than 1 that admit two different prime factorizations shows that there must be a *least* positive integer greater than 1, call it a, that admits two different prime factorizations. Suppose that

$$a = p_1^{n_1} \cdots p_s^{n_s} = q_1^{m_1} \cdots q_t^{m_t},$$

where the p_i $(i = 1, \ldots, s)$ are distinct primes, and the q_j $(j = 1, \ldots, t)$ are distinct primes, and the n_i and the m_j are positive integers. (By "distinct primes" we mean that p_1, p_2, $\ldots$, p_s are all different from one another, and similarly, q_1, q_2, $\ldots$, q_t are all different from one another.) Since p_1 divides a, and since $a = q_1^{m_1} \cdots q_t^{m_t}$, p_1 must divide $q_1^{m_1} \cdots q_t^{m_t}$. Now, by Exercise 1.18 above (which simply generalizes Lemma 1.17), we find that since p_1 divides the product $q_1^{m_1} \cdots q_t^{m_t}$, it must divide one of the factors of this product, that is, it must divide one of the q_j. Relabeling the primes q_j if necessary (remember, we do not consider a rearrangement of primes to be a different factorization), we may assume that p_1 divides q_1. Since the only positive divisors of q_1 are 1 and q_1, we find $p_1 = q_1$.

Since now $p_1 = q_1$, consider the integer $a' = a/p_1 = a/q_1$. If $a' = 1$, this means that $a = p_1 = q_1$, and there is nothing to prove, the factorization of a is already unique. So assume that $a' > 1$. Then a' is a positive integer greater than 1 and less than a, so by our assumption about a, any prime factorization of a' must be unique (that is, except for rearrangement of factors). But then, since a' is obtained by dividing a by p_1 $(= q_1)$, we find that a' has the prime factorizations

$$a' = p_1^{n_1-1} \cdots p_s^{n_s} = q_1^{m_1-1} \cdots q_t^{m_t}$$

So, by the uniqueness of prime factorization of a', we find that $n_1-1 = m_1-1$

(so $n_1 = m_1$), $s = t$, and after relabeling the primes if necessary, $p_i = q_i$, and similarly, $n_i = m_i$, for $i = 2, \ldots, s(= t)$. This establishes that the two prime factorizations of a we began with are indeed the same, except perhaps for rearrangement.

$\square$

Remark 1.21. While Theorem 1.19 only talks about integers greater than 1, a similar result holds for integers less than -1 as well: every integer less than -1 can be factored as -1 times a product of primes, and these primes are unique, except perhaps for order. This is clear, since, if a is a negative integer less than -1, then $a = -1 \cdot |a|$, and of course, $|a| > 1$ and therefore admits unique prime factorization.

The following result follows easily from studying prime factorizations and will be useful in the exercises:

Proposition 1.22. *Let a and b be integers greater than 1. Then b divides a if and only if the prime factors of b are a subset of the prime factors of a and if a prime p occurs in the factorization of b with exponent y and in the factorization of a with exponent x, then $y \leq x$.*

Proof. Let us assume that $b|a$, so $a = bc$ for some integer c. If $c = 1$, then $a = b$, and there is nothing to prove, the assertion is obvious. So suppose $c > 1$. Then c also has a factorization into primes, and multiplying together the prime factorizations of b and c, we get a factorization of bc into a product of primes. On the other hand, bc is just a, and a has its own prime factorization as well. By the uniqueness of prime factorizations, the prime factorization of bc that we get from multiplying together the prime factorizations of b and c *must* be *the* prime factorization of a. In particular, the prime factors of b (and c) must be a subset of the prime factors of a. Now suppose that a prime p occurs to the power x in the factorization of a, to the power y in the factorization of b, and to the power z in the factorization of c. Multiplying together the factorizations of b and c, we find that p occurs to the power $y+z$ in the factorization of bc. Since the factorization of bc is just the factorization of a and since p occurs to the power x in the factorization of a, we find that $x = y + z$. In particular, $y \leq x$. This proves one half of the proposition.

As for the converse, assume that b has the prime factorization $b = p_1^{n_1} \cdots p_s^{n_s}$. Then, by the hypothesis, the primes $p_1, \ldots, p_s$ must all appear in the prime factorization of a with exponents at least $n_1, \ldots, n_s$ (respectively). Thus, the prime factorization of a must look like $a = p_1^{m_1} \cdots p_s^{m_s} p_{s+1}^{m_{s+1}} \cdots p_t^{m_t}$, where $m_i \geq n_i$ for $i = 1, \ldots, s$, and where $p_{s+1}, \ldots, p_t$ are other primes. Writing c for $p_1^{m_1-n_1} \cdots p_s^{m_s-n_s} p_{s+1}^{m_{s+1}} \cdots p_t^{m_t}$ and noting that $m_i - n_i \geq 0$ for $i = 1, \ldots, s$ by hypotheses, we find that c is an integer, and of course, clearly, $a = (p_1^{n_1} \cdots p_s^{n_s})c$, i.e, $a = bc$. This proves the converse.

□

We have proved the Fundamental Theorem of Arithmetic, but there remains the question of showing that there are infinitely many primes. The proof that we provide is due to Euclid, and is justly celebrated for its beauty.

Theorem 1.23. *(Euclid) There exist infinitely many prime numbers.*

Proof. Assume to the contrary that there are only finitely many primes. Label them $p_1,\ p_2,\ \ldots,\ p_n$. (Thus, we assume that there are n primes.) Consider the integer $a = p_1 p_2 \cdots p_n + 1$. Since $a > 1$, a admits a prime factorization by Theorem 1.19. Let q be any prime factor of a. Since the set $\{p_1,\ p_2,\ \ldots, p_n\}$ contains all the primes, q must be in this set, so q must equal, say, p_i. But then, $a = q(p_1 p_2 \cdots p_{i-1} p_{i+1} \cdots p_n) + 1$, so we get a remainder of 1 when we divide a by q. In other words, q cannot divide a. This is a contradiction. Hence there must be infinitely many primes! □

Question 1.24. What is wrong with the following proof of Theorem 1.23?—There are infinitely many positive integers. Each of them factors into primes by Theorem 1.19. Hence there must be infinitely many primes.

Further Exercises

Exercise 1.25. In this exercise, we will formally prove the validity of various quick tests for divisibility that we learn in high school!

1. Prove that an integer is divisible by 2 if and only if the digit in the units place is divisible by 2. (Hint: Look at a couple of examples: $58 = 5 \cdot 10 + 8$, while $57 = 5 \cdot 10 + 7$. What does Lemma 1.2 suggest in the context of these examples?)

2. Prove that an integer (with two or more digits) is divisible by 4 if and only if the integer represented by the tens digit and the units digit is divisible by 4. (To give you an example, the "integer represented by the tens digit and the units digit" of 1024 is 24, and the assertion is that 1024 is divisible by 4 if and only if 24 is divisible by 4—which it is!)

3. Prove that an integer (with three or more digits) is divisible by 8 if and only if the integer represented by the hundreds digit and the tens digit and the units digit is divisible by 8.

4. Prove that an integer is divisible by 3 if and only if the sum of its digits is divisible by 3. (For instance, the sum of the digits of 1024 is $1+0+2+4 = 7$, and the assertion is that 1024 is divisible by 3 if and only if 7 is divisible by 3—and therefore, since 7 is not divisible by 3, we can conclude that 1024 is not divisible by 3 either! Here is a hint in the context of this example: $1024 = 1 \cdot 1000 + 0 \cdot 100 + 2 \cdot 10 + 4 = 1 \cdot (999 + 1) + 0 \cdot (99 + 1) + 2 \cdot (9 + 1) + 4$. What can you say about the terms containing 9, 99, and 999 as far as divisibility by 3 is concerned? Then, what does Lemma 1.2 suggest?)

5. Prove that an integer is divisible by 9 if and only if the sum of its digits is divisible by 9.

6. Prove that an integer is divisible by 11 if and only if the difference between the sum of the digits in the units place, the hundreds place, the ten thousands place, ... (the places corresponding to the even powers of 10) and the sum of the digits in the tens place, the thousands place, the hundred thousands place, ... (the places corresponding to the odd powers of 10) is divisible by 11. (Hint: $10 = 11 - 1$, $100 = 99 + 1$, $1000 = 1001 - 1$, $10000 = 9999 + 1$, etc. What can you say about the integers 11, 99, 1001, 9999, etc. as far as divisibility by 11 is concerned?)

Exercise 1.26. Given nonzero integers a and b, with $b > 0$, write $a = bq + r$ (division algorithm). Show that $\mathrm{gcd}(a, b) = \mathrm{gcd}(b, r)$.

(This exercise forms the basis for the Euclidean algorithm for finding the greatest common divisor of two nonzero integers. For instance, how do we find the greatest common divisor of, say, 48 and 30 using this algorithm? We divide 48 by 30 and find a remainder of 18, then we divide 30 by 18 and find a remainder of 12, then we divide 18 by 12 and find a remainder of 6, and finally, we divide 12 by 6 and find a remainder of 0. Since 6 divides 12 evenly, we claim that $\mathrm{gcd}(48, 30) = 6$. What is the justification for this claim? Well, applying the statement of this exercise to the first division, we find that $\mathrm{gcd}(48, 30) = \mathrm{gcd}(30, 18)$. Applying the statement to the second division, we find that $\mathrm{gcd}(30, 18) = \mathrm{gcd}(18, 12)$. Applying the statement to the third division, we find that $\mathrm{gcd}(18, 12) = \mathrm{gcd}(12, 6)$. Since the fourth division shows that 6 divides 12 evenly, $\mathrm{gcd}(12, 6) = 6$. Working our way backwards, we obtain $\mathrm{gcd}(48, 30) = \mathrm{gcd}(30, 18) = \mathrm{gcd}(18, 12) = \mathrm{gcd}(12, 6) = 6$.)

Exercise 1.27. Given nonzero integers a and b, let $h = a/\mathrm{gcd}(a, b)$ and $k = b/\mathrm{gcd}(a, b)$. Show that $\mathrm{gcd}(h, k) = 1$.

Exercise 1.28. Show that if a and b are nonzero integers with $\mathrm{gcd}(a, b) = 1$, and if c is an arbitrary integer, then $a|c$ and $b|c$ together imply $ab|c$. Give a counterexample to show that this result is false if $\mathrm{gcd}(a, b) \neq 1$. (Hint: Just as in the proof of Lemma 1.14, use the fact that $\mathrm{gcd}(a, b) = 1$ to write $1 = xa + yb$ for suitable integers x and y, and then multiply both sides by c. Now stare hard at your equation!)

Exercise 1.29. The *Fibonacci Sequence,* $1, 1, 2, 3, 5, 8, 13, \cdots$ is defined as follows: If a_i stands for the ith term of this sequence, then $a_1 = 1$, $a_2 = 1$, and for $n \geq 3$, a_n is given by the formula $a_n = a_{n-1} + a_{n-2}$. Prove that for all $n \geq 2$, $\mathrm{gcd}(a_n, a_{n-1}) = 1$.

Exercise 1.30. Given an integer $n \geq 1$, recall that $n!$ is the product $1 \cdot 2 \cdot 3 \cdots (n-1) \cdot n$. Show that the integers $(n+1)! + 2, (n+1)! + 3, \ldots, (n+1)! + (n+1)$ are all composite.

Exercise 1.31. Use Exercise 1.30 to prove that given any positive integer n, one can always find *consecutive* primes p and q such that $q - p \geq n$.

Exercise 1.32. If m and n are odd integers, show that 8 divides $m^2 - n^2$.

Exercise 1.33. Show that 3 divides $n^3 - n$ for any integer n. (Hint: Factor $n^3 - n$ as $n(n^2 - 1) = n(n-1)(n+1)$. Write n as $3q + r$, where r is one of 0, 1, or 2, and examine, for each value of r, the divisibility of each of these factors by 3. This result is a special case of Fermat's Little Theorem , which you will encounter as Theorem 4.42 in Chapter 4 ahead.)

Exercise 1.34. Here is another instance of Fermat's Little Theorem : show that 5 divides $n^5 - n$ for any integer n. (Hint: As in the previous exercise, factor $n^5 - n$ appropriately, and write $n = 5q + r$ for $0 \leq r < 5$.)

Exercise 1.35. . Show that 7 divides $n^7 - n$ for any integer n.

Exercise 1.36. Use Proposition 1.22 to show that the number of positive divisors of n is $(n_1 + 1)(n_2 + 1) \cdots (n_k + 1)$.

Exercise 1.37. Let m and n be positive integers. By allowing the exponents in the prime factorizations of m and n to equal 0 if necessary, we may assume that $m = p_1^{m_1} p_2^{m_2} \cdots p_k^{m_k}$ and $n = p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$, where for $i = 1, \cdots, k$, p_i is prime, $m_i \geq 0$, and $n_i \geq 0$. (For instance, we can rewrite the factorizations $84 = 2^2 \cdot 3 \cdot 7$ and $375 = 3 \cdot 5^3$ as $84 = 2^2 \cdot 3 \cdot 5^0 \cdot 7$ and $375 = 2^0 \cdot 3 \cdot 5^3 \cdot 7^0$.) For each i, let $d_i = \min(m_i, n_i)$. Prove that $\gcd(m, n) = p_1^{d_1} p_2^{d_2} \cdots p_k^{d_k}$.

Exercise 1.38. Given two (nonzero) integers a and b, the *least common multiple* of a and b (written as $\mathrm{lcm}(a, b)$) is defined to be the smallest of all the positive common multiples of a and b.

1. Show that this definition makes sense, that is, show that the set of positive common multiples of a and b has a smallest element.

2. Retaining the notation of Exercise 1.37 above, let $l_i = \max(m_i, n_i)$ $(i = 1, \ldots, k)$. Show that $\mathrm{lcm}(m, n) = p_1^{l_1} p_2^{l_2} \cdots p_k^{l_k}$.

3. Use Exercise 1.37 and Part 2 above to show that $\mathrm{lcm}(a, b) = ab/\gcd(a, b)$.

4. Conclude that if if $\gcd(a, b) = 1$, then $\mathrm{lcm}(a, b) = ab$.

Exercise 1.39. Let $a = p^n$, where p is a prime and n is a positive integer. Prove that the number of integers x such that $1 \leq x \leq a$ and $\gcd(x, a) = 1$ is $p^n - p^{n-1}$.

(More generally, if a is *any* integer greater than 1, one can ask for the number of integers x such that $1 \leq x \leq a$ and $\gcd(x, a) = 1$. This number is denoted by $\phi(a)$, and is referred to as *Euler's ϕ-function*. It turns out that if a has the prime factorization $p_1^{m_1} p_2^{m_2} \cdots p_k^{m_k}$, then $\phi(a) = \phi(p_1^{m_1}) \cdot \phi(p_2^{m_2}) \cdot \cdots \cdot \phi(p_k^{m_k})$! Delightful as this statement is, we will not delve deeper into it in this book, but you are encouraged to read about it in any introductory textbook on number theory.)

Exercise 1.40. The series $1+1/2+1/3+\cdots$ is known as the *harmonic series.* This exercise concerns the partial sums (see below) of this series.

1. Fix an integer $n \geq 1$, and let S_n denote the set $\{1, 2, \ldots, n\}$ Let 2^t be the highest power of 2 that appears in S_n. Show that 2^t does not divide any element of S_n other than itself.

2. For any integer $n \geq 1$, the *nth partial sum* of the harmonic series is the sum of the first n terms of the series, that is, it is the number $1 + 1/2 + 1/3 + \cdots 1/n$. Show that if $n \geq 2$, the nth partial sum is *not* an integer as follows:

   (a) Clearing denominators, show that the nth partial sum may be written as a/b, where $b = n!$ and $a = (2 \cdot 3 \cdots n) + (2 \cdot 4 \cdots n) + (2 \cdot 3 \cdot 5 \cdots n) + \cdots + (2 \cdot 3 \cdots n - 1)$.

   (b) Let S_n and 2^t be as in part 1 above. Also, let 2^m be the highest power of 2 that divides $n!$. Show that $m \geq t \geq 1$ and that $m \geq m - t + 1 \geq 1$.

   (c) Conclude from part 2b above that 2^{m-t+1} divides b.

   (d) Use part 1 to show that 2^{m-t+1} divides all the summands in the expression in part 2a above for a except the term $(2 \cdot 3 \cdots 2^t - 1 \cdot 2^t + 1 \cdots n)$.

   (e) Conclude that 2^{m-t+1} does not divide a.

   (f) Conclude that the nth partial sum is not an integer.

Exercise 1.41. Fix an integer $n \geq 1$, and let S_n denote the set $\{1, 3, 5, \ldots, 2n-1\}$. Let 3^t be the highest power of 3 that appears in S_n. Show that 3^t does not divide any element of S_n other than itself. Can you use this result to show that the nth partial sums ($n \geq 2$) of a series analogous to the harmonic series (see Exercise 1.40 above) are not integers?

Exercise 1.42. Prove using the unique prime factorization theorem that $\sqrt{2}$ is not a rational number. Using essentially the same ideas, show that $\sqrt{p}$ is not a rational number for any prime p. (Hint: Suppose that $\sqrt{2} = a/b$ for some two integers a and b with $b \neq 0$. Rewrite this as $a^2 = 2b^2$. What can you say about the exponent of 2 in the prime factorizations of a^2 and $2b^2$?)

Notes

Remarks on Definition 1.6 The alert reader may wonder why we have restricted both integers a and b to be nonzero in Definition 1.6 above. Let us explore this question further: Suppose first that a and b are both zero. Note that every nonzero integer divides 0, since given any nonzero integer n, we certainly have the relation $n \cdot 0 = 0$. Thus, if a and b are both zero, we find that every nonzero integer is a common divisor of a and b, and thus, there is no *greatest* common divisor at all. The concept of the greatest common divisor therefore has no meaning in this situation. Next, let us assume just *one* of a and b is nonzero. For concreteness, let us assume $a \neq 0$ and $b = 0$. Then, as we have seen in the discussions preceding Defintion 1.6, $|a|$ is a divisor of a, and is the largest of the divisors of a. Also, since every nonzero integer divides 0 and we have assumed $b = 0$, we find $|a|$ divides b. It follows that $|a|$ is a common divisor of a and b, and since $|a|$ is the largest among the divisors of a, it has to be the greatest of the common divisors of a and b. We find therefore that if exactly one of a and b, say a, is nonzero, then the concept of $\gcd(a, b)$ has meaning, and the gcd in this case equals $|a|$. However, this situation may be viewed as somewhat less interesting, since every integer anyway divides b. The more interesting case, therefore, is when both a and b are nonzero, and we have chosen to focus on that situation in Definition 1.6.

Remarks on Theorem 1.9 and Exercise 1.11. It is very crucial that d be the *least* positive linear combination of a and b for you to be able to conclude that $\gcd(a, b) = d$. For instance, if you only know that there exist integers x and y such that $xa + yb = 2$, you cannot conclude that $\gcd(a, b) = 2$—for all you know, there may exist two other integers x' and y' such that $x'a + y'b = 1$!

Notice though that if you know that there exist integers x' and y' such that $x'a+y'b = 1$, you *can* conclude that $\gcd(a, b) = 1$. For 1 *has* to be the least positive linear combination of a and b, since there is no positive integer smaller than 1.

Remarks on the definition of the greatest common divisor. We have defined the greatest common divisor of two nonzero integers a and b to be the largest of their common divisors (Definition 1.6), and we have noted that $\gcd(a, b)$ must be positive. On the other hand, Corollary 1.10 showed that every common divisor of a and b must divide $\gcd(a, b)$. Putting these together, we find that $\gcd(a, b)$ has the following specific properties:

1. $\gcd(a, b)$ is a positive integer.
2. $\gcd(a, b)$ is a common divisor of a and b.
3. Every common divisor of a and b must divide $\gcd(a, b)$.

You will find that many textbooks have turned these properties around and *have used these properites to define the greatest common divisor!* Thus, these textbooks define the greatest common divisor of a and b to be that integer d which has the following properties:

1. d is a positive integer.
2. d is a common divisor of a and b.
3. Every common divisor of a and b must divide d.

Of course, it is not immediately clear that such an integer d must exist, nor is it clear that it must be unique, and these books then give a proof of the existence and uniqueness of such a d. Furthermore, it is not immediately clear that the integer d yielded by this definition is the same as the greatest common divisor as we have defined it (although it will be clear if one takes a moment to think about it). The reason why many books prefer to define the greatest common divisor as above is that this definition applies (with a tiny modification) to other number systems where the concept of a "largest" common divisor may not exist.

In the case of the integers, however, we prefer our Definition 1.6, since the largest of the common divisors of a and b is exactly what we would intuitively expect $\gcd(a, b)$ to be!

Chapter 2

Rings and Fields

2.1 Rings: Definition and Examples

Abstract algebra begins with the observation that several sets that occur naturally in mathematics, such as the set of integers, the set of rationals, the set of 2×2 matrices with entries in the reals, the set of functions from the reals to the reals, all come equipped with certain operations that allow one to combine any two elements of the set and come up with a third element. These operations go by different names, such as addition, multiplication, or composition (you would have seen the notion of composing two functions in calculus). Abstract algebra studies mathematics from the point of view of these operations, asking, for instance, what properties of a given mathematical set can be deduced just from the existence of a given operation on the set with a given list of properties. We will be dealing with some of the more rudimentary aspects of this approach to mathematics in this book.

However, do not let the abstract nature of the subject fool you into thinking that mathematics no longer deals with concrete objects! Abstraction grows only from extensive studies of the concrete, it is merely a device (albeit an extremely effective one) for codifying phenomena that simultaneously occur in several concrete mathematical sets. In particular, to understand an abstract concept well, you must work with the specific examples from which the abstract concept grew (remember the advice on active learning).

Let us look at $\mathbb{Z}$, focusing on the operations of addition and multiplication.

Given a set S, recall that a *binary operation* on S is a process that takes an ordered pair of elements from S and gives us a third member of the set. It is helpful to think of this in more abstract terms—a binary operation on S is just a function $f\colon S \times S \to S$, that is, a rule that assigns to each ordered pair (a, b), a third element $f(a, b)$. Given an aribitrary set S, it is quite easy to define binary operations on it, but it is much harder to define binary operations that satisfy additional properties.

> **Question 2.1.** How many different binary operations can be defined on the set $\{0, 1\}$? Now select some of these binary operations and check whether they are associative or commutative. How many binary operations can be constructed on a set T that has n elements?

What will be crucial to us is that addition and multiplication are *special* binary operations on $\mathbb{Z}$ that satisfy certain extra properties.

First, why are addition and multiplication binary operations? The process of adding two integers is of course familiar to us, but suppose we view addition abstractly as a rule that assigns to each ordered pair of integers (m, n) the integer $m+n$. (For instance, addition assigns to the ordered pair $(2, 3)$ the integer 5, to the ordered pair $(3, -4)$ the integer -1, to the ordered pair $(1, 0)$ the integer 1, etc.) It is clear then that addition is indeed a binary operation—it takes an ordered pair of integers, namely (m, n), and gives us a third uniquely determined integer, namely $m+n$. Similarly, multiplication too is a binary operation—it is a rule that assigns to every ordered pair of integers (m, n) the uniquely determined integer $m \cdot n$.

What are the properties of these binary operations? Let us consider addition first. It is customary to write $(\mathbb{Z}, +)$ to emphasize the fact that we are considering $\mathbb{Z}$ not just as a set of objects, but as a set with the binary operation of addition. (We will temporarily ignore the fact that $\mathbb{Z}$ has a second binary operation, namely multiplication, defined on it.) The first property that $(\mathbb{Z}, +)$ has is that $+$ is *associative*. That is, for all integers a, b, and c, $(a+b)+c = a+(b+c)$. The second property that $(\mathbb{Z}, +)$ has is the existence of an *identity element with respect to* $+$. This is the integer 0—it satisfies the condition $a + 0 = 0 + a = a$ all integers a. The third property of $(\mathbb{Z}, +)$ is the existence of *inverses with respect to* $+$. For every integer a, there is an integer b (depending on a) such that $a + b = b + a = 0$. (It is clear what this integer b is, it is just the integer $-a$.)

What these observations show is that the integers form a *group with respect to addition.* We will study groups in detail in Chapter 4 ahead, but let us introduce the concept here. It turns out that the situation we have encountered above (namely, a set equipped with a binary operation with certain properties) arises in several different areas of mathematics. Precisely because the same situation appears in so many different contexts, it has been given a name and has been studied extensively as a subject in its own right.

Definition 2.2. A *group* is a set S with a binary operation "$*$" : $S \times S \to S$ such that

1. $*$ is associative, i.e., $a * (b * c) = (a * b) * c$ for all a, b, and c in S,
2. S has an identity element with respect to $*$, i.e., an element "id" such that $a * id = id * a = a$ for all a in S, and
3. every element of S has an inverse with respect to $*$, i.e., for every element a in S there exists an element "a^{-1}" such that $a * a^{-1} = a^{-1} * a = id$.

To emphasize that there are two ingredients in this definition—the set S and the operation $*$ with these special properties—the group is sometimes written as $(S, *)$, and S is often referred to as *a group with respect to the operation* $*$.

The reason that the integers form a group with respect to addition is that if we take the set "S" of this definition to be $\mathbb{Z}$, and if we take the binary operation "$*$" to be $+$, then the three conditions of the definition are met. There is a vast and beautiful theory about groups, the *beginnings* of which we will pursue in Chapter 4 ahead.

Observe that there is one more property of addition that we have not listed yet, namely *commutativity*. This is the property that for all integers a and b, $a + b = b + a$. In the language of group theory, this makes $(\mathbb{Z}, +)$ an abelian group:

Definition 2.3. An *abelian group* is one in which the function "$*$" in Definition 2.2 above satisfies the additional condition $a * b = b * a$ for all a and b in S.

Commutativity of addition is a crucial property of the integers; the only reason we delayed introducing it was to allow us first to introduce the notion of a group.

Now let us consider multiplication. As with addition, we write $(\mathbb{Z}, \cdot)$ to emphasize the fact that we are considering $\mathbb{Z}$ as a set with the binary operation of multiplication, temporarily ignoring the operation addition. As with addition, we find that multiplication is *associative*, that is, for all integers a, b, and c, $(a \cdot b) \cdot c = a \cdot (b \cdot c)$. Also, $\mathbb{Z}$ has an *identity with respect to multiplication*. This is the integer 1; it satisfies $a \cdot 1 = 1 \cdot a = a$ for all integers a.

> **Question 2.4.** Is $(\mathbb{Z}, \cdot)$ a group? In other words, do the integers form a group with respect to multiplication? To answer this question, you would check whether the three group axioms above hold for $(\mathbb{Z}, \cdot)$. What is the inverse with respect to multiplication of 1? What is the inverse of 2? What is the inverse of 0?

There are two more properties of multiplication of integers we wish to consider. The first is that multiplication is *commutative*, that is, $a \cdot b = b \cdot a$ for all integers a and b. The second, which is not a property of just multiplication alone, but rather a property that connects multiplication and addition together, is the *distributivity of multiplication over addition*, that is, for all integers a, b, and c, $a \cdot (b + c) = a \cdot b + a \cdot c$, and $(a + b) \cdot c = a \cdot c + b \cdot c$. (Notice that since multiplication of integers is commutative, the second relation in the previous sentence follows from the first!)

There are other properties of these operations of course (for instance $a \cdot b = 0$ implies that either $a = 0$ or $b = 0$), but we will study these later. Let us meanwhile reflect on the properties that we have considered so far. Studying them closely, one gets the sense that these properties are somehow rather "natural." For instance, if one were to think of the integers as (intellectual) counting tools, then it is clear that addition must necessarily be commutative, since commutativity of addition corresponds to the fact that if you have a certain number of objects in one pile and a certain number in another, then the total number of objects can be obtained either by counting all the objects in the first pile and then all the objects in the second pile, or by counting all the objects in the second pile and then all the objects in the first pile.

This sense of these properties being "natural" is further reinforced when we consider other "number systems" that we encounter in mathematics. For instance, consider the set of all polynomials in one variable whose coefficients are real numbers, a set with which you are already very familiar. (The real numbers are traditionally denoted by $\mathbb{R}$, and the set of all polynomials in one variable whose coefficients are real numbers is traditionally denoted by $\mathbb{R}[x]$.) This set, too, is more than just a collection of objects. Just as with the integers, $\mathbb{R}[x]$ has two binary operations, also called *addition* and *multiplication*. Recall that given two polynomials $g(x) = \sum_{i=0}^{n} g_i x^i$ and $h(x) = \sum_{j=0}^{m} h_i x^i$, we add g and h by adding together the coefficients of the same powers of x, and we multiply g and h by multiplying each monomial $g_i x^i$ of g by each monomial $h_j x^j$ of h and adding the results together. (For instance, $(1+x+x^2)+(x+\sqrt{3}x^3)$ is $1+2x+x^2+\sqrt{3}x^3$, and $(1+x+x^2)\cdot(x+\sqrt{3}x^3)$ is $x + x^2 + (1 + \sqrt{3})x^3 + \sqrt{3}x^4 + \sqrt{3}x^5$.) Furthermore, it is our experience that these binary operations on $\mathbb{R}[x]$ satisfy the very same properties above that the corresponding operations on $\mathbb{Z}$ satisfied.

It turns out that these properties of addition and multiplication are shared not just by $\mathbb{Z}$ and $\mathbb{R}[x]$, but by a whole host of "number systems" in mathematics. Because of the importance of such sets with two binary operations with these special properties, there is a special term for them—they are called *rings*.

Definition 2.5. A *ring* is a set R with two binary operations $+$ and $\cdot$ such that

1. $a + (b + c) = (a + b) + c$ for all a, b, c in R.
2. There exists an element in R, denoted "0", such that $a + 0 = 0 + a = a$ for all a in R.
3. For each a in R there exists an element in R, denoted "$-a$", such that $a + (-a) = (-a) + a = 0$.
4. $a + b = b + a$ for all elements a, b in R.
5. $a \cdot (b \cdot c) = (a \cdot b) \cdot c$ for all elements a, b, c in R.
6. There exists an element in R, denoted "1", such that $a \cdot 1 = 1 \cdot a = a$ for all a in R.
7. $a \cdot (b + c) = a \cdot b + a \cdot c$ and $(a + b) \cdot c = a \cdot c + b \cdot c$ for all elements a, b, c in R.

Remark 2.6. The binary operation $+$ is usually referred to as *addition* and the binary operation $\cdot$ is usually referred to as *multiplication*, in keeping with the terminology for the integers and other familiar rings. As is the usual practice in high school algebra, one often suppresses the multiplication symbol, that is, one often writes ab for $a \cdot b$.

Remark 2.7. Just as we did earlier with the integers, if we temporarily ignore the operation $\cdot$ on R and write $(R, +)$ to indicate that we are focusing on just the operation $+$, then the first four conditions in the definition of the ring R show that $(R, +)$ is an abelian group.

Remark 2.8. We have used the term "number system" at several places in the book without really being explicit about what a number system is. We did not have the language before this point to make our meaning precise, but what we had intended to convey loosely by this term is the concept of a set with two binary operations with properties much like those of the integers. But now that we have the language, let us be precise: a number system is just a ring as defined above!

It must be borne in mind however that “number system” is a nonstandard term: it is not used very widely, and when used at all, different authors mean different things by the term! So it is better to stick to “rings,” which is standard.

Observe that we left out one important property of the integers in our definition of a ring, namely the commutativity of multiplication. And correspondingly, we have included both *left distributivity* ($a \cdot (b+c) = a \cdot b + a \cdot c$) and *right distributivity* ($(a+b) \cdot c = a \cdot c + b \cdot c$) of multiplication over addition. While this may seem strange at first, think about the set of 2×2 matrices with entries in $\mathbb{R}$. Convince yourselves that this is a ring with respect to the usual definitions of matrix addition and multiplication—see Example 2.16 ahead. Yet, even in this extremely familiar number system, multiplication is not commutative; for instance,

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \neq \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

Rings in which multiplication is not commutative are fairly common in mathematics, and hence requiring commutativity of multiplication in the definition of a ring would be too restrictive. On the other hand, there is no denying that a significant proportion of the rings that we come across indeed have multiplication that is commutative. Thus, it is reasonable to single them out as special cases of rings, and we have the following:

Definition 2.9. A *commutative ring* is a ring R in which $a \cdot b = b \cdot a$ for all a and b in R.

(Rings in which the multiplication is not commutative are referred to as *noncommutative* rings.)

The following are various examples of rings. (Once again, recall the advice in the preliminary chapter To the Student, page ix, on reading actively.)

Example 2.10. The set of rational numbers, $\mathbb{Q}$, with the usual operations of addition and multiplication forms a ring. We know how to add and multiply two rational numbers very well, and we *know* that all the ring axioms hold for the rationals. (One can take a more advanced perspective and *prove* that the ring axioms hold for the rationals, starting from the fact that they hold for the integers. Although sound, such an approach is unduly technical for a first course.) $\mathbb{Q}$ is, in fact, a commutative ring.

> **Question 2.10.1.** $\mathbb{Q}$ has one crucial property (with respect to multiplication) that $\mathbb{Z}$ does not have. Can you discover what that might be? (See the remarks on page 86 in the notes, but only after you have thought about this question on your own!)

Example 2.11. In a like manner, both the reals, $\mathbb{R}$, and the complexes, usually denoted $\mathbb{C}$, are rings under the usual operations of addition and multiplication. Again, we will not try to *prove* that the ring axioms hold; we will just invoke our intimate knowledge of $\mathbb{R}$ and $\mathbb{C}$ to recognize that they are rings.

Example 2.12. Let $\mathbb{Q}[\sqrt{2}]$ denote the set of all real numbers of the form $a + b\sqrt{2}$, where a and b are arbitrary rational numbers. For instance, this includes numbers like $1/2 + 3\sqrt{2}$, $-1/7 + (1/5)\sqrt{2}$, etc. You *know* from your experience with real numbers how to add and multiply two elements $a + b\sqrt{2}$ and $c + d\sqrt{2}$ of this set. Under these operations, this set indeed forms a ring–let us see why:

> **Question 2.12.1.** Here is the first point you need to check: under this method of addition and multiplication, do the sum and product of any two elements of this set also lie in this set? (Remember, a binary operation should take an ordered pair of elements to another element in the *same* set. If, say, the usual product of some two elements $a + b\sqrt{2}$ and $c + d\sqrt{2}$ of this set does not belong to this set, then our usual product will not be a valid binary operation on this set, and hence we cannot claim that this set is a ring!)

> **Question 2.12.2.** Why should associativity of addition and multiplication and distributivity of multiplication over addition all follow from the fact that this set is contained in $\mathbb{R}$?

Question 2.12.3. Are all other ring axioms satisfied? Check!

Question 2.12.4. You know that $\sqrt{2}$ is not a rational number (see Chapter 1, Exercise 1.42). Why does it follow that if a and b are rational numbers, then $a + b\sqrt{2} = 0$ if and only if both a and b are zero.

See the notes on page 86 (but as always, *after* you have played with this example yourselves!). Also, see the notes on page 90 (in particular, Example 2.141 on page 92) for an explanation of the notation $\mathbb{Q}[\sqrt{2}]$.

Example 2.13. Now let us generalize Example 2.12 above. Let m be *any* rational number. Note that if m is negative, $\sqrt{m}$ will not be a real number but a complex (non-real) number. Let $\mathbb{Q}[\sqrt{m}]$ denote the set of all complex numbers of the form $a+b\sqrt{m}$, where a and b are arbitrary rational numbers. (Of course, if $m \geq 0$, then $\mathbb{Q}[\sqrt{m}]$ will actually be contained in the reals.)

Question 2.13.1. What familiar set of numbers does $\mathbb{Q}[\sqrt{m}]$ reduce to if m is the square of a rational number?

Question 2.13.2. More generally, compare the sets $\mathbb{Q}[\sqrt{m}]$ and $\mathbb{Q}[\sqrt{m'}]$ when m and m' satisfy the relation $m = q^2 m'$ for some rational number q. Are these the same sets?

Question 2.13.3. Under the usual addition and multiplication of complex numbers, does $\mathbb{Q}[\sqrt{m}]$ form a ring? (Follow the same steps as in Example 2.12 above.)

Question 2.13.4. Is it true that if a and b are rational numbers, then $a + b\sqrt{m} = 0$ if and only if $a = b = 0$? (As always, if you claim something is not true, give a counterexample!)

Example 2.14. As a specific example of Example 2.13, take $m = -1$. We get the ring $\mathbb{Q}[\imath]$, which is the set of all complex numbers of the form $a + b\imath$, where a and b are arbitrary rational numbers and $\imath$ stands for $\sqrt{-1}$

Exercise 2.14.1. Show that if a and b are real numbers, then $a + b\imath = 0$ if and only if both a and b are zero. (See the notes on page 87 for a clue.)

Example 2.15. Consider the set of rational numbers q that have the property that when q is written in the reduced form a/b with a, b integers and $\gcd(a,b) = 1$ the denominator b is *odd.* This set is usually denoted by $\mathbb{Z}_{(2)}$, and contains elements like $1/3$, $-5/7$, $6/19$, etc., but does not contain $1/4$ or $-5/62$.

Question 2.15.1. Does $\mathbb{Z}_{(2)}$ contain $2/6$?

Notice that every element of $\mathbb{Z}_{(2)}$ is just a fraction (albeit of a particular kind). We know how to add and multiply two fractions together, so we can use this knowledge to add and multiply any two elements of $\mathbb{Z}_{(2)}$. Here is the punch line: $\mathbb{Z}_{(2)}$ forms a ring under the usual operation of addition and multiplication of fractions! Strange as this ring may seem at first, it plays an important role in number theory.

Question 2.15.2. Check that if you add (or multiply) two fractions in $\mathbb{Z}_{(2)}$ you get a fraction that is not an arbitrary rational number but one that also lives in $\mathbb{Z}_{(2)}$. What role does the fact that the denominators are odd play in ensuring this? (The role of the odd denominators is rather crucial; make sure that you understand it!)

Question 2.15.3. Why do associativity and distributivity follow from the fact that $\mathbb{Z}_{(2)} \subseteq \mathbb{Q}$?

Question 2.15.4. Do the other ring axioms hold? Check!

Question 2.15.5. Can you generalize this construction to other subsets of $\mathbb{Q}$ where the denominators have analogous properties?

(See the notes on page 87 for some comments.)

Example 2.16. The set of $n \times n$ matrices with entries in $\mathbb{R}$ ($M_n(\mathbb{R})$), where n is a positive integer, forms a ring with respect to the usual operations of matrix addition and multiplication. For almost all values of n, matrix multiplication is not commutative.

Question 2.16.1. What is the exception?

Checking associativity of addition and multiplication and the distributivity of multiplication over addition is tedious, but you should check at least one of them so as to be familiar with the process.

Exercise 2.16.1. For example, prove that for any three matrices A, B, and C, $(A + B) + C = A + (B + C)$.

What is important is that you get a feel for how associativity and distributivity in $M_n(\mathbb{R})$ *derives* from the fact that associativity and distributivity hold for $\mathbb{R}$.

Question 2.16.2. What about the ring axioms other than associativity and distributivity: do they hold?

Question 2.16.3. What are the additive and multiplicative identities?

Question 2.16.4. Let $e_{i,j}$ denote the matrix with 1 in the (i, j)-th slot and 0 everywhere else. Study the case of 2×2 matrices and guess at a formula for the product $e_{i,j} \cdot e_{k,l}$. (You need not try to prove formally that your formula is correct, but after you have made your guess, substitute various values for i, j, k, and l and test your guess.)

Question 2.16.5. Would the ring axioms still be satisfied if we only considered the set of $n \times n$ matrices whose entries came from $\mathbb{Q}$? From $\mathbb{Z}$?

Question 2.16.6. Now suppose R is any ring. Let us consider the set $M_n(R)$ of $n \times n$ matrices with entries in R with the usual definitions of matrix addition and multiplication. Is $M_n(R)$ with these operations a ring? What if R is not commutative? Does this affect whether $M_n(R)$ is a ring or not?

(See the notes on page 88 for some hints.)

Example 2.17. $\mathbb{R}[x]$, the set of polynomials in one variable with coefficients from $\mathbb{R}$, forms a ring with respect to the usual operations of polynomial addition and multiplication. (We have considered this before.) Here, x denotes the variable. Of course, one could use *any* letter to represent the variable. For instance, one could refer to the variable as t, in which case the set of polynomials with coefficients in $\mathbb{R}$ would be denoted by $\mathbb{R}[t]$. Sometimes, to emphasize our choice of notation for the variable, we refer to $\mathbb{R}[x]$ as the set of polynomials *in the variable* x with coefficients in $\mathbb{R}$, and we refer to $\mathbb{R}[t]$ as the set of polynomials *in the variable* t with coefficients in $\mathbb{R}$. Both $\mathbb{R}[x]$ and $\mathbb{R}[t]$, of course, refer to the same set of objects. Likewise, we often write $f(x)$ (or $f(t)$) for a polynomial, rather than just "f," to emphasize that the variable is x (or t).

If $f(x) = a_0 + a_1x + a_2x^2 + \cdots$ is a *nonzero* polynomial in $\mathbb{R}[x]$, the *degree* of $f(x)$ is the largest value of n for which $a_n \neq 0$, a_nx^n is known as the *highest term*, and a_n is known as the *highest coefficient*. Thus, the polynomials of degree 0 are precisely the nonzero constants. Polynomials of degree 1 are called *linear*, polynomials of degree 2 are called *quadratic*, polynomials of degree 3 are called *cubic*, and so on. *Note that we have not defined the degree of the zero polynomial.* This is on purpose—it will be convenient for the formulation of certain theorems if the zero polynomial does not have a degree!

It is worth recalling an elementary property of polynomials that we will use frequently (in fact, in a more formal treatment of polynomials, this fact is built into the definitions of polynomials): two polynomials are equal if and only if their coefficients are equal. That is, $\sum f_ix^i = \sum g_ix^i$ if and only if $f_i = g_i$ $(i = 0, 1, \dots)$. In particular, a polynomial $\sum f_ix^i$ equals 0 if and only if each $f_i = 0$.

> **Exercise 2.17.1.** Now just as with Example 2.16, prove that if f, g, and h are any three polynomials in $\mathbb{R}[x]$, then $(f + g) + h = f + (g + h)$. Your proof should invoke the fact that associativity holds in $\mathbb{R}$.

Example 2.18. Instead of polynomials with coefficients from $\mathbb{R}$, we can consider polynomials in the variable x with coefficients from an arbitrary ring R, with the usual definition of addition and multiplication of polynomials. We get a ring, denoted $R[x]$. Thus, if we were to consider polynomials in the variable x whose coefficients are all integers, we get the ring $\mathbb{Z}[x]$.

> **Question 2.18.1.** As always, convince yourself that for a general ring R, the set of polynomials $R[x]$ forms a ring. For arbitrary R, is $R[x]$ commutative?

(See the notes on page 88 for some hints and more remarks.)

Example 2.19. Generalizing Example 2.17, the set $\mathbb{R}[x, y]$ of polynomials in two variables x and y, forms a ring. A polynomial in x and y is of the form $\sum_{i,j} f_{i,j}x^iy^j$. (For example, consider the polynomial $4+2x+3y+x^2y+5xy^3$– here, $f_{0,0}$ is the coefficient of x^0y^0, i.e., the coefficient of 1, so $f_{0,0} = 4$. Similarly, $f_{1,3}$ is the coefficient of x^1y^3, so it equals 5. On the other hand, $f_{1,1}$ is zero, since there is no xy term.) Two polynomials $\sum_{i,j} f_{i,j}x^iy^j$ and $\sum_{i,j} g_{i,j}x^iy^j$ are equal if and only if for each pair (i, j), $f_{i,j} = g_{i,j}$.

In the same manner, we can consider $\mathbb{R}[x_1, \ldots, x_n]$, the set of polynomials in n variables x_1, $\ldots$, x_n with coefficients in $\mathbb{R}$. These too form a ring. More generally, if R is any ring we may consider $R[x_1, \ldots, x_n]$, the set of polynomials in n variables x_1, $\ldots$, x_n with coefficients in R. Once again, we get a ring.

Example 2.20. Here is a ring with only two elements! Divide the integers into two sets, the even integers and the odd integers. Let $[0]_2$ denote the set of even integers, and let $[1]_2$ denote the set of odd integers. (Notice that $[0]_2$ and $[1]_2$ are precisely the equivalence classes of $\mathbb{Z}$ under the equivalence relation defined by $a \sim b$ iff $a-b$ is even.) Denote by $\mathbb{Z}/2\mathbb{Z}$ the set $\{[0]_2, [1]_2\}$. Each element of $\{[0]_2, [1]_2\}$ is itself a set containing an infinite number of integers, but we will ignore this fact. Instead, we will view all the even integers together as one "number" of $\mathbb{Z}/2\mathbb{Z}$, and we will view all the odd integers together as another "number" of $\mathbb{Z}/2\mathbb{Z}$. How should we add and multiply these new numbers? Recall that if we add two even integers we get an even integer, if we add an even and an odd integer we get an odd integer, and if we add two odd integers we get an even integer. This suggests the following addition rules in $\mathbb{Z}/2\mathbb{Z}$:

"+"	$[0]_2$	$[1]_2$
$[0]_2$	$[0]_2$	$[1]_2$
$[1]_2$	$[1]_2$	$[0]_2$

(There is an obvious way to interpret this table: if you want to know what "a" + "b" is, you go to the cell corresponding to row a and column b.) Similarly, we know that the product of two even integers is even, the product of an even integer and an odd integer is even, and the product of two odd integers is odd. This gives us the following multiplication rules:

"·"	$[0]_2$	$[1]_2$
$[0]_2$	$[0]_2$	$[0]_2$
$[1]_2$	$[0]_2$	$[1]_2$

Later in this chapter (see Example 2.83 and the discussions preceding that example), we will interpret the ring $\mathbb{Z}/2\mathbb{Z}$ differently: as a *quotient ring* of $\mathbb{Z}$. This interpretation, in particular, will prove that $\mathbb{Z}/2\mathbb{Z}$ is indeed a ring under the given operations. Just accept for now the fact that we get a ring, and play with the it to develop a feel for it.

> **Question 2.20.1.** How would you get a ring with three elements in it? With four?

Example 2.21. Here is the answer to the previous two questions! We have observed that $[0]_2$ and $[1]_2$ are just the equivalence classes of $\mathbb{Z}$ under the equivalence relation $a \sim b$ iff $a - b$ is even. Analogously, let us consider the equivalence classes of $\mathbb{Z}$ under the equivalence relation aRb iff $a - b$ is divisible by 3. Since $a - b$ is divisible by 3 exactly when a and b each leaves the same remainder when divided by 3, there are three equivalence classes: (i) $[0]_3$, the set of all those integers that yield a remainder of 0 when you divide them by 3. In other words, $[0]_3$ consists of all multiples of 3, that is, all integers of the form $3k$, $k \in \mathbb{Z}$. (ii) $[1]_3$ for the set of all those integers that yield a remainder of 1, so $[1]_3$ consists of all integers of the form $3k+1$, $k \in \mathbb{Z}$. (iii) $[2]_3$ for the set of all those integers that yield a remainder of 2, so $[2]_3$ consists of all integers of the form $3k+2$, $k \in \mathbb{Z}$. Write $\mathbb{Z}/3\mathbb{Z}$ for the set $\{[0]_3, [1]_3, [2]_3\}$. Just as in the case of $\mathbb{Z}/2\mathbb{Z}$, every element of this set is itself a set consisting of an infinite number of integers, but we will ignore this fact. How would you add two elements of this set? In $\mathbb{Z}/2\mathbb{Z}$, we defined addition using observations like "an odd integer plus an odd integer gives you an even integer." The corresponding observations here are "an integer of the form $3k+1$ plus another integer of the form $3k+1$ gives you an integer of the form $3k+2$," "an integer of the form $3k+1$ plus another integer of the form $3k+2$ gives you an integer of the form $3k$," "an integer of the form $3k+2$ plus another integer of the form $3k+2$ gives you an integer of the form $3k+1$," etc. We thus get the following addition table:

"+"	$[0]_3$	$[1]_3$	$[2]_3$
$[0]_3$	$[0]_3$	$[1]_3$	$[2]_3$
$[1]_3$	$[1]_3$	$[2]_3$	$[0]_3$
$[2]_3$	$[2]_3$	$[0]_3$	$[1]_3$

Exercise 2.21.1. Similarly, study how the remainders work out when we multiply two integers. (For instance, we find that "an integer of the form $3k+2$ times an integer of the form $3k+2$ gives you an integer of the form $3k+1$," etc.) Derive the following multiplication table:

"·"	$[0]_3$	$[1]_3$	$[2]_3$
$[0]_3$	$[0]_3$	$[0]_3$	$[0]_3$
$[1]_3$	$[0]_3$	$[1]_3$	$[2]_3$
$[2]_3$	$[0]_3$	$[2]_3$	$[1]_3$

This process can easily be generalized to yield a ring with n elements ($\mathbb{Z}/n\mathbb{Z}$) for any $n \geq 2$.

Exercise 2.21.2. Construct the addition and multiplication tables for the ring $\mathbb{Z}/4\mathbb{Z}$.

Example 2.22. Suppose R and S are two rings. (For example, take $R = \mathbb{Z}/2\mathbb{Z}$, and take $S = \mathbb{Z}/3\mathbb{Z}$.) Consider the Cartesian product $T = R \times S$, which is the set of ordered pairs (r, s) with $r \in R$ and $s \in S$. Define addition in T by $(r, s)+(r', s') = (r+r', s+s')$. Here, "$r+r'$" refers to the addition of two elements of R according to the defintion of addition in R, and similarly, "$s+s'$" refers to the addition of two elements of S according to the definition of addition in S. For instance, in $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$, $([0]_2, [1]_3) + ([1]_2, [2]_3) = ([1]_2, [0]_3)$. Similarly, define multiplication in T by $(r, s)\cdot(r', s') = (r\cdot r', s\cdot s')$. Once again, "$r\cdot r'$" refers to the multiplication of two elements of R according to the definition of multiplication in R, and "$s\cdot s'$" refers to the multiplication of two elements of S according to the definition of multiplication in S. Thus, in $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$ again, $([0]_2, [1]_3) \cdot ([1]_2, [2]_3) = ([0]_2, [2]_3)$.

> **Question 2.22.1.** Do these definitions of addition and mulitplication make T a ring? Check!

Definition 2.23. Given two rings R and S, the ring T constructed above is known as the *direct product* of R and S.

> **Question 2.23.1.** What are the identity elements with respect to addition and multiplication?
>
> **Question 2.23.2.** Now take $R = S = \mathbb{Z}$. Can you find pairs of *nonzero* elements a and b in the ring $T = \mathbb{Z}\times\mathbb{Z}$ such that $a\cdot b = 0$? (Note that $\mathbb{Z}$ itself does not contain pairs of such elements.) If R and S are arbitrary rings, can you find a pair of nonzero elements a and b in $T = R \times S$ such that $a \cdot b = 0$?

(See the notes on page 89 for hints.)

Remark 2.24. The examples above should have convinced you that our definition of a ring (Definition 2.5 above) is rather natural, and that it very effectively models several number systems that arise in mathematics. Here is further evidence that our axioms are the "correct" ones. Notice that in all the rings that we have come across, the following properties hold:

1. The additive identity is unique, that is, there is precisely one element 0 in the ring that has the property that $a + 0 = a$ for all elements a in the ring.

2. The multiplicative identity is unique, that is, there is precisely one element 1 in the ring that has the property that $a \cdot 1 = 1 \cdot a = a$ for all elements a in the ring.

3. $a + b = a + c$ implies $b = c$ for any elements a, b, and c in the ring.

4. For every element a in the ring, there is precisely one element $-a$ that satisfies the condition that $a + (-a) = 0$.

5. For every element a in the ring, $-(-a)$ is just a.

6. $a \cdot 0 = 0 \cdot a = 0$ for all elements a.

7. $(-1) \cdot a = a \cdot (-1) = -a$ for all elements a.

8. More generally, $a \cdot (-b) = (-a) \cdot b = -(ab)$ for all elements a and b.

9. $(-1) \cdot (-1) = 1$.

10. More generally, $(-a) \cdot (-b) = ab$ for all elements a and b.

Now these properties all seem extremely natural, and we would certainly like them to hold in all rings. (More strongly, a ring in which any of these properties fail would appear very pathological to us!) Now, if our ring axioms were the "correct" ones, then the properties above would be deducible *from the ring axioms themselves,* thereby showing that they hold in all rings. As it turns out, this is indeed true: they *are* deducible from the axioms, and therefore, they *do* hold in every ring R. Although we will not verify this in the text, it is good practice for you to verify that at least some of these properties above follow from the axioms, so we have included the verification as Exercise 2.114 in the exercises at the end of this chapter (see also the remarks on page 89).

Property 3 above is known as *additive cancellation.* It is actually a consequence of the fact that if R is any ring, then $(R, +)$ is a group: *in any group* $(G, *)$*,* if $a * b = a * c$ for elements a, b, and c in G then $b = c$ (see

Exercise 4.18 in Chapter 4 ahead). In much the same way, properties 1, 4, and 5 are really consequences of the fact that these properties hold for any group. (See Exercises 4.15, 4.16, and 4.17 in Chapter 4.)

In fact, after you do Exercise 2.115 at the end of this chapter, you will realize that property 2 above is also a property that comes from a particular group structure on a particular subset of R!

Notice that there is one property that is very similar to additive cancellation, namely *multiplicative cancellation*: $a \cdot b = a \cdot c$ implies $b = c$, which we have not listed above. The reason for its absence is very simple: multiplicative cancellation *cannot* be deduced from the ring axioms. In turn, the reason that it cannot be deduced from the axioms is because *multiplicative cancellation does not hold in all rings!*

Question 2.25. Can you think of an example of a ring R and elements a, b, and c in R such that $ab = ac$ yet $b \neq c$?

2.2 Subrings

In Examples 2.12, 2.13, 2.14, and 2.15 above, we came across the following phenomenon: A ring R and a subset S of R that had the following two properties: For any s_1 and s_2 in S, $s_1 + s_2$ was in S and $s_1 s_2$ was in S. In Example 2.12, the ring R was $\mathbb{R}$, and the subset S was the set of all real numbers of the form $a + b\sqrt{2}$ with a and b rational numbers. In Example 2.14, R was $\mathbb{C}$ and S was the set of all complex numbers of the form $a + bi$ with a and b rational numbers. In Example 2.15, R was $\mathbb{Q}$, and S was the set of all reduced fractions with odd denominator. Moreover, in all three examples, we endowed S with binary operations in the following way: Given s_1 and s_2 in S, we viewed them as elements of R, and formed the sum $s_1 + s_2$ (the sum being defined according to the definition of addition in R). Next, we observed that $s_1 + s_2$ was actually in S (this is one of the two properties alluded to above). Similarly, we observed that $s_1 s_2$ (the product being formed according to the definition of multiplication in R) was also in

S. These two facts hence gave us two binary operation on S. We then found that with respect to these binary operations, S was not just an arbitrary subset of R, it was actually a ring in its own right.

The crucial reason (although not the only reason) why the set S in all our examples was itself a ring was that S had the properties described at the beginning of the previous paragraph. We give these properties a name.

Definition 2.26. Given an arbitrary nonempty subset S of a ring R, we say that S is *closed under addition* if for any s_1 and s_2 in S, $s_1 + s_2$ is also in S. Similarly, we say that S is *closed under multiplication* if for any s_1 and s_2 in S, $s_1 s_2$ is also in S.

As we have observed, if a subset S of a ring R is closed under addition, then the addition operation on R, when restricted to ordered pairs of elements of S, yields a binary operation on S (which we also call addition), and we say that the addition on S is *induced* by the addition on R. Similarly, when S is closed under multiplication, we get a binary operation on S (also called multiplication) that we say is *induced* by the multiplication on R.

Now suppose that S is a subset of a ring R that is closed with respect to addition and multiplication, and just as in our examples above, suppose that with respect to the induced operations, S is itself a ring. We will give a special name to this situation:

Definition 2.27. Let S be a subset of a ring R that is closed with respect to addition and multiplication. Suppose that $1 \in S$. Suppose further that with respect to these addition and multiplication operations on S that are induced from those on R, S is itself a ring. We say that S is a *subring* of R. We also describe R as a *ring extension* of S, and refer to R and S jointly as the *ring extension* R/S.

Examples 2.12, 2.13, 2.14, and 2.15 above are therefore all instances of subrings: $\mathbb{Q}[\sqrt{2}]$ is a subring of $\mathbb{R}$, $\mathbb{Q}[\imath]$ is a subring of $\mathbb{C}$, and $\mathbb{Z}_{(2)}$ is a subring of $\mathbb{Q}$. (See the notes on page 90 for a remark on Definition 2.27 above.)

Question 2.28. Consider the subset S of $\mathbb{Z}$ consisting of the positive even integers, that is, the set $\{2n|n \in \mathbb{Z}$ and $n > 0\}$. Check that S is closed with respect to both addition and multiplication. Does this make S a subring of $\mathbb{Z}$? Next, consider the set T of all nonnegative integers. Check that T is also closed with respect to addition and multiplication. Clearly, T contains 1. Does this make T a subring of $\mathbb{Z}$?

Here is a quick exercise, which is really a special case of Exercise 4.22 in Chapter 4 ahead:

Exercise 2.29. Let S be a subring of the ring R. Thus, by definition $(S, +)$ is an abelian group. Let 0_S denote the identity element of this group, and write 0_R for the usual "0" of R. Show that $0_S = 0_R$. (See also Exercise 3.54 in Chapter 3 ahead.)

Before we proceed to look at further examples of subrings, let us first consider a criterion that will help us decide whether a given subset of a ring is actually a subring.

Lemma 2.30. *Let S be a subset of a ring R which has the following properties:*

1. *S is closed under addition,*
2. *S is closed under multiplication,*
3. *1 is in S, and*
4. *For all $a \in S$, $-a$ is also in S.*

Then S is a subring of R.

Proof. As discussed above, since S is closed with respect to addition and multiplication, the addition and multiplication operations on R induce addition and multiplication operations on S. Now consider addition. For any a, b, and c in S, we may view a, b, and c as elements of R, and since addition is associative in R, we find $(a + b) + c = a + (b + c)$. Viewing a, b, and c back as elements of S in this equation, we find that the induced addition operation on S is associative. Similarly, since addition is commutative in R, the induced addition on S is commutative. Now we are given that $1 \in S$, so property (4) shows that -1 is also in S. From the fact that S is closed under addition, we find that $1 + (-1)$ is also in S, so 0 is in S. The relation $s + 0 = s$ holds for all $s \in S$, since it holds more generally for any $s \in R$. Thus, S has an additive identity, namely 0. For every $s \in S$, we are given that $-s$ is also in S, so every element of S has an additive inverse. As for multiplication, given a, b, and c in S, we may view these as elements of R, and since multiplication in R is associative, we find that $(ab)c = a(bc)$. As before, viewing a, b, and c back as elements of S in this equation, we find that the induced multiplication operation on S is associative. Since $s \cdot 1 = 1 \cdot s = s$ for all $s \in S$ (as this is true more generally for all $s \in R$), and since $1 \in S$, we find that S has a multiplicative identity, namely 1. Finally, exactly as in the arguments for associativity above, the relations $a(b+c) = ab+ac$ and $(a+b)c = ac+bc$ hold for all a, b, and c in S because they hold in R, so distributivity is satisfied. S is hence a ring in its own right with respect to the induced operations of addition and multiplication and it contains 1. Thus, S is a subring of R. □

The following are further examples of subrings. Play with these examples to gain familiarity with them. *Check that they are indeed examples of subrings of the given rings by applying Lemma 2.30.*

Example 2.31. The set of all real numbers of the form $a + b\sqrt{2}$ where a and b are *integers* is a subring of $\mathbb{Q}[\sqrt{2}]$. Why? It is denoted by $\mathbb{Z}[\sqrt{2}]$.

Example 2.32. The set of all complex numbers of the form $a + bi$ where a and b are integers is a subring of $\mathbb{Q}[\imath]$. It is denoted by $\mathbb{Z}[\imath]$. (It is often called the ring of *Gaussian integers.*)

Example 2.33. Let $\mathbb{Z}[1/2]$ denote the set of all rational numbers that are such that when written in the reduced form a/b with $\gcd(a, b) = 1$, the denominator b is a power of 2. (Contrast this set with $\mathbb{Z}_{(2)}$.) This is a subring of $\mathbb{Q}$.

> **Question 2.33.1.** What are the rational numbers that this ring has in common with $\mathbb{Z}_{(2)}$?
> (See the notes on page 90 for clues.)

Example 2.34. Let $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$ denote the set of all real numbers of the form $a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6}$, where a, b, c, and d are all rational numbers. This is a subring of the reals. (See the notes on page 93, in particular see Example 2.146 there, for an explanation of this notation.)

> **Question 2.34.1.** Is the set of all real numbers of the form $a + b\sqrt{2} + c\sqrt{3}$ where a, b, and c are rationals a subring of the reals?

Example 2.35. If S is a subring of a ring R, then $S[x]$ is a subring of $R[x]$.

> **Exercise 2.35.1.** Prove this assertion!

In particular, this shows that $\mathbb{Q}[x]$ is a subring of $\mathbb{R}[x]$, which in turn is a subring of $\mathbb{C}[x]$.

Example 2.36. Similarly, If S is a subring of a ring R, then $M_n(S)$ is a subring of $M_n(R)$.

Example 2.37. Let $U_n(\mathbb{R})$ denote the *upper triangular matrices*, that is, the subset of $M_n(\mathbb{R})$ consisting of all matrices whose entries below the main diagonal are all zero. Thus, $U_n(\mathbb{R})$ is the set of all $(a_{i,j})$ in $M_n(\mathbb{R})$ with $a_{i,j} = 0$ for $i > j$. (You may have seen the notation "$(a_{i,j})$" before: it denotes the matrix whose entry in the ith row and jth column is the element $a_{i,j}$.) Then $U_n(\mathbb{R})$ is a subring of $M_n(\mathbb{R})$.

> **Question 2.37.1.** Why?

> **Question 2.37.2.** For what values of n will $U_n(\mathbb{R})$ be the same as $M_n(\mathbb{R})$?

Question 2.37.3. Suppose we considered the set of *strictly upper triangular matrices*, namely the set of all $(a_{i,j})$ in $M_n(\mathbb{R})$ with $a_{i,j} = 0$ for $i \leq j$. Would we still get a subring of $M_n(\mathbb{R})$?

Example 2.38. Here is another subring of $M_n(\mathbb{R})$. For each real number r, let $\mathrm{diag}(r)$ denote the matrix in which each diagonal entry is just r and in which the off-diagonal entries are all zero. The set of matrices in $M_n(\mathbb{R})$ of the form $\mathrm{diag}(r)$ (as r ranges through $\mathbb{R}$) is then a subring.

Question 2.38.1. What observations can you make about the function from $\mathbb{R}$ to $M_n(\mathbb{R})$ that sends r to $\mathrm{diag}(r)$?
(See Example 2.106 ahead.)

2.3 Integral Domains and Fields

In passing from the concrete example of the integers to the abstract definition of a ring, observe that we have introduced some phenomena that at first seem pathological. The first, which we have already pointed out explicitly and is already present in $M_2(\mathbb{R})$, is noncommutativity of multiplication. The second, which is also present in $M_2(\mathbb{R})$, and examples of which you have seen as far back as in the preliminary chapter To the Student, page ix, is the existence of zero-divisors.

Definition 2.39. A *zero-divisor* in a ring R is a nonzero element a for which there exists a nonzero element b such that either $a \cdot b = 0$ or $b \cdot a = 0$.

Just as noncommutativity of multiplication, on closer observation, turns out to be quite a natural phemomenon after all, the existence of zero-divisors is really not very pathological either. It merely *seems* so because most of our experience has been restricted to various rings that appear as subrings of the complex numbers.

Besides matrix rings (try to discover lots of zero-divisors in $M_2(\mathbb{R})$ for yourselves), zero-divisors occur in several rings that arise naturally in mathematics, including many *commutative* ones. For instance, the direct product of two rings always contains zero-divisors (see Example 2.22 above). Also, (see Exercise 2.21.2), $\mathbb{Z}/4\mathbb{Z}$ contains zero-divisors: $[2]_4 \cdot [2]_4 = [0]_4$! In fact, as long as n is not prime, you should be able to discover zero-divisors in any of the rings $\mathbb{Z}/n\mathbb{Z}$ (see 2.58 ahead). (It can be proved, however, that $\mathbb{Z}/n\mathbb{Z}$ *cannot* have zero-divisors if n is prime, see 2.59 ahead.)

On the other hand, there is no doubt that the absence of zero-divisors in a ring indeed makes the ring relatively easy to work with. If, in addition, such a ring is also commutative, it becomes *exceptionally* nice to work with. With this in mind, we make the following definition:

Definition 2.40. An integral domain is a commutative ring with no zero-divisors.

(Alternatively, an integral domain is a commutative ring R with the property that whenever $a \cdot b = 0$ for two elements a and b in R, then either a must be 0 or else b must be 0.)

$\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are all obvious examples of integral domains. (Again, we are simply invoking our knowledge of these rings when we make this claim.)

> **Question 2.41.** Is $\mathbb{R}[x]$ an integral domain? More generally, if R is an arbitrary ring, can you determine necessary and sufficient conditions on R that will guarantee that $R[x]$ has no zero-divisors? (See the notes on page 88 for a definition of $R[x]$, and for some discussions that may help you answer this question.)

Notice that any subring S of an integeral domain R must itself be an integral domain. (If $ab = 0$ holds in S for some nonzero elements a and b, then viewing a and b as elements of R, we would find $ab = 0$ in R, which is a contradiction, since R is an integral domain.) In particular, *any subring of* $\mathbb{C}$ *is an integral domain.*

> **Question 2.42.** Now suppose S is a subring of R and suppose that S (note!) is an integral domain. Must R also be an integral domain? (Hint: Look at Example 2.38 above for inspiration!)

Integral domains have one nice property: one can always cancel elements from both sides of an equation, i.e., multiplicative cancellation holds! More precisely, we have the following:

Lemma 2.43. *(Multiplicative Cancellation in Integral Domains:) Let R be an integral domain, and let a be a nonzero element of R. If $ab = ac$ for two elements b and c in R, then $b = c$.*

Proof. Write $ab = ac$ as $a(b - c) = 0$. Since $a \neq 0$ and since R is an integral domain, $b - c = 0$, or $b = c$! □

Now, integral domains are definitely very nice rings, but one can go out on a limb and require that rings be even nicer! We can require that we be able to *divide* any element a by any nonzero element b. This would certainly make the ring behave much more like $\mathbb{Q}$ or $\mathbb{R}$.

To understand division better, let us look at the process of dividing two integers a little closer. To divide 3 by 5 is really to multiply together 3 and $1/5$ (just as to subtract, say, 6 from 9 is really to add together 9 and -6). The reason this cannot be done within the context of the integers is that $1/5$ is not an integer. (After all, if $1/5$ were an integer, then the product of 3 and $1/5$ would also be an integer.) Now let us look at $1/5$ a different way. $1/5$ has the property that $1/5 \cdot 5 = 5 \cdot 1/5 = 1$. In other words, $1/5$ *is the inverse of* 5 *with respect to multiplication* (just as -6 is the inverse of 6 with respect to addition). First, let us pause to give a name to this:

Definition 2.44. If R is an *arbitrary* ring, a nonzero element a is said to be *invertible* or to have a *multiplicative inverse* if there exists an element $b \in R$ such that $ab = ba = 1$. In such a situation, b is known as the multiplicative inverse of a, and a is known as the multiplicative inverse of b.

Invertible elements of a ring R are also known as *units* of R.

(Notice that for an arbitrary ring, it is not enough in the definition of invertibility to insist that $ab = 1$, we also need ba to equal 1. It is certainly possible to have two elements a and b in a ring R such that $ab = 1$ but $ba \neq 1$: see Exercise 3.103 in Chapter 3 ahead.)

Question 2.45. What are the units of $\mathbb{Z}$?

Putting all this together, the reason that we cannot divide within the context of the integers is that given an arbitrary (nonzero integer) m, it need not be invertible. With this in mind, we have the following definition:

Definition 2.46. A *field* is an integral domain in which every nonzero element a is invertible. The multiplicative inverse of a nonzero element a is usually denoted either by "$1/a$" or by "a^{-1}."

(For a comment on this definition, see the notes on page 93.)

Here is a quick exercise:

> **Exercise 2.47.** Let R be a commutative ring. Suppose that every nonzero element in R is invertible. Prove that R cannot have zero-divisors, and hence, R must a field. Give an example of a commutative ring to show that, conversely, if R is commutative and has no zero-divisors, then all nonzero elements *need not* be invertible.

We will often use the letter F to denote a field. The set of nonzero elements of a field F is often denoted by F^*.

> **Question 2.48.** If F is a field, is F^* a group with respect to multiplication?
> (See also Exercise 2.115 at the end of the chapter.)

Remark 2.49. Notice that 0 can never have a multiplicative inverse, since $a \cdot 0 = 0$ for any a. (See Remark 2.24.) We describe this by saying that *division by* 0 *is not defined.*

Perhaps the most familiar example of a field is $\mathbb{Q}$. We have already seen that it is a ring (Example 2.10) The multiplicative inverse of the nonzero rational number m/n is, of course, n/m. Here are more examples:

Example 2.50. The reals, $\mathbb{R}$.

Example 2.51. $\mathbb{Q}[\sqrt{2}]$.

> **Question 2.51.1.** $\mathbb{Q}[\sqrt{2}]$ is a subring of $\mathbb{R}$, and hence an integral domain. Explicitly exhibit the multiplicative inverse of the nonzero number $a + b\sqrt{2}$ as $c + d\sqrt{2}$ for suitable rational numbers c and d. (Think in terms of "rationalizing" denominators.)

Question 2.51.2. Is $\mathbb{Z}[\sqrt{2}]$ a field?

Example 2.52. The complex numbers, $\mathbb{C}$.

Question 2.52.1. What is the inverse of the nonzero number $a + \imath b$? (Give the inverse as $c + \imath d$ for suitable real numbers c and d: think in terms of "real-izing" denominators.)

Example 2.53. $\mathbb{Q}[\imath]$.

Question 2.53.1. Why is $\mathbb{Q}[\imath]$ a field?

Question 2.53.2. Is $\mathbb{Z}[\imath]$ a field?

Example 2.54. Here is a new example: the set of rational functions with coefficients from the reals, $\mathbb{R}(x)$. (Note the parentheses around the x.) This is the set of all quotients of polynomials with coefficients from the reals, that is, the set $\left\{\dfrac{f(x)}{g(x)}\right\}$, where $f(x)$ and $g(x)$ are elements of $\mathbb{R}[x]$, and $g(x) \neq 0$. (Of course, we take $f(x)/g(x) = f'(x)/g'(x)$ if $f(x)g'(x) = g(x)f'(x)$.) Addition and multiplication in $\mathbb{R}(x)$ are similar to addition and multiplication in $\mathbb{Q}$—

$$\frac{f_1(x)}{g_1(x)} + \frac{f_2(x)}{g_2(x)} = \frac{f_1(x) \cdot g_2(x) + f_2(x) \cdot g_1(x)}{g_1(x) \cdot g_2(x)},$$

and

$$\frac{f_1(x)}{g_1(x)} \cdot \frac{f_2(x)}{g_2(x)} = \frac{f_1(x) \cdot f_2(x)}{g_1(x) \cdot g_2(x)}.$$

The multiplicative inverse of the nonzero element $\dfrac{f(x)}{g(x)}$ is just $\dfrac{g(x)}{f(x)}$.

Example 2.55. More generally, if F is *any* field, we may consider the set of rational functions with coefficients from F, denoted $F(x)$. This is analogous to $\mathbb{R}(x)$: it is the set $\left\{\dfrac{f(x)}{g(x)}\right\}$, where $f(x)$ and $g(x)$ are now elements of $F[x]$ instead of $\mathbb{R}[x]$, and $g(x) \neq 0$. (As with $\mathbb{R}(x)$, we take $f(x)/g(x) = f'(x)/g'(x)$ if $f(x)g'(x) = g(x)f'(x)$.) Addition and multiplication are defined just as in $\mathbb{R}(x)$, and we can check that we get a field.

Example 2.56. The ring $\mathbb{Z}/2\mathbb{Z}$ is a field! This is easy to see from the multiplication table for $\mathbb{Z}/2\mathbb{Z}$ (see page 36): since the only nonzero element is $[1]_2$, and since $[1]_2 \cdot [1]_2 = [1]_2 \neq [0]_2$, it is clear that $\mathbb{Z}/2\mathbb{Z}$ is an integral domain. But we can read one more fact from the relation $[1]_2 \cdot [1]_2 = [1]_2$: the only nonzero element in $\mathbb{Z}/2\mathbb{Z}$ is actually invertible! Thus, $\mathbb{Z}/2\mathbb{Z}$ is indeed a field.

> **Question 2.56.1.** How many elements does the ring $M_2(\mathbb{Z}/2\mathbb{Z})$ have? Which of these elements are invertible?
> (It would be helpful to recall from more elementary courses that a matrix with entries in, say the real numbers, is invertible if and only if its determinant is nonzero. You may accept for now that this same result holds for matrices with entries in *any* field.)

Example 2.57. The ring $\mathbb{Z}/3\mathbb{Z}$ is also a field!

> **Question 2.57.1.** Study the multiplication table in $\mathbb{Z}/3\mathbb{Z}$ on page 37. There are no zeros in the table other than in the first row and in the first column (which correspond to multiplication by zero). Why does this show that there are no zero-divisors in this ring? Now notice that every row and every column (other than the first) has $[1]_3$ in it. Why does this show that every nonzero element is invertible?

(After this example and the previous example of $\mathbb{Z}/2\mathbb{Z}$, you may find Exercise 2.127 at the end of the chapter illuminating.)

Example 2.58. It would be tempting to jump to the conclusion that $\mathbb{Z}/m\mathbb{Z}$ is a field for all $m \geq 2$. However, we have already seen on page 45 that $\mathbb{Z}/4\mathbb{Z}$ has a zero-divisor. This shows that $\mathbb{Z}/4\mathbb{Z}$ is not an integral domain, and hence most definitely not a field.

> **Question 2.58.1.** Study the observation on page 45 that shows that $\mathbb{Z}/4\mathbb{Z}$ is not an integral domain. How should you generalize this observation to prove that for any *composite* integer $m \geq 4$, $\mathbb{Z}/m\mathbb{Z}$ is not an integral domain?

Example 2.59. However, Examples 2.56 and 2.57 do generalize suitably: it turns out that for any *prime* p, the ring $\mathbb{Z}/p\mathbb{Z}$ is a field (with p elements).

Recall from the discussions in Examples 2.20 and 2.21 that the elements of $\mathbb{Z}/p\mathbb{Z}$ are equivalence classes of integers under the relation $a \sim b$ if and only if $a - b$ is divisible by p. The equivalence class $[a]_p$ of an integer a is thus the set of integers of the form $a \pm p$, $a \pm 2p$, $a \pm 3p$, $\ldots$. Addition and multiplication in $\mathbb{Z}/p\mathbb{Z}$ are defined by the rules

1. $[a]_p + [b]_p = [a + b]_p$
2. $[a]_p \cdot [b]_p = [a \cdot b]_p$

Exercise 2.59.1. Show that addition and multiplication are well-defined, that is, if $a \sim a'$ and $b \sim b'$, then $a + b \sim a' + b'$ and $a \cdot b \sim a' \cdot b'$.

Exercise 2.59.2. Show that the zero in this ring is $[0]_p$, and the 1 in this ring is $[1]_p$. (In particular, $[a]_p$ is nonzero in $\mathbb{Z}/p\mathbb{Z}$ precisely when a is not divisible by p.)

Exercise 2.59.3. Now let $[a]_p$ be a nonzero element in $\mathbb{Z}/p\mathbb{Z}$. Show that $[a]_p$ is invertible. (Hint: Invoking the fact that a and p are relatively prime, we find that there must exist integers x and y such that $xa + yp = 1$. So?)

Exercise 2.59.4. Now conclude using Exercise 2.47 and Exercise 2.59.3 above that $\mathbb{Z}/p\mathbb{Z}$ is a field.

We end this section with the concept of a subfield. The idea is very simple (compare with Definition 2.27 above):

Definition 2.60. A subset F of a field K is called a *subfield of* K if F is a subring of K and is itself a field. In this situation, we also describe K as a *field extension* of F, and refer to F and K jointly as the *field extension* K/F.

The difference between being a sub*ring* of K and a sub*field* of K is as follows: Suppose R is a subring of K. Given a nonzero element a in R, its multiplicative inverse $1/a$ certainly exists in K (why?). *However, $1/a$ may not live inside R.* If $1/a$ happens to live inside R, we say that a has a multiplicative inverse in R itself. Now, if every nonzero a in R has a multiplicative inverse in R itself, then by Definition 2.46 (why is R an integral domain?), R is a field. Therefore, by Definition 2.60 above, R is then a subfield of K.

Thus, $\mathbb{Q}$ is a subfield of $\mathbb{R}$, but $\mathbb{Z}$ is only a subring of $\mathbb{R}$; it is not a subfield of $\mathbb{R}$. Similarly, $\mathbb{R}$ is a subfield of $\mathbb{C}$. (Is $\mathbb{R}$ a subring of $\mathbb{C}$?) $\mathbb{Q}[\sqrt{2}]$ is a subfield of $\mathbb{C}$. In fact, more is true—$\mathbb{Q}$ is a subfield of $\mathbb{Q}[\sqrt{2}]$, which in turn is a subfield of $\mathbb{R}$, which in turn is a subfield of $\mathbb{C}$.

Question 2.61. By contrast, is $\mathbb{Q}[i]$ a subfield of $\mathbb{R}$? Of $\mathbb{C}$?

Question 2.62. Is $\mathbb{Z}[i]$ a subfield of $\mathbb{R}$? Of $\mathbb{C}$?

(See Exercise 2.128 at the end of the chapter for a situation in which we can conclude that a subring of a field must actually be a subfield.)

2.4 Ideals

Consider the ring $\mathbb{Z}$, and consider the subset of even integers, denoted (suggestively) $2\mathbb{Z}$. The set $2\mathbb{Z}$ is closed under addition (the sum of two even integers is again an even integer), and in fact, $(2\mathbb{Z}, +)$ is even an abelian group (this is because (i) 0 is an even integer and hence in $2\mathbb{Z}$, (ii) for any even integer a, $-a$ is also an even integer and hence in $2\mathbb{Z}$, (iii) and of course, addition of integers, restricted to $2\mathbb{Z}$ is both an associative and commutative operation). Moreover, the set $2\mathbb{Z}$ has one extra property that will be of interest: for any integer $a \in 2\mathbb{Z}$ and for any *arbitrary* integer m, am is also an even integer and hence in $2\mathbb{Z}$. Subsets such as these play a crucial role in the structure of rings, and are given a special name: they are referred to as *ideals*.

Definition 2.63. Let R be a ring. A subset I of R is called an *ideal* of R if I is closed under the addition operation of R and under this induced binary operation $(I, +)$ is an abelian group, and if for any $i \in I$ and *arbitrary* $r \in R$, both $ri \in I$ and $ir \in I$. An ideal I is called *proper* if $I \neq R$.

Remark 2.64. Of course, if R is commutative, as in the example of $\mathbb{Z}$ and $2\mathbb{Z}$ above, $ri \in I$ if and only if $ir \in I$, but in an arbitrary ring, one must specify in the definition that both ri and ir be in I.

Remark 2.65. Notice in the definition of ideals above that if $ir \in I$ for all $r \in R$, then in particular, taking r to come from I, we find that I must be closed under multiplication as well, that is, for any i and j in I, ij must also be in I. Once we find that ideals are closed under multiplication, the associative and distributive laws will then be inherited from R, so ideals seem like they should be the same as subrings. However, they differ from subrings in one crucial aspect—ideals do not have to contain the multiplicative identity 1. (Recall the definition of subrings, and see the example of $2\mathbb{Z}$ above–it certainly does not contain 1.)

> **Exercise 2.66.** Show that if I is an ideal of a ring R, then $1 \in I$ implies $I = R$.

Here is an alternative characterization of ideals:

Lemma 2.67. *Let I be a subset of a ring R. Then I is an ideal of R if and only if*

1. *I is nonempty,*
2. *I is closed under addition, and*
3. *for all $i \in I$ and $r \in R$, both ir and ri are in I.*

Proof. If I is an ideal of R, then by definition, I is closed under addition, and for all $i \in I$ and $r \in R$, both ir and ri are in I. Moreover, by definition of being an ideal, $(I, +)$ is an abelian group, so it has at least one element (the identity element). This shows that I is nonempty.

Now assume that I is nonempty, closed under addition, and for all $i \in I$ and $r \in R$, both ir and ri are in I. Since I is nonempty, there exists at least one element in I, call it a. Then, by the hypotheses, $a \cdot 0 = 0$ must be in I. Also, for any $i \in I$, $i \cdot (-1) = -i \in I$. Since commutativity and associativity of addition in I follows from that in R, we find that indeed $(I, +)$ is an abelian group. □

Exercise 2.68. If I is an ideal of R, then by definition, $(I, +)$ is an abelian group. Consequently, it has an identity element, call it 0_I, that satisfies the property that $i + 0_I = 0_I + i = i$ for all $i \in I$. On the other hand, the element "0" in R is the identity element for the group $(R, +)$. Prove that the element 0_I must be the same as the element 0.
(See Exercise 4.22 in Chapter 4 ahead.)

The significance of ideals will become clear when we study quotient rings and ring homomorphisms a little ahead, but first let us consider several examples of ideals in rings:

Example 2.69. Convince yourselves that if R is any ring, then both R and the set $\{0\}$ are both ideals of R. The ideal $\{0\}$ is often referred to informally as the *zero ideal*.

Example 2.70. Just as with the set $2\mathbb{Z}$, we may consider, for any integer m, the set of all multiples of m, denoted $m\mathbb{Z}$.

Exercise 2.70.1. Prove that $m\mathbb{Z}$ is an ideal of $\mathbb{Z}$.

Question 2.70.1. What does $m\mathbb{Z}$ look like when $m = 1$?

Question 2.70.2. What does $m\mathbb{Z}$ look like when $m = 0$?

Example 2.71. In the ring $\mathbb{R}[x]$, let $\langle x\rangle$ denote the set of all polynomials that are a *multiple* of x, i.e. the set $\{xg(x) \mid g(x) \in \mathbb{R}[x]\}$.

Exercise 2.71.1. Prove that $\langle x\rangle$ is an ideal of $\mathbb{R}[x]$.

Exercise 2.71.2. More generally, let $f(x)$ be an arbitrary polynomial, and let $\langle f(x)\rangle$ denote the set of all polynomials that are a multiple of $f(x)$, i.e. the set $\{f(x)g(x) \mid g(x) \in \mathbb{R}[x]\}$. Show that $\langle f(x)\rangle$ is an ideal of $\mathbb{R}[x]$.

Example 2.72. In the ring $\mathbb{R}[x, y]$, let $\langle x, y\rangle$ denote the set of all polynomials that can be expressed as $xf(x, y) + yg(x, y)$ for suitable polynomials $f(x, y)$ and $g(x, y)$. For example, the polynomial $x + 2y + x^2y + xy^3$ is in $\langle x, y\rangle$ because it can be rewritten as $x(1 + xy) + y(2 + xy^2)$. (Note that this rewrite is not unique–it can also be written as $x(1 + xy + y^3) + 2y$–but this will not be an issue.)

> **Exercise 2.72.1.** Show that $\langle x, y\rangle$ is an ideal of $\mathbb{R}[x, y]$.
>
> **Exercise 2.72.2.** More generally, given two arbitrary polynomials $p(x, y)$ and $q(x, y)$, let $\langle p(x, y), q(x, y)\rangle$ denote the set of all polynomials that can be expressed as $p(x, y)f(x, y)+q(x, y)g(x, y)$ for suitable polynomials $f(x, y)$ and $g(x, y)$. Show that $\langle p(x, y), q(x, y)\rangle$ is an ideal of $\mathbb{R}[x, y]$.

Example 2.73. Fix an integer $n \geq 1$. In the ring $M_n(\mathbb{Z})$ (see Exercise 2.16.5), the subset $M_n(2\mathbb{Z})$ consisting of all matrices all of whose entries are even, is an ideal.

> **Exercise 2.73.1.** Prove this.
>
> **Question 2.73.1.** Given an arbitrary integer m, is the subset $M_n(m\mathbb{Z})$ consisting of all matrices all of whose entries are a multiple of m an ideal of $M_n(\mathbb{Z})$?

Example 2.74. Let R be an arbitrary ring, and let I be an ideal of R. Fix an integer $n \geq 1$. In $M_n(R)$, let $M_n(I)$ denote the subset of all matrices all of whose entries come from I.

> **Exercise 2.74.1.** Prove that $M_n(I)$ is an ideal of $M_n(R)$.

Example 2.75. In the ring $\mathbb{Z}_{(2)}$, denote by $\langle 2\rangle_{(2)}$ the set of all fractions of the (reduced) form a/b where b is odd and a is even.

> **Question 2.75.1.** Study Example 2.15 carefully. Is $\langle 2\rangle_{(2)}$ a *proper* subset of $\mathbb{Z}_{(2)}$?
>
> **Exercise 2.75.1.** Prove that $\langle 2\rangle_{(2)}$ is an ideal of $\mathbb{Z}_{(2)}$.

Example 2.76. Let R and S be rings, and let I_1 be an ideal of R and I_2 an ideal of S. Let $I_1 \times I_2$ denote the set $\{(a,b) \mid a \in I_1,\ b \in I_2\}$.

> **Exercise 2.76.1.** Prove that $I_1 \times I_2$ is an ideal of $R \times S$.

(See Exercise 2.129 ahead.)

Example 2.77. *For simplicity, we will restrict ourselves in this example to commutative rings.* First, just to point out terminology that we have already introduced in Example 2.71, by a *multiple of* r in a general *commutative* ring R, we mean the set $\{ra \mid a \in R\}$. (This obviously generalizes the notion of multiple that we use in $\mathbb{Z}$.) In Examples 2.70 and 2.71, we considered the set of all multiples of a given element of our ring (multiples of m in the case of $\mathbb{Z}$, multiples of $f(x)$ in the case of $\mathbb{R}[x]$), and observed that these formed an ideal. In Example 2.72, we considered something more general: the set $\langle p(x,y), q(x,y)\rangle$ is the set of *sums of multiples of* $p(x,y)$ *and* $q(x,y)$. This process can be generalized even further. If $a_1, \ldots, a_n$ are elements of a *commutative* ring R, we denote by $\langle a_1, \ldots, a_n\rangle$ the set of all elements of R that are expressible as $a_1r_1 + \cdots + a_nr_n$ for suitable elements $r_1, \ldots, r_n$ in R. Thus, the elements of $\langle a_1, \ldots, a_n\rangle$ are sums of multiples of the a_i. (As in Example 2.72, the r_i may not be uniquely determined, but this will not be an issue.)

> **Exercise 2.77.1.** Show that $\langle a_1, \ldots, a_n\rangle$ is an ideal of R.

The ideal $\langle a_1, \ldots, a_n\rangle$ is known as the *ideal generated by* $a_1, \ldots, a_n$. An ideal generated by a *single* element is known as a *principal ideal.* Thus, the ideal $2\mathbb{Z}$ is a principal ideal in $\mathbb{Z}$. (Of course, the ideal $2\mathbb{Z}$ could just as easily have been denoted by $\langle 2\rangle$.) See Exercise 2.130 ahead.

> **Exercise 2.77.2.** Show that $\langle a_1, \ldots, a_n\rangle$ is the smallest ideal containing $a_1, \ldots, a_n$, in the sense that if J is any ideal of R that contains $a_1, \ldots, a_n$, then $\langle a_1, \ldots, a_n\rangle \subseteq J$.

Question 2.77.1. Convince yourselves that $\langle 1\rangle = R$ and $\langle 0\rangle$ is just the zero ideal $\{0\}$.

Exercise 2.77.3. Suppose that R is a field, and let a be a nonzero element of R. Show that $\langle a\rangle = R$. (Hint: play with the fact that a^{-1} exists in R and that $\langle a\rangle$ is an ideal.)

Exercise 2.77.4. Conclude that the only ideals in a field F are the set $\{0\}$ and F.

2.5 Quotient Rings

We now come to a fundamental method of constructing a new ring from a given ring and an ideal in the ring, namely, the *quotient ring* construction.

Let R be a ring (not necessarily commutative) and let I be an ideal in R. We define a relation $\sim$ on R by declaring $a \sim b$ if and only if $a - b \in I$. It is immediate that $\sim$ is an equivalence relation:(i) certainly, for any a, $a - a = 0 \in I$; (ii) if $a \sim b$, then by definition $a - b \in I$, but since I is an ideal, $-1(a - b) = b - a \in I$, so $b \sim a$ as well; (iii) finally, if $a \sim b$ and $b \sim c$, then by definition, $a - b \in I$ and $b - c \in I$, so again because I is an ideal, $(a - b) + (b - c) = a - c \in I$, so $a \sim c$ as well.

Let us denote the equivalence class of an element a as $[a]$. (Recall what this means: it is the set of all elements in R that are related to a under this equivalence relation.) Let us also denote by $a + I$ the set of all elements of the ring of the form $a + i$ as i varies in I. The set denoted $a + I$ is called the *coset of I with respect to a*. We have the following:

Lemma 2.78. *The equivalence class $[a]$ is precisely the coset $a + I$.*

Proof. Take $b \in [a]$. Then $b \sim a$, so by definition, $b - a \in I$. Thus, $b - a = i$ for some $i \in I$, or written differently, $b = a + i$. Thus, $b \in a + I$, and since b was arbitrary, we find $[a] \subseteq a + I$. Conversely, take any element $b \in a + I$. Then by definition of the set $a + I$, we find $b = a + i$ for some $i \in I$. But this just means $b - a \in I$, that is $b \sim a$. Thus, $b \in [a]$ and since b was arbitrary, we find $a + I \subseteq [a]$. This proves that the two sets are equal.

□

Let us write R/I ("$R \bmod I$") for the set of equivalence classes of R under the relation $\sim$ above. Thanks to Lemma 2.78 we know that the equivalence class of $r \in R$ is the same as the coset $r + I$, so we will use the notation $[r]$ and $r+I$ interchangeably for the equivalence class of r. The key observation we make is that the set R/I can be endowed with two binary operations $+$ (addition) and $\cdot$ (multiplication) by the following rather natural definitions:

Definition 2.79. $[a] + [b] = [a + b]$ and $[a] \cdot [b] = [a \cdot b]$ for all $[a]$ and $[b]$ in R/I. (In coset notation, this would read $(a + I) + (b + I) = (a + b) + I$, and $(a + I)(b + I) = ab + I$.) As always, if the context is clear, we will often omit the "$\cdot$" sign and write $[a][b]$ for $[a] \cdot [b]$.

Before proceeding any further, we need to settle the issue of whether these definitions make sense, in other words, whether these operations are *well-defined*. Observe that the definition of addition, for instance, depends on which representative we use for the equivalence classes. Now recall that if $a' \sim a$, then $[a] = [a']$. Similarly, if $b' \sim b$, then $[b] = [b']$. If we use a and b as representatives for the equivalence classes to which they belong, our definition of the sum of the two classes is the class to which $a + b$ belongs. However, if we use a' and b' as representatives for the classes $[a]$ and $[b]$, our definition says that the sum of the two classes is the class to which $a' + b'$ belongs. Can we be certain that the class to which $a+b$ belongs is the same as the class to which $a' + b'$ belongs? If yes, then we can be certain that our definition of addition is independent of which representative we use for the equivalence class. We have the following:

Lemma 2.80. *The operations of addition and multiplication on R/I described above in Definition 2.79 are indeed well-defined. Moreover, the addition operation is commutative.*

Proof. As in the paragraph above, suppose that $a' \sim a$ and $b' \sim b$. Then, by definition, $a' - a = i$ for some $i \in I$, and $b' - b = j$ for some $j \in I$. Thus,

$a' + b' = (a + i) + (b + j) = (a + b) + (i + j)$. Since I is an ideal and hence closed under addition, $i + j$ is also in I. Thus, we find that $(a' + b') - (a + b)$ is in I, that is, $a' + b'$ is related to $a + b$. Put differently, this just means that $[a' + b'] = [a + b]$, so indeed, addition is well-defined.

As for multiplication, note that $a'b' = (a + i)(b + j) = ab + aj + ib + ij$. Since I is an ideal, $j \in I$ implies that $aj \in I$ and $ij \in I$, and again since I is an ideal, $i \in I$ implies that $ib \in I$. Thus, $aj + ib + ij \in I$ as well (as I is closed under addition). It follows that $a'b' - ab \in I$, or put differently, $[a'b'] = [ab]$. This shows that multiplication is well-defined.

Finally, note that $[a] + [b] = [a + b] = [b + a]$ (the last equality is because $a + b = b + a$ in the ring R), and of course, $[b + a] = [b] + [a]$. Hence, $[a] + [b] = [b] + [a]$. □

Remark 2.81. The proof above illustrates why we require in the definition of ideals that they be closed under addition and that $ir \in I$ and $ri \in I$ for all i in I and all $r \in R$ (see Lemma 2.67). It was this that allowed us to say that addition and multiplication are well-defined: we needed to know above that $i + j \in I$ in the proof that addition is well-defined, and that $aj \in I$ and $ib \in I$ and $ij \in I$ and then $aj + ib + ij \in I$ in the proof that multiplication is well-defined, and for this, we invoked the corresponding properties of ideals.

Having proved that the operations $+$ and $\cdot$ on R/I are well-defined, let us proceed to prove that all ring axioms hold in R/I:

Theorem 2.82. $(R/I, +, \cdot)$ *is a ring.*

Proof. We proceed to check all axioms one by one:

1. *Associativity of* $+$*:* Given elements $[a]$, $[b]$, and $[c]$ in R/I, we need to check that $([a] + [b]) + [c] = [a] + ([b] + [c])$. Now $([a] + [b]) + [c] = [a+b]+[c]$ by definition of $[a]+[b]$, and similarly $[a+b]+[c] = [(a+b)+c]$. But by the associativity of addition *in R,* $(a+b)+c = a+(b+c)$. Hence, $[(a + b) + c] = [a + (b + c)]$. But applying the definition of addition of

two elements of R/I in reverse, $[a + (b + c)]$ is just $[a] + [b + c]$, which is then $[a] + ([b] + [c])$. Thus, $+$ is associative in R/I.

2. *Existence of identity element for $+$:* The element $[0]$ is the additive identity, since for any element $[a]$, $[a]+[0] = [a+0] = [a]$, and $[0]+[a] = [0 + a] = [a]$.

3. *Existence of inverses under $+$:* For any element $[a]$, the element $[-a]$ is the inverse of $[a]$ under $+$, since $[a] + [-a] = [a + (-a)] = [0]$, and similarly, $[-a] + [a] = [-a + a] = [0]$.

4. *Commutativity of $+$:* This was already observed in Lemma 2.80 above.

5. *Associativity of $\cdot$:* This proof is similar to the proof of associativity of $+$ above.

6. *Existence of identity for $\cdot$:* The element $[1]$ acts as the "1" of R/I since for any element $[a]$, $[a] \cdot [1] = [a \cdot 1] = [a]$, and similarly, $[1] \cdot [a] = [1 \cdot a] = [a]$.

7. *Distributivity of $\cdot$ over $+$:* For any elements $[a]$, $[b]$, and $[c]$ in R/I, we have $[a] \cdot ([b] + [c]) = [a] \cdot [b + c] = [a \cdot (b + c)] = [a \cdot b + a \cdot c]$ (this last equality is because of the distributive property in R). And of course, $[a \cdot b + a \cdot c] = [a \cdot b] + [a \cdot c] = [a] \cdot [b] + [a] \cdot [c]$. Putting it together, we find $[a] \cdot ([b] + [c]) = [a] \cdot [b] + [a] \cdot [c]$. The proof that $([a] + [b]) \cdot [c] = [a] \cdot [c] + [b] \cdot [c]$ is similar.

□

Definition 2.83. $(R/I, +, \cdot)$ is called the *quotient ring* of R by the ideal I.

How should one visualize R/I? Here is one intuitive way. Note that the zero of R/I is the element $[0]$, which is just the coset $0+I$ (see Lemma 2.78). But the coset $0 + I$ is the set of all elements of R of the form $0 + i$ for some $i \in I$, and of course, the set of all such elements is just I. Thus, we may view the quotient construction as something that takes the ring R and simply converts all elements in I to zero—more colloquially, the construction "kills" all elements in I, or "divides out" all elements in I. This last description explains the term "quotient ring," and pushing the analogy one step further, R/I can then be thought of as the set of all "remainders" after dividing out by I, endowed with the natural "quotient" binary operations of Definition 2.79.

Example 2.84. As our first example, take R to be $\mathbb{R}[x]$, and I to be $\langle x\rangle$ (Example 2.71). What does R/I "look like" here? Any polynomial in $\mathbb{R}[x]$ is of the form $a_0+a_1x+a_2x^2+\cdots+a_nx^n$ for some $n\geq 0$ and some $a_i\in\mathbb{R}$. The monomials a_1x, a_2x^2, …, a_nx^n are all in I since each of these is a multiple of x. If we "set these to zero", we are left with simply a_0 which is a real number. Thus, $\mathbb{R}[x]/\langle x\rangle$ is just the set of constant terms (the coefficients of x^0) as we range through all the polynomials in $\mathbb{R}[x]$. But the set of constant terms is precisely the set of all real numbers, since every constant term is just a real number, and every real number shows up as the constant term of some polynomial. Thus, $\mathbb{R}[x]/\langle x\rangle$ "equals" $\mathbb{R}$. But this equality is more than just an equality of sets: it is an equality that preserves the ring structure. (We will make the notion of "preserving ring structure" more precise in the next section—see Example 2.104; Example 2.96 is also relevant.)

Example 2.85. Here is another example that would help us understand how to visualize R/I. Consider $\mathbb{R}[x]$ again, but this time take I to be $\langle x^2+1\rangle$ (Example 2.71, Exercise 2.71.2). Notice that x^2 is in the same equivalence class as -1, since $x^2-(-1)=x^2+1$ is clearly in I. What this means is in the quotient ring $\mathbb{R}[x]/\langle x^2+1\rangle$, we may represent the coset x^2+I by $-1+I$. (Another way of thinking about this is to note that x^2 may be written as $(x^2+1)+(-1)$. If we "kill off" the first summand x^2+1, which is in I, we arrive at the representative $-1+I$ for x^2+I.) But there is more. As we have seen while proving the well-definedness of multiplication in R/I (Lemma 2.80 above), if $x^2\sim -1$, then $x^2\cdot x^2\sim$

$(-1)\cdot(-1)$. Thus, $x^4 \sim 1$, so we may replace $x^4 + I$ by $1 + I$. Proceeding, we find x^6+I is the same as $-1+I$, x^8+I is the same as $1+I$, etc. Moreover, $x^3 + I = (x+I)(x^2+I) = (x+I)(-1+I) = (-x+I)$, etc. The coset of any monomial x^n is thus either $\pm 1+I$ or $\pm x+I$. For instance, while considering the equivalence class of a polynomial such as $2 - 5x + 3x^2 + 2x^3 - 2x^4 + x^5$, which is $(2+I)-(5+I)(x+I)+(3+I)(x^2+I)+(2+I)(x^3+I)-(2+I)(x^4+I)+(x^5+I)$, we may make the replacements above to find that it is the same as $(2+I)-(5+I)(x+I)+(3+I)(-1+I)+(2+I)(-x+I)-(2+I)(1+I)+(x+I)$. Multiplying out, we find this is the same as $(2 - 5x - 3 - 2x - 2 + x) + I$, which simplifies to $(-3-6x)+I$ or $(-3+I)-6(x+I)$. Temporarily writing $\overline{x}$ for $x + I$, we loosely think of $(-3 - 6x) + I$ as the element "$-3 - 6\overline{x}$ subject to the relation $\overline{x}^2 + 1 = 0$", or what is the same thing, "$\overline{x}^2 = -1$." But if $\overline{x}^2 = -1$, then $\overline{x}$ behaves like our familiar complex number $\imath$! Thus, we appear to have obtained the complex number $-3 - 6\imath$ as the equivalence class of $2 - 5x + 3x^2 + 2x^3 - 2x^4 + x^5$ mod $\langle x^2 + 1\rangle$. Indeed this is true: it turns out that the ring $\mathbb{R}[x]/\langle x^2 + 1\rangle$ is "the same as" the set of complex numbers. We will make all this precise and justify these heuristics in the next section (where this example will appear in Exercise 2.112).

Let us revisit a couple of quotient rings that we have already considered.

Example 2.86. The ring $\mathbb{Z}/2\mathbb{Z}$ is really just the quotient ring of $\mathbb{Z}$ by the ideal $2\mathbb{Z}$. Recall that $[0]_2$ and $[1]_2$ are precisely the equivalence classes of $\mathbb{Z}$ under the equivalence relation $a \sim b$ iff $a - b$ is even (see 2.20). Since the even integers constitute the ideal $2\mathbb{Z}$, this is precisely the sort of equivalence relation we have considered in this section.

> **Question 2.86.1.** Write down the addition and multiplication tables for the ring operations on $\mathbb{Z}/2\mathbb{Z}$ that we have introduced in Definition 2.79 of this section. Can you see that these are precisely the ring operations we defined early on in Example 2.20?

Example 2.87. In a similar manner, the ring $\mathbb{Z}/3\mathbb{Z}$ of Example 2.21 is the quotient ring of $\mathbb{Z}$ by the ideal $3\mathbb{Z}$.

Question 2.87.1. Do you see this?

More generally, one can consider the ideal $m\mathbb{Z}$ for $m \geq 4$ and construct the ring $\mathbb{Z}/n\mathbb{Z}$ with operations as in Definition 2.79.

Exercise 2.87.1. Redo Exercise 2.21.2 in this new light.

2.6 Ring Homomorphisms and Isomorphisms

The process of forming the quotient ring of a ring R by an ideal I is worth studying from an alternative perspective. Intuitively speaking, the ring operations in R/I are "essentially the same" as the operations in R except that the elements of R have all been divided out by I. What do we mean by this? Let us take addition: Suppose we have two elements $r + I$ and $s + I$ in R/I, and we wish to know how to add them. By the very definition of addition in R/I, to add $r + I$ and $s + I$ in R/I is the same as adding r and s first in the ring R and then "pushing the answer down" to R/I to obtain the coset $(r+s)+I$. It is in this sense that adding in R/I is "essentially the same" as adding in R. We can view this in terms of the function $f : R \to R/I$ that "pushes" $r \in R$ "down" to $r + I$. Since $r + I = f(r)$, $s + I = f(s)$, and $(r + s) + I = f(r + s)$, we find $f(r) + f(s) = f(r + s)$. The function f that sends r to $r + I$, along with the property $f(r) + f(s) = f(r + s)$ for all r and s in R, precisely captures the notion that addition in R/I and R are "essentially the same."

Similarly, the definition of multiplication in R/I: $(r + I)(s + I) = rs + I$ gives the feeling that multiplication in R/I is "the same" as the multiplication in R except for dividing out by I: once again this intuition is captured by the function f above along with the property $f(r)f(s) = f(rs)$ for all r and s in R.

We will turn this situation around. Suppose one has a function f from one ring R to another ring S which has the two properties described above (along with one another: $f(1_R) = 1_S$, where 1_R and 1_S are the multiplicative identities in R and S respectively). In the paragraphs above, the map $f : R \to R/I$ was surjective, but let us be more general, and not assume that our map f from R to S is surjective. It will turn out that the image of f will, all the same, be a subring of S (see Lemma 2.103 ahead). In such a situation too, it will turn out, the ring operations in the ring R and in the image of f (a subring of S) will "essentially be the same" except perhaps for dividing out by some ideal. We will give this a name:

Definition 2.88. Let R and S be two rings, and let $f \colon R \to S$ be a function. Suppose that f has the following properties:

1. $f(a) + f(b) = f(a + b)$ for all a, b, in R,
2. $f(a)f(b) = f(ab)$ for all a, b, in R,
3. $f(1_R) = 1_S$.

Then f is said to be a *ring homomorphism* from R to S.

Remark 2.89. There are some features of this definition that are worth noting:

1. In the equation $f(a) + f(b) = f(a + b)$, note that the operation on the left side represents addition in the ring S, while the operation on the right side represents addition in the ring R. Loosely, we say that any function $f : R \to S$ satisfying $f(a) + f(b) = f(a + b)$ "preserves addition."

2. Similarly for the equation $f(a)f(b) = f(ab)$: the operation on the left side represents multiplication in S, while the operation on the right side represents multiplication in R. Loosely, we say that any function $f : R \to S$ satisfying $f(a)f(b) = f(ab)$ "preserves multiplication."

3. By the very definition of a function, f is defined on all of R. The image of R under f, however, need not be all of S (i.e, f need not be surjective). We will see examples of this ahead (see Example 2.97 and Example 2.98 for instance). However, the image of R under f is

not an arbitrary subset of S. The definition of a ring homomorphism ensures that the image of R under f is actually a *subring* of S (see Lemma 2.103 later in this section).

4. In fact, the stipulation $f(1_R) = 1_S$ in the definition of a ring homomorphism is made precisely to ensure that the image of f is a subring of S.

5. Writing 0_R and 0_S for the additive identities of R and S respectively, note that it is not necessary to stipulate that $f(0_R) = 0_S$—this property holds automatically, as we will prove in Lemma 2.90 below.

Lemma 2.90. *Let $f : R \to S$ be a ring homomorphism. Then $f(0_R) = 0_S$.*

Proof. We start with the fact that $f(0_R) = f(0_R + 0_R) = f(0_R) + f(0_R)$, where the first equality is because $0_R = 0_R + 0_R$, and the second equality is because $f(a + b) = f(a) + f(b)$ for all a and b in R. We now have an equality in S: $f(0_R) = f(0_R) + f(0_R)$. Since $f(0_R) = 0_S + f(0_R)$, we find $0_S + f(0_R) = f(0_R) + f(0_R)$. By additive cancellation (see Remark 2.24) we find $0_S = f(0_R)$, thereby proving the lemma.

□

There is an immediate corollary to this that will be useful (see Corollary 4.60 in Chapter 4):

Corollary 2.91. *Let $f : R \to S$ be a ring homomorphism. Then, for all $a \in R$, $f(-a) = -f(a)$. In particular, $f(-1_R) = -1_S$.*

Proof. Note that $0_R = a+(-a)$. Hence, $f(0_R) = f(a+(-a)) = f(a)+f(-a)$. Since $f(0_R) = 0_S$ by Lemma 2.90, we find $0_S = f(a) + f(-a)$ (and by commutativity of addition in S, $0_S = f(-a) + f(a)$ as well). It follows that $f(-a) = -f(a)$. In particular, taking $a = 1_R$ and noting that $f(1_R) = 1_S$ by definition of a ring homomorphism, the last line of the corollary follows.

□

Before proceeding to examples of ring homomorphisms, let us consider one remaining issue: the concept of a ring homomorphism was introduced to capture the notion of operations on two rings being "the same" except for dividing out by some ideal. What is the natural candidate for this ideal? To divide out by an ideal in R is to make it zero in S (recall our discussion after Definition 2.83 on how to view R/I). This leads naturally to the following:

Definition 2.92. Given a ring homomorphism $f : R \to S$, the *kernel* of f is the set $\{r \in R \mid f(r) = 0_S\}$. It is denoted $ker(f)$.

(Thus, the kernel of f is the set of all elements of R that get mapped to 0_S under f.)

After these discussions, the following statement should come as no surprise:

Proposition 2.93. *The kernel of a ring homomorphism $f : R \to S$ is an ideal of R.*

Proof. Given a and b in $ker(f)$, note that $f(a+b) = f(a)+f(b) = 0_S+0_S = 0_S$, so $a + b$ is also in the kernel of f. Hence $ker(f)$ is closed under the addition operation on R. We first wish to show that $(ker(f), +)$ is an abelian group. Both associativity and commutativity of $+$ follow from the fact that these properties hold for the addition operation on all of R, so we only need to show that 0_R is in $ker(f)$ and that for all $a \in ker(f)$, $-a$ is also in $ker(f)$. The fact that $0_R \in ker(f)$ is just a restatement of Lemma 2.90. Now, for any $a \in ker(f)$, $f(a) = 0_S$ by definition of the kernel. By Corollary 2.91, $f(-a) = -f(a)$, so $f(-a) = -0_S = 0_S$. This shows that $-a$ is in $ker(f)$ as well. Thus, $(ker(f), +)$ is indeed an abelian group.

Next, note that for any $r \in R$ and $a \in ker(f)$, $f(ra) = f(r)f(a) = f(r) \cdot 0_S = 0_S$ (for the last equality, recall the properties in Remark 2.24), and similarly, $f(ar) = f(a)f(r) = 0_S \cdot f(r) = 0_S$, so both ra and ar are also in $\ker(f)$. This proves that $ker(f)$ is indeed an ideal of R.

□

Here are some examples of ring homomorphisms:

Example 2.94. Let us revisit the example that started this discussion on ring homomorphisms: a ring R, an ideal I in R, the quotient ring R/I, and the function $f : R \to R/I$ that sends r to its equivalence class modulo I, i.e., $[r]$, or what is the same thing, $r + I$. We have already observed in our discussion above that $f(r + s) = f(r) + f(s)$ and $f(rs) = f(r)f(s)$— in fact, it is these properties of f that led us to the definition of a ring homomorphism. It is immediate that the third property of Definition 2.88 also holds: By Theorem 2.82, the multiplicative identity in R/I is $1+I$, and indeed, $f(1) = 1 + I$. Thus, $f(1_R) = 1_{R/I}$ as desired. What is the kernel of f? We expect it be I, since our entire discussion of kernels was modeled on how, in this very example, we are "dividing out by I". Let us formally verify this: the kernel of f is all $r \in R$ such that $f(r) = 0_{R/I}$. Now, by Theorem 2.82, the zero in R/I is $0 + I$. Thus, $f(r) = 0$ if and only if $r + I = 0 + I$, i.e., if and only if $r + I = I$. Now if $r + I = I$, this means in particular that r $(= r + 0)$, which is an element of $r + I$, must be in I. Conversely, if $r \in I$, then it is easy to see (prove it!) that the set $r + I$ must equal I. Putting this together, we find that the kernel of f is precisely I.

Example 2.95. As a special case of Example 2.94, we have, for any $m \geq 2$, a ring homomorphism from $\mathbb{Z}$ to $\mathbb{Z}/m\mathbb{Z}$ defined by $f(a) = a + m\mathbb{Z}$, whose kernel if precisely the ideal $m\mathbb{Z}$ (see Example 2.87).

Example 2.96. Consider the function $f\colon \mathbb{R}[x] \to \mathbb{R}$ that sends x to 0 and more generally, a polynomial $p(x)$ to $p(0)$. (Thus, given a polynomial $p(x) = a_0 + a_1x + a_2x^2 + \cdots a_kx^k$, f simply "sets x to zero", so f sends $p(x)$ to a_0.)

> **Exercise 2.96.1.** Prove that f is a ring homomorphism from $\mathbb{R}[x]$ to $\mathbb{R}$.
>
> **Exercise 2.96.2.** Prove that the kernel of f is precisely the ideal $\langle x\rangle$.

See the discussion on page 61. We will have more to say on this example ahead (see Example 2.104 and Theorem 2.110). See also Example 2.101 ahead for a generalization.

Example 2.97. Consider $\mathbb{Z}$ as a subset of $\mathbb{Q}$. The function $f : \mathbb{Z} \to \mathbb{Q}$ that sends $n \in \mathbb{Z}$ to the fraction $n/1$ is a ring homomorphism.

> **Exercise 2.97.1.** Prove this.

> **Exercise 2.97.2.** Prove that the kernel of f is the zero ideal in $\mathbb{Z}$.

Note that the image of f is just the integers, and in particular, f is not surjective.

Example 2.98. More generally, if R is a subring of S, the function $f : R \to S$ that sends r to r is a ring homomorphism. The image of f is just R, so if R is a proper subset of S, then f will not be surjective.

Example 2.99. Consider the function $f : \mathbb{Q}[x] \to \mathbb{Q}[\sqrt{2}]$ that sends x to $\sqrt{2}$ and more generally $p(x)$ to $p(\sqrt{2})$. (Thus, given a polynomial $p(x) = a_0 + a_1x + a_2x^2 + \cdots a_kx^k$, f simply "sets x to $\sqrt{2}$", so f sends $p(x)$ to the element $a_0 + a_1\sqrt{2} + a_2(\sqrt{2})^2 + \cdots a_k(\sqrt{2})^k$. Of course, this horrible expression simplifies into one of the form $a + b\sqrt{2}$, by using the fact that $(\sqrt{2})^2 = 2$, $(\sqrt{2})^3 = 2\sqrt{2}$, etc.)

> **Exercise 2.99.1.** Prove that f is a ring homomorphism.

> **Exercise 2.99.2.** Prove that f is surjective.
> (Hint: Given rationals a and b what is the image of $a + bx$?)

Let us determine the kernel of this homomorphism. Since x^2 goes to 2, $x^2 - 2$ is certainly in the kernel. Since the kernel is an ideal of $\mathbb{Q}[x]$, the set of multiples of $x^2 - 2$ (which is the principal ideal denoted $\langle x^2 - 2\rangle$, see Example 2.77), will also be in the kernel. We will show that there are no other elements in the kernel, that is, $ker(f) = \langle x^2 - 2\rangle$. To this end, let us invoke polynomial long division that is taught in high school (and which we will revisit in Exercise 2.131 at the end of this chapter). So, suppose we are given an arbitrary polynomial $p(x)$ that is in $ker(f)$. We wish to show that $p(x)$ is a multiple of x^2-2. Dividing $p(x)$ by x^2-2 using long division, we can write $p(x)=q(x)(x^2-2)+r(x)$ for some quotient polynomial $q(x)$ and some remainder $r(x)$ that is at most of degree 1. We wish to show that $r(x)$ is actually zero. Since $r(x)$ is at most of degree 1, we may write it as $a+bx$ for some a and b in $\mathbb{Q}$. Since f is a ring homomorphism, $f(p(x))=f(q(x))f(x^2-2)+a+b\sqrt{2}$, and since $p(x)$ goes to zero under f, we find $p(\sqrt{2})=0=q(\sqrt{2})\cdot 0+a+b\sqrt{2}$. Thus, we find $a+b\sqrt{2}=0$. But we have seen in Exercise 2.12.4 that this is impossible unless $a=b=0$. Thus, $r(x)$ must be zero, thereby showing that $ker(f)=\langle x^2-2\rangle$.

> **Question 2.99.1.** After all, x goes to $\sqrt{2}$ under f, so why is $x-\sqrt{2}$ not in the kernel of f?

Example 2.100. Here is an example similar in spirit to Example 2.99 above.

> **Exercise 2.100.1.** Show that the function $f:\mathbb{Q}[x]\to\mathbb{Q}[\imath]$ that sends x to i and more generally $p(x)$ to $p(i)$ is a surjective ring homomorphism, whose kernel is the ideal $\langle x^2+1\rangle$.

Example 2.101. After seeing in Examples 2.99 and 2.100 above how long division can be used to determine kernels of homomorphisms from $\mathbb{Q}[x]$ to other rings, the following should be easy:

> **Exercise 2.101.1.** Let F be any field, and let $a\in F$ be arbitrary. Show that the function $f: F[x]\to F$ that sends x to a and more generally $p(x)$ to $p(a)$ is a surjective ring homomorphism whose kernel is the ideal generated by $\langle x-a\rangle$.
> Notice that this example generalizes 2.96 above. The process of sending x to a is also known as *evaluation at* a, and hence this homomorphism is also known as the *Evaluation Homomorphism*.

We now come to a very special family of ring homomorphisms, namely, ring isomorphisms. While ring homomorphisms capture the notion that somehow the addition and multiplication in two rings are "essentially the same" except perhaps for "dividing out" by some ideal, isomorphisms capture a stronger notion: that multiplication in two rings are "essentially the same" *without even having to divide out by any ideal.*

First, we need a couple of lemmas:

Lemma 2.102. *Let $f : R \to S$ be a ring homomorphism. Then f is an injective function if and only if $ker(f) = \{0_R\}$ (the zero ideal in R).*

Proof. Suppose f is injective. Suppose that $r \in ker(f)$, so $f(r) = 0_S$. By Lemma 2.90, $f(0_R) = 0_S$. Since both r and 0_R map to the same element in S and since f is injective, we find $r = 0$. Thus, the kernel of f consists of just the element 0_Rl. Conversely, suppose that $ker(f) = \langle 0_R \rangle$. Suppose that $f(r_1) = f(r_2)$ for r_1, r_2 in R. Since f is a ring homomorphism, we find $f(r_1 - r_2) = f(r_1) + f(-r_2) = f(r_1) - f(r_2) = 0_S$ (the last but one equality is because of Corollary 2.91). Thus, $r_1 - r_2 \in ker(f)$. But $ker(f)$ is the zero ideal, so $r_1 - r_2 = 0$, i.e., $r_1 = r_2$. Hence, f is injective.

□

Lemma 2.103. *Let $f : R \to S$ be a ring homomorphism. Write $f(R)$ for the image of R under f. Then $f(R)$ is a subring of S.*

Proof. We will apply Lemma 2.30 to $f(R)$. Given arbitrary s_1 and s_2, in $f(R)$, note that by definition of being in the image of R, $s_1 = f(r_1)$ and $s_2 = f(r_2)$ for some elements r_1 and r_2 in R (these elements are not necessarily uniquely determined in R). Then $s_1 + s_2 = f(r_1) + f(r_2) = f(r_1 + r_2)$ (the last equality is because f is a ring homomorphism), thus showing that s_1+s_2 is also in the image of R. Hence $f(R)$ is closed under addition. Similarly, $s_1 s_2 = f(r_1)f(r_2) = f(r_1 r_2)$, so $f(R)$ is also closed under multiplication. By definition, $f(1_R) = 1_S$, so $1_S \in f(R)$. Finally, we need to show that $-s_1 \in f(R)$ (recall that s_1 is an arbitrary element of $f(R)$). But this is easy: thanks to Corollary 2.91, $f(-r_1) = -f(r_1) = -s_1$, so indeed $-s_1 \in f(R)$. Hence, $f(R)$ is a subring of S.

□

We now quantify our observation (see the discussion on page 61) that somehow, the rings $\mathbb{R}/\langle x\rangle$ and $\mathbb{R}$ are "equal." Let us revisit this example again in a new light:

Example 2.104. Let us define $\tilde{f} : \mathbb{R}[x]/\langle x\rangle \to \mathbb{R}$ by $\tilde{f}(p(x) + \langle x\rangle) = p(0)$. Let us explain this: the equivalence class of a polynomial $p(x)$ under the equivalence relation that defines the ring R/I is the coset $p(x) + I$ (see Lemma 2.78). Our function sends the equivalence class of $p(x)$ to the constant term of $p(x)$. We first need to check that this function is well-defined: we have defined $\tilde{f}$ in terms of one representative of an equivalence class, what if we had used another representative? So, suppose $p(x) + \langle x\rangle = q(x) + \langle x\rangle$, then if we had used $q(x)$, we would have defined $\tilde{f}(p(x) + \langle x\rangle) = \tilde{f}(q(x) + \langle x\rangle) = q(0)$. Earlier, we had defined $\tilde{f}(p(x) + \langle x\rangle)$ to be $p(0)$: are these definitions the same? In other words, is $p(0) = q(0)$? The answer is yes! For, the fact that $p(x) + \langle x\rangle = q(x) + \langle x\rangle$ means that $p(x) - q(x) \in \langle x\rangle$ (why?), or alternatively, $p(x) - q(x)$ is a multiple of x. Hence, the constant term of $p(x) - q(x)$, which is $p(0) - q(0)$, must be zero, i.e., $p(0) = q(0)$. It follows that $\tilde{f}$ is indeed well-defined.

Now that we know $\tilde{f}$ is well-defined, it is easy to check that $\tilde{f}$ is a ring homomorphism (do it!). What is the kernel of $\tilde{f}$? It consists of all equivalence classes $p(x) + \langle x\rangle$ such that the constant term $p(0)$ is zero. But to say that $p(0)$ is zero is to say that $p(x)$ is divisible by x (why?), or in other words, that $p(x)$ is already in $\langle x\rangle$. Thus, the kernel of $\tilde{f}$ consists of just the equivalence class $\langle x\rangle$—*but this is the zero element in the ring* $\mathbb{R}[x]/\langle x\rangle$. Thus, the kernel of $\tilde{f}$ is just the zero ideal, so by Lemma 2.102, $\tilde{f}$ is injective. Moreover, $\tilde{f}$ is clearly surjective, since every real number r arises as the constant term of some polynomial in $\mathbb{R}[x]$ (for example, the polynomial $r + 0x + 0x^2 + \cdots$).

The function $\tilde{f}$ quantifies why $\mathbb{R}[x]/\langle x\rangle$ and $\mathbb{R}$ are really "equal to each other." There are two ingredients to this: the function $\tilde{f}$, being injective and surjective, provides a one-to-one correspondence between $\mathbb{R}[x]/\langle x\rangle$ and

$\mathbb{R}$ *as sets*, and the fact that $\tilde{f}$ is a ring homomorphism tells us that the addition and multiplication in $\mathbb{R}$ is "essentially the same" as that in $\mathbb{R}[x]/\langle x\rangle$. Moreover, since $\tilde{f}$ has kernel zero, we do not even have to divide out by any ideal in $\mathbb{R}[x]/\langle x\rangle$ to realize this "sameness" of ring operations. Thus, $\mathbb{R}/\langle x\rangle$ and $\mathbb{R}$ are really the same rings, even though they look different. We say that $\mathbb{R}[x]/\langle x\rangle$ is *isomorphic* to $\mathbb{R}$ via the map $\tilde{f}$.

Definition 2.105. Let $f : R \to S$ be a ring homomorphism. If f is both injective and surjective, then f is said to be an *isomorphism* between R and S. Two rings R and S are said to be *isomorphic* (written $R \cong S$) if there is some function $f : R \to S$ that is an isomorphism between R and S.

Let us look at some examples of ring isomorphisms:

Example 2.106. Let us revisit Example 2.38. Denote the function that sends $r \in \mathbb{R}$ to $\text{diag}(r)$ by f.

> **Exercise 2.106.1.** Check that f is bijective as a function from $\mathbb{R}$ to the subring of $M_n(\mathbb{R})$ consisting of matrices of the form $\text{diag}(r)$.
>
> **Exercise 2.106.2.** Also, check that $f(r+s) = f(r) + f(s)$, and $f(rs) = f(r)f(s)$.

Moreover $f(1)$ is clearly the identity matrix. Thus, the function f is indeed a ring homomorphism from $\mathbb{R}$ to the subring of $M_n(\mathbb{R})$ consisting of matrices of the form $\text{diag}(r)$ that is both injective and surjective, or described alternatively, f is an isomorphism between these two rings. Intuitively, these two rings are "the same," even though one appears as a set of ordinary numbers, while the other appears in the form of special matrices.

Example 2.107. Define a function $\tilde{f}$ from the quotient ring $\mathbb{Q}[x]/\langle x^2-2\rangle$ to $\mathbb{Q}[\sqrt{2}]$ by the rule $\tilde{f}(p(x) + \langle x^2-2\rangle) = p(\sqrt{2})$. We will prove that $\tilde{f}$ is an isomorphism between $\mathbb{Q}[x]/\langle x^2-2\rangle$ and $\mathbb{Q}[\sqrt{2}]$.

> **Exercise 2.107.1.** Show that $\tilde{f}$ is well defined. (Hint: If $p(x) + \langle x^2-2\rangle = q(x) + \langle x^2-2\rangle$, then $p(x) - q(x) \in \langle x^2-2\rangle$, so $p(x) - q(x) = g(x)(x^2-2)$ for some polynomial $g(x) \in \mathbb{Q}[x]$. What happens if you set $x = \sqrt{2}$ in this?)

Exercise 2.107.2. Show that $\tilde{f}$ is a ring homomorphism.

Exercise 2.107.3. Show that $\tilde{f}$ is surjective.

Exercise 2.107.4. Show that $\tilde{f}$ is injective. (Hint: Recall that we have proved in Example 2.99 that $p(\sqrt{2})$ is zero precisely when $p(x)$ is divisible by $x^2 - 2$.)

Thus, $\tilde{f}$ provides an isomorphism between $\mathbb{Q}[x]/\langle x^2 - 2\rangle$ and $\mathbb{Q}[\sqrt{2}]$. Intuitively, the two rings are "the same," even though one appears as a quotient ring of polynomials, while the other appears as a subring of the reals.

Example 2.108. The following examples show that well-known fields can show up as subrings of matrices!

Exercise 2.108.1. Let S denote the subset of $M_2(\mathbb{Q})$ consisting of all matrices of the form

$$\begin{pmatrix} a & 2b \\ b & a \end{pmatrix}$$

where a and b are arbitrary rational numbers.

1. Show that S is a subring of $M_2(\mathbb{Q})$.

2. Prove that the map $f : \mathbb{Q}[\sqrt{2}] \to S$ that sends $a + b\sqrt{2}$ to the matrix above is an isomorphism between $\mathbb{Q}[\sqrt{2}]$ and S.

Exercise 2.108.2. Let S denote the subset of $M_2(\mathbb{R})$ consisting of all matrices of the form

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$$

where a and b are arbitrary real numbers.

1. Show that S is a subring of $M_2(\mathbb{R})$.

2. Prove that the map $f : \mathbb{C} \to S$ that sends $a + \imath b$ to the matrix above is an isomorphism between $\mathbb{C}$ and S.

The two examples in the exercises above are referred to as the *regular representation* of $\mathbb{Q}[\sqrt{2}]$ in $M_2(\mathbb{Q})$ and $\mathbb{C}$ in $M_2(\mathbb{R})$ (respectively). More generally, let K/F be any field extension (see Definition 2.60). Then K can be considered as a vector space over F (we will study this in Example 3.7 in Chapter 3 ahead). When the dimension of K over F is finite, say n, then one can always find a subring of $M_n(F)$ that is isomorphic to K: this is considered in Exercise 3.105 in Chapter 3.

Example 2.109. It is not necessary that the rings R and S in the definition of a ring isomorphism be different rings. A ring isomorphism $f : R \to R$ is to be thought of as a one-to-one onto map from R to R that preserves the ring structure. (Such a map is also known as an *automorphism* of R.) Here are some examples:

> **Exercise 2.109.1.** Prove that the map $f : \mathbb{Q}[\sqrt{2}] \to \mathbb{Q}[\sqrt{2}]$ that sends $a + b\sqrt{2}$ (for a, b, in the rationals) to $a - b\sqrt{2}$ is a ring isomorphism. What are the elements of $\mathbb{Q}[\sqrt{2}]$ on which f acts as the identity map?
>
> **Exercise 2.109.2.** Let F be a field, and let a be a nonzero element of F. Let b be an arbitrary element of F. Prove that the map $f : F[x] \to F[x]$ that sends x to $ax + b$ and more generally, a polynomial $p_0 + p_1x + \cdots + p_nx^n$ to the polynomial $p_0 + p_1(ax + b) + \cdots + p_n(ax + b)^n$ is an automorphism of $F[x]$.
>
> **Exercise 2.109.3.** Prove that the *complex conjugation* map $f : \mathbb{C} \to \mathbb{C}$ that sends $a+\imath b$ (given real number a and b) to the complex number $a - \imath b$ is an automorphism of $\mathbb{C}$. Determine the set of complex numbers on which f acts as the identity map.

We now come to a fundamental result that connects homomorphisms and isomorphisms. To motivate this, compare Examples 2.96 and 2.104. In the first example, we defined a function $f\colon \mathbb{R}[x] \to \mathbb{R}$ that sends $p(x)$ to $p(0)$ and observed that it was a ring homomorphism whose image was all of $\mathbb{R}$ and whose kernel was the ideal $\langle x\rangle$, while in the second example, we defined a function $\tilde{f} : \mathbb{R}[x]/\langle x\rangle \to \mathbb{R}$ by $\tilde{f}(p(x) + \langle x\rangle) = p(0)$, and observed that it was well-defined and that it gave us an isomorphism between $\mathbb{R}[x]/\langle x\rangle$ and

$\mathbb{R}$. Observe the close connection between how the functions f and $\tilde{f}$ are defined in the two examples, and observe that the ring $\mathbb{R}[x]/\langle x\rangle$ is obtained by modding $\mathbb{R}[x]$ by the kernel of f. Now as another instance, compare Examples 2.99 and 2.107. Here too, in the first example, we defined a function $f : \mathbb{Q}[x] \to \mathbb{Q}[\sqrt{2}]$ that sends $p(x)$ to $p(\sqrt{2})$, and we observed that it was a surjective ring homomorphism whose kernel was $\langle x^2 - 2\rangle$. In the second example, we defined a function $\tilde{f} : \mathbb{Q}[x]/\langle x^2 - 2\rangle \to \mathbb{Q}[\sqrt{2}]$ by $\tilde{f}(p(x) + \langle x^2 - 2\rangle) = p(\sqrt{2})$, and observed that it was well-defined and that it gave us an isomorphism between $\mathbb{Q}[x]/\langle x^2 - 2\rangle$ and $\mathbb{Q}[\sqrt{2}]$. Once again, observe the close connection between how the functions f and $\tilde{f}$ are defined in the two examples, and observe that the ring $\mathbb{Q}[x]/\langle x^2 - 2\rangle$ is obtained by modding out $\mathbb{Q}[x]$ by the kernel of f.

These connections in the two pairs of examples above are not accidental, and are merely instances of a more general phenomenon, captured by the following:

Theorem 2.110. *(Fundamental Theorem of Homomorphisms of Rings.)* *Let $f : R \to S$ be a homomorphism of rings, and write $f(R)$ for the image of R under f. Then the function $\tilde{f} : R/ker(f) \to f(R)$ defined by $\tilde{f}(r + ker(f)) = f(r)$ is well-defined, and provides an isomorphism between $R/ker(f)$ and $f(R)$.*

Proof. The idea of the proof is already contained in the two sets of examples 2.96 and 2.104, and 2.99 and 2.107 discussed above.

We check that $\tilde{f}$ is well-defined. Suppose $r + ker(f) = s + ker(f)$. Then $r - s \in ker(f)$, so $f(r - s) = f(r) - f(s) = 0$, so $f(r) = f(s)$. Thus, $\tilde{f}(r + ker(f)) = \tilde{f}(s + ker(f))$, i.e., $\tilde{f}$ is well-defined.

Now let us go through the three ingredients in Definition 2.88 and check that $\tilde{f}$ is a ring homomorphism. We have $\tilde{f}\left((r + ker(f)) + (s + ker(f))\right) = \tilde{f}\left((r + s) + ker(f)\right) = f(r+s) = f(r)+f(s) = \tilde{f}(r+ker(f))+\tilde{f}(s+ker(f))$.

Exercise 2.110.1. Justify all the equalities above.

Similarly, $\tilde{f}\left((r + ker(f)) \cdot (s + ker(f))\right) = \tilde{f}\left((r \cdot s) + ker(f)\right) = f(r \cdot s) = f(r) \cdot f(s) = \tilde{f}(r + ker(f)) \cdot \tilde{f}(s + ker(f))$.

Exercise 2.110.2. Again, justify all the equalities above.

Finally, $\tilde{f}(1_R + ker(f)) = f(1_R) = 1_S$.

Question 2.110.1. Why?

Hence $\tilde{f}$ is a ring homomorphism.

We check that $\tilde{f}$ is surjective as a function from $R/ker(f)$ to $f(R)$. Note that any element of $f(R)$ is, by definition, of the form $f(r)$ for some $r \in R$. But then, by the way we have defined $\tilde{f}$, we find $f(r) = \tilde{f}(r + ker(f))$, so indeed $\tilde{f}$ is surjective.

Finally, we check that $\tilde{f}$ is injective. Note that $\tilde{f}(r + ker(f)) = f(r) = 0$ means that $r \in ker(f)$. But this means that $r + ker(f) = ker(f)$ (why?), so $r + ker(f)$ is the zero element of $R/ker(f)$. Thus $\tilde{f}$ is injective.

Putting this together, we find that $\tilde{f}$ provides an isomorphism between $R/ker(f)$ and $f(R)$.

□

Here are examples of applications of this theorem, all built from examples of ring homomorphisms that we have already seen.

Example 2.111. We have the isomorphism (see Example 2.100) $\mathbb{Q}[x]/\langle x^2 + 1\rangle \cong \mathbb{Q}[\imath]$.

Example 2.112. By the same token, we find $\mathbb{R}[x]/\langle x^2 + 1\rangle \cong \mathbb{C}$.

Exercise 2.112.1. Mimic Example 2.100 and construct a homomorphism from $\mathbb{R}[x]$ to $\mathbb{C}$ that sends $p(x)$ to $p(i)$ and prove that it is surjective with kernel $\langle x^2 + 1\rangle$. Then apply Theorem 2.110 to establish the claim that $\mathbb{R}[x]/\langle x^2 + 1\rangle \cong \mathbb{C}$.

Example 2.113. Example 2.101 along with Theorem 2.110 above establishes that for any field F and any $a \in F$, $F[x]/\langle x - a\rangle \cong F$.

2.7 Further Exercises

Exercise 2.114. Starting from the ring axioms, prove that the properties stated in Remark 2.24 hold for any ring R.

(See the notes on page 89 for some hints.)

Exercise 2.115. This generalizes Exercise 2.48: If R is a ring, let R^* denote the set of invertible elements of R. Prove that R^* forms a group with respect to multiplication.

Exercise 2.116. This exercise determines the units of the ring $\mathbb{Z}[\imath]$:

1. Define a function $N : \mathbb{Z}[\imath] \to \mathbb{Z}$ by $N(a + b\imath) = a^2 + b^2$. Show that $N(xy) = N(x)N(y)$ for all x and y in $\mathbb{Z}[\imath]$.

2. If x is invertible in $\mathbb{Z}[\imath]$, show that $N(x)$ must equal 1.

3. Conclude that the only units of $\mathbb{Z}[\imath]$ are ± 1 and $\pm\imath$.

Exercise 2.117. Consider the ring $\mathbb{Q}[\sqrt{m}]$ of Example 2.12. Now assume for this exercise that m is not a perfect square. Show that $a + b\sqrt{m} = 0$ (for a and b in $\mathbb{Q}$) if and only if $a = b = 0$. Show that $\mathbb{Q}[\sqrt{m}]$ is a field.

Exercise 2.118. The following concerns the ring $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$ of Example 2.34, and is designed to show that if a, b, c, and d are rational numbers, then $a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6} = 0$ if and only if a, b, c, and d are all zero.

1. Show that $\sqrt{3/2}$ is not rational. (This is similar to showing that $\sqrt{p}$ is not rational for any prime p.

2. Show that $\sqrt{3} \notin \mathbb{Q}[\sqrt{2}]$. (Hint: Assume that $\sqrt{3} \in \mathbb{Q}[\sqrt{2}]$. Then there must exist rational numbers x and y such that $\sqrt{3} = x + y\sqrt{2}$. Square both sides and arrive at a contradiction. You will need to invoke a fact about $\mathbb{Q}[\sqrt{2}]$ that you were asked to prove in Example 2.12, as well as the results of Chapter 1, Exercise 1.42, and part 1 above.)

3. Now assume that $a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6} = 0$ for some choice of rational numbers a, b, c, and d. Write this as $(a + b\sqrt{2}) + \sqrt{3}(c + d\sqrt{2}) = 0$. Prove that $c + d\sqrt{2}$ must be zero. (Hint: Argue that otherwise we can write $\sqrt{3} = -\dfrac{a + b\sqrt{2}}{c + d\sqrt{2}}$. Why is this last equality a contradiction?)

4. Conclude that this forces $a = b = c = d = 0$.

5. Observe that if $a = b = c = d = 0$ then $a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6} = 0$ trivially. This proves the assertion stated at the beginning.

Exercise 2.119. We will prove in this exercise that $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$ is actually a field.

1. You know that if a and b are rational numbers, then $(a+b\sqrt{2})\cdot(a-b\sqrt{2})$ is also rational. (Why?) Similarly, if c and d are rational numbers, then $(c+d\sqrt{3})\cdot(c-d\sqrt{3})$ is also rational. Now show the following: if a, b, c, and d are all rational numbers, then the product of the four terms

$$\begin{array}{ccccc}(a+b\sqrt{2}+c\sqrt{3}+d\sqrt{6}) & \cdot & (a+b\sqrt{2}-c\sqrt{3}-d\sqrt{6}) & \cdot \\ (a-b\sqrt{2}+c\sqrt{3}-d\sqrt{6}) & \cdot & (a-b\sqrt{2}-c\sqrt{3}+d\sqrt{6}) & \end{array}$$

   is also rational. (This just involves multiplying out all the terms above—do it! However, you can save yourselves a lot of work by multiplying the first two terms together using the formula $(x+y)(x-y) = x^2 - y^2$, and then multiplying the remaining two terms together, and looking out for patterns.)

2. Now show using part (1) above that $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$ is a field. (Hint: Given a nonzero element $a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6}$ in $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$, first note that by Exercise 2.118 above, none of $(a + b\sqrt{2} - c\sqrt{3} - d\sqrt{6})$, $(a - b\sqrt{2} + c\sqrt{3} - d\sqrt{6})$ or $(a - b\sqrt{2} - c\sqrt{3} + d\sqrt{6})$ can be zero—why? Now, in the case of $\mathbb{Q}[\sqrt{2}]$, one finds the inverse of $x + y\sqrt{2}$ by multiplying both the numerator and the denominator of the fraction $\frac{1}{x+y\sqrt{2}}$ by $x - y\sqrt{2}$ and taking advantage of the fact that $(x + y\sqrt{2})(x - y\sqrt{2})$ is rational. What ideas do you get from part (1) above?)

Exercise 2.120. Let R be an integral domain. Show that an element in $R[x]$ is invertible, if and only if it is the constant polynomial $r(= r+0x+0x^2+\cdots)$ for some *invertible element* $r \in R$. In particular, if R is a field, then a polynomial in $R[x]$ is invertible *if and only if* it is a nonzero element of R. (See the notes on Page 88 for a discussion on polynomials with coefficients from an arbitrary ring.)

By contrast, show that the (nonconstant) polynomial $1 + [2]_4x$ in the polynomial ring $\mathbb{Z}/4\mathbb{Z}[x]$ is invertible, by explicitly finding the inverse of $1 + [2]_4x$. Repeat the exercise by finding the inverse of $1 + [2]_8x$ in the polynomial ring $\mathbb{Z}/8\mathbb{Z}[x]$. (Hint: Think in terms of the usual Binomial Series for $1/(1+t)$ from your Calculus courses. Do not worry about convergence issues. Instead, think about what information would you glean from this series if, due to some miracle, $t^n = 0$ for some positive integer n?)

Exercise 2.121. We will revisit some familiar identities from high school in the context of rings! Let R be a ring:

1. Show that $a^2 - b^2 = (a-b)(a+b)$ for all a and b in R if and only if R is commutative.

2. Show that $(a+b)^2 = a^2 + 2ab + b^2$ for all a and b in R if and only if R is commutative.

3. More generally, if R is a commutative ring, prove that the *Binomial Theorem* holds in R: for all a and b in R and for all positive integers n,

$$(a+b)^n = \binom{n}{0}a^n + \binom{n}{1}a^{n-1}b + \binom{n}{2}a^{n-2}b^2 + \cdots + \binom{n}{n-1}ab^{n-1} + \binom{n}{n}b^n$$

Exercise 2.122. An element a in a ring is said to be *nilpotent* if $a^n = 0$ for some positive integer n.

1. Show that if a is nilpotent, then $1-a$ and $1+a$ are both invertible. (Hint: Just as in Exercise 2.120 above, think in terms of the Binomial Series for $1/(1-t)$ and $1/(1+t)$. Do not worry about convergence, but ask yourself what you can learn from the series if $t^n = 0$ for some positive integer n.)

2. Let R be a *commutative* ring. Show that the set of all nilpotent elements in R forms an ideal in R. (Hint: Suppose that $a^n = 0$ and $b^m = 0$. What can you say about $(a+b)^{n+m-1}$, given your knowledge of the Binomial Theorem for commutative rings from Exercise 2.121 above?

Exercise 2.123. Let S denote the set of all functions $f\colon \mathbb{R} \to \mathbb{R}$. Given f and g in S, define two binary operations "+" and "·" on S by the rules

$$\begin{aligned} (f+g)(x) &= f(x) + g(x) \\ (f\cdot g)(x) &= f(x)g(x) \end{aligned}$$

(These are referred to, respectively, as the *pointwise addition and multiplication* of functions.)

1. Convince yourselves that $(S, +, \cdot)$ is a ring. What is the "0" of S? What is the "1" of S?

2. Show that S is not an integral domain. (Hint: Play with functions like $f(x) = x + |x|$ or $g(x) = x - |x|$.)

3. More generally, show that every nonzero $f \in S$ is either a unit or a zero-divisor by showing:

   (a) f is a unit if and only if $f(x) \neq 0$ for all $x \in \mathbb{R}$.

   (b) f is a zero-divisor if and only if $f(x) = 0$ for at least one $x \in \mathbb{R}$.

4. Let $s : \mathbb{R} \to S$ be the function that sends the real number r to the function s_r defined by $s_r(x) = r$ for all $x \in R$. Show that s is an injective ring homomorphism from $\mathbb{R}$ to S The image of s in R is therefore a subring of R that is isomorphic to $\mathbb{R}$. It is known as the set of *constant functions.*

Exercise 2.124. Let R be a ring.

> **Definition 2.124.1.** The *center* of R, written $Z(R)$, is defined to be the set $\{r \in R \mid rx = xr \text{ for all } x \in R\}$.

1. Show that $Z(R)$ is a subring of R.

2. If R is commutative, what is $Z(R)$?

3. Determine $Z(M_2(\mathbb{Z}))$. (Hint: Invoke the fact that a matrix in the center must commute with the four matrices $e_{i,j}$, where $e_{i,j}$ is as defined in Exercise 2.16.4.)

Exercise 2.125. Let R be a ring.

1. If I and J are ideals of R, show that $I \cap J$ is an ideal of R. (Is $I \cup J$ an ideal of R?)

2. If S and T are subrings of R, show that $S \cap T$ is a subring of R. (Is $S \cup T$ a subring of R?)

3. If R is a field, and if S and T are subfields of R, show that $S \cap T$ is a subfield of R.

Exercise 2.126. Here is an example of a ring in which elements do not factor uniquely into a product of primes! Consider the subring of $\mathbb{C}$ generated by $\mathbb{Z}$ and $\sqrt{-5}$, namely, $\mathbb{Z}[\sqrt{-5}]$. By arguments nearly identical to those that you must have used in Exercise 2.117 above, every element in this ring can be written *uniquely* as $a + b\sqrt{-5}$ for suitable integers a and b. We define a function N from $\mathbb{Z}[\sqrt{-5}]$ to $\mathbb{Z}$ as follows: $N(a + b\sqrt{-5}) = a^2 + 5b^2$. (Notice that $a^2 + 5b^2$ is just $(a + b\sqrt{-5}) \cdot (a - b\sqrt{-5})$.)

1. Show that N is *multiplicative,* that is, $N(xy) = N(x)N(y)$ for any two elements x and y of $\mathbb{Z}[\sqrt{-5}]$.

2. Show that if x in $\mathbb{Z}[\sqrt{-5}]$ is such that $N(x) = 1$, then x must be ± 1.

3. Use part 1 and Question 2.45 to show that if x is a unit in $\mathbb{Z}[\sqrt{-5}]$, then $N(x)$ must be 1.

4. Use parts 2 and 3 above to show that if x is a unit in $\mathbb{Z}[\sqrt{-5}]$, then x can only be ± 1.

5. If R is a commutative ring, an *irreducible* in R is a nonzero element x such that if $x = bc$ for two elements b and c, then either b or c must be a unit. (It turns out that this is the correct generalization of the concept of primes that is needed to study unique factorization in arbitrary rings.) Also, just as in $\mathbb{Z}$, we say an element b in an arbitrary commutative ring R *divides* an element a (or is a *divisor* of a) if there exists an element c in R such that $a = bc$. Using part 4, show that if x is an irreducible element in $\mathbb{Z}[\sqrt{-5}]$, then the only divisors of x are ± 1 and $\pm x$. (Thus, at least in $\mathbb{Z}[\sqrt{-5}]$, it is clear that irreducible elements are just like primes.)

6. Show that if x in $\mathbb{Z}[\sqrt{-5}]$ is such that $N(x)$ is a prime integer, then x must be irreducible.

7. Show that there is no element x in $\mathbb{Z}[\sqrt{-5}]$ with $N(x) = 2$. Similarly, show that there is no element x with $N(x) = 3$.

8. Show that 2 is irreducible in $\mathbb{Z}[\sqrt{-5}]$. (Hint: Suppose $2 = xy$. Then $4 = N(2) = N(x)N(y)$, as N is multiplicative. Study the various factorizations of 4 and use the previous parts.)

9. Similarly, show that 3 is irreducible in $\mathbb{Z}[\sqrt{-5}]$.

10. Study the various factors of $N(1+\sqrt{-5})$ and of $N(1-\sqrt{-5})$ and show that both $1+\sqrt{-5}$ and $1-\sqrt{-5}$ are irreducible.

11. Two irreducibles x and y in a commutative ring R are said to be *associates* if $x = yu$ for some unit u. Part 4 shows that in the ring $\mathbb{Z}[\sqrt{-5}]$, two elements x and y are associates if and only if $x = \pm y$. Now use the fact that every element in $\mathbb{Z}[\sqrt{-5}]$ is uniquely expressible as $a+b\sqrt{-5}$ to show that neither 2 nor 3 is an associate of either $1+\sqrt{-5}$ or $1-\sqrt{-5}$.

12. A commutative ring R is said to possess unique prime factorization if every element $a \in R$ that is not a unit factors into a product of irreducibles,

and if $a = x_1x_2\cdots x_s$ and $a = y_1y_2\cdots y_t$ are two factorizations of a into irreducibles, then s must equal t, and after relabeling if necessary, each x_i must be an associate of the corresponding y_i. (Again, it turns out that this is the correct generalization of the concept of unique prime factorization in the integers to arbitrary commutative rings.) Prove that $\mathbb{Z}[\sqrt{-5}]$ does not possess unique prime factorization by considering two different factorizations of 6 into irreducibles. (Hint: Look at parts 8, 9, 10, and 11.)

Exercise 2.127. Prove that any finite integral domain must be a field. (Hint: Write R for the integral domain. Given a nonzero $a \in R$, you need to show that a is invertible. What can you say about the function $f_a : R \to R$ that sends any r to ar? Is it injective? Is it surjective? So?)

Exercise 2.128. Let K be a field, and let R be a subring of K. Assume that every element of K satisfies a *monic* polynomial with coefficients in R: this means that given any k in K, there exists a positive integer n and elements $r_0, r_1, \ldots, r_{n-1}$ in R such that $k^n + r_{n-1}k^{n-1} + \cdots + r_1k + r_0 = 0$. (The term *monic* refers to the fact that the coefficient of k^n in the relation above is 1.) Show that R must also be a field.

(Hint: Since R is already an integral domain, you only need to show that every nonzero element of R is invertible. Given a nonzero $r \in R$, note that r is invertible as an element of K since K is a field. In particular, r^{-1} exists in K. Use the hypothesis to show that r^{-1} actually lives in R.)

Exercise 2.129. Let R and S be two rings. This exercise studies ideals in the direct product R and S. Let I be an ideal of $R \times S$.

1. Let $I_1 = \{a \in R\}$ such that $(a, b) \in I$ for some $b \in S$. Show that I_1 is an ideal of R.

2. Similarly define I_2 to be the set $\{b \in S\}$ such that $(a, b) \in I$ for some $a \in R$. Show that I_2 is an ideal of S.

3. Recall from Example 2.76 the meaning of $I_1 \times I_2$. Show that $I = I_1 \times I_2$. (Hint: $(a, b) = (1_R, 0_S)(a, b) + (0_R, 1_S)(a, b)$.)

We saw in Example 2.76 that if I_1 and I_2 are ideals of R and S respectively, then $I_1 \times I_2$ is an ideal of $R \times S$. This exercise therefore shows the converse: every ideal of $R \times S$ is of the form $I_1 \times I_2$ where I_1 and I_2 are ideals of R and S respectively.

Exercise 2.130. We will prove in this exercise that every ideal in $\mathbb{Z}$ is principal (see Example 2.77). Let I be an ideal of $\mathbb{Z}$. If I consists of just the element 0, then $I = \langle 0 \rangle$ and is already principal. So, assume in what follows that I is a nonzero ideal.

1. Show that I contains at least one positive integer.

2. Let d be the least positive integer in I (Well-Ordering Principle). Given an arbitrary $i \in I$, write $i = dq + r$ for some $0 \le r \le d$ by the division algorithm. Show that $r \in I$.

3. Conclude that r must be zero.

4. Conclude that $I = \langle d \rangle$.

An integral domain in which every ideal is principal is called a *principal ideal domain.* This exercise therefore establishes that $\mathbb{Z}$ is a principal ideal domain.

Exercise 2.131. The purpose of this exercise is to prove the following result, which is known as the *division algorithm for polynomials:* Let F be a field, and let $f(x)$ and $g(x)$ be two polynomials in $F[x]$ with $g(x) \neq 0$. Then, there exist *unique* polynomials $q(x)$ and $r(x)$, such that $f(x) = g(x)q(x) + r(x)$ with either $r(x) = 0$ or else $deg(r(x)) < deg(g(x))$.

Note the similarity of this result with Lemma 1.4 of Chapter 1. This result, of course, simply codes the output of the familiar process of dividing $f(x)$ by $g(x)$ using long division, and in fact, the long division process can be turned around to furnish a proof of this result. However, we will prove this result instead in a manner analogous to how we proved Lemma 1.4 of Chapter 1, to underscore certain similarities between the integers and polynomials. (You will notice that the core of this proof, however, invokes a crucial ingredient of the long division process in Part 5a below!)

1. First prove the uniqueness of $q(x)$ and $r(x)$ as follows: Suppose that $f(x) = g(x)q(x) + r(x)$ and as well, $f(x) = g(x)q'(x) + r'(x)$, for polynomials $q(x)$, $r(x)$, $q'(x)$, and $r'(x)$ with either $r(x) = 0$ or $deg(r(x)) < deg(g(x))$, and similarly, either $r'(x) = 0$ or $deg(r'(x)) < deg(g(x))$. Rewrite this as $g(x)(q(x) - q'(x)) = r'(x) - r(x)$. Show that if $r'(x) - r(x) \neq 0$ then the degree of the right side must be less than the degree of the left side, and hence conclude that $r(x) = r'(x)$ and $q(x) = q'(x)$. This establishes the uniqueness of $q(x)$ and $r(x)$.

2. Now for the existence of $q(x)$ and $r(x)$. First, let S^* denote the set $\{f(x) - g(x)h(x) \mid h(x) \in F[x]\}$. Show that S^* is nonempty.

3. If S^* contains 0, show that we have proved the existence of $q(x)$ and $r(x)$ with the required properties.

4. So assume from now on that S^* does not contain 0. Show that among the elements of S^* there must be an element of least degree.

5. Let $r(x)$ denote an element of least degree in S^*, and let $q(x) \in F[x]$ be that polynomial such that $f(x) - g(x)q(x) = r(x)$. First show that $deg(r(x)) < deg(g(x))$ as follows:

   (a) Assume to the contrary that $deg(r(x)) \geq deg(g(x))$. Let $m = deg(r(x))$ and $n = deg(g(x))$, so $m \geq n$. Let r_m and g_n (respectively) be the highest coefficients of $r(x)$ and $g(x)$ (so, by definition of highest coefficient, r_m and g_n are nonzero). Show that $r(x) - (r_m/g_n)x^{m-n}g(x)$ has degree less than $r(x)$.

   (b) Now show that the element $f(x) - g(x)(q(x) + (r_m/g_n)x^{m-n})$ is an element of S^* that has degree less than that of $r(x)$. Conclude that $deg(r(x)) < deg(g(x))$.

6. Conclude that we have proved the existence of $q(x)$ and $r(x)$ with the required properties in the case where S^* does not contain 0, and have hence proved our result in all cases.

Exercise 2.132. We saw in Exercise 2.130 that $\mathbb{Z}$ is a principal ideal domain. The key to that proof was the division algorithm in the integers. Now that we have established a corresponding division algorithm in the ring $F[x]$, where F is any field (see Exercise 2.131), we will use it to show that $F[x]$ is also a principal ideal domain.

Accordingly, let I be an ideal of $F[x]$. If I consists of just the element 0, then $I = \langle 0 \rangle$ and is already principal. Similarly, if $I = R$, then $I = \langle 1 \rangle$ (see Exercise 2.77.1) so I is principal in this case as well. So, assume in what follows that I is a nonzero *proper* ideal of R ("proper" simply means that $I \neq R$). In particular, I cannot contain any constant polynomials other than 0, since, if some nonzero $a \in F$ is in I, then $a \cdot a^{-1} = 1$ is also in I, contradicting what we have assumed about I.

Let $f(x)$ be a polynomial in I whose degree is least among all (nonzero) polynomials in I. (Such a polynomial exists by the Well-Ordering Principle.) Note that $f(x)$ must have positive degree by our assumption about I. Let $g(x)$ be an arbitrary polynomial in I. Apply the division algorithm and, using similar ideas as in Exercise 2.131, prove that $g(x)$ must a multiple of $f(x)$. Conclude that $I = \langle f(x) \rangle$.

Exercise 2.133. By contrast with the situation in Exercise 2.132 above, $\mathbb{Z}[x]$ is not a principal ideal domain! Prove this. (Hint: Show that the ideal $\langle 2, x\rangle$ of $\mathbb{Z}[x]$ cannot be generated by a single polynomial $f(x)$ with coefficients in $\mathbb{Z}$.)

Exercise 2.134. Let R be a ring, and let I and J be two ideals of R. The *sum* of I and J, denoted $I + J$, is the set $\{i + j \mid i \in I \text{ and } j \in J\}$ (i.e., it consists of all elements of R that are expressible as a sum of an element from I and an element from J). Show that $I + J$ is an ideal of R.

Exercise 2.135. Let R be a ring, and for simplicity, assume throughout this exercise that R is *commutative.* A proper ideal I is said to be *maximal* if for any other *proper* ideal J, $I \subseteq J$ implies that $I = J$.

1. Show that if I is maximal then for any other ideal J, $J \not\subseteq I$ implies $I + J = R$. (Hint: Assume that I is maximal and J is another ideal with $J \not\subseteq I$. Pick an element $j \in J - I$. Show that the set $K = \{i + rj \mid i \in I \text{ and } r \in R\}$ is an ideal of R. Now invoke the fact $I \subseteq K$.)

2. Show that the converse is true as well: if I is a proper ideal such that for any other ideal J, $J \not\subseteq I$ implies $I + J = R$, then I is maximal. (Hint: Assume that I has this property, and let J be a proper ideal with $I \subseteq J$. If $I \neq J$, then clearly $J \not\subseteq I$. The hypothesis then says $I + J = R$. But what else can you say about $I + J$ that then gives a contradiction?)

   (Thus either property could be used to define maximal ideals.)

3. Show that a proper ideal I is maximal if and only if R/I is a field. (Hint: Assume that I is maximal. Pick a nonzero element $[x]$ in R/I. Since $[x]$ is nonzero, $x \notin I$. Study the set $K = \{i + rx \mid i \in I \text{ and } r \in R\}$, which you showed in part (1) to be an ideal of R. By maximality of I show that there must be some $i \in I$ and $r \in R$ such that $i + rx = 1$. What does this relation read in R/I? Now invoke Exercise 2.47. A similar argument should also establish that if R/I is a field, then I must be maximal.)

It is instructive to note that maximal ideals always exist—see Theorem B.6 in Appendix B.

Exercise 2.136. Let R be a commutative ring. A proper ideal I of R is said to be *prime* if whenever $ab \in I$ for a and b in R, then either a or b must be in I.

1. Show that I is a prime ideal if and only if R/I is an integral domain. (Hint: This is just a matter of translating the definition of a prime ideal over to the ring R/I: for instance, assume that I is prime. If we have a relation $[a][b] = 0_{R/I}$ in R/I, then this means that $ab \in I$ in R.)

2. Show that every maximal ideal is necessarily prime.

3. Show that if p is a prime integer, then the ideal $\langle p \rangle$ in $\mathbb{Z}/p\mathbb{Z}$ is a prime ideal.

4. Let F be a field, and let $p(x)$ be an *irreducible* polynomial in $F[x]$. (This means that whenever $p(x) = q(x)r(x)$ for two polynomials in $F[x]$, then either $q(x)$ or $r(x)$ must be a constant polynomial.) Show that the ideal $\langle p(x)) \rangle$ is a prime ideal in $F[x]$.

Exercise 2.137. Let R be any ring containing the rationals. Prove that the only ring homomorphism $f : \mathbb{Q} \to R$ is the identity map that sends any rational number to itself. (Hint: given that $f(1)$ must be 1, what can you say about $f(2)$, $f(3)$, etc.? Next, what can you say about $f(1/2)$, $f(1/3)$, etc.? So now, what can you say about $f(m/n)$ for arbitrary integers m and n with $n \neq 0$?

Exercise 2.138. Prove that the following are all ring isomorphisms from $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$ to itself. Here, a, b, c, and d are, as usual, rational numbers.

1. The map that sends $a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6}$ to $a - b\sqrt{2} + c\sqrt{3} - d\sqrt{6}$.

2. The map that sends $a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6}$ to $a + b\sqrt{2} - c\sqrt{3} - d\sqrt{6}$.

3. The map that sends $a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6}$ to $a - b\sqrt{2} - c\sqrt{3} + d\sqrt{6}$.

(Of course, the identity map that sends $a+b\sqrt{2}+c\sqrt{3}+d\sqrt{6}$ to $a+b\sqrt{2}+c\sqrt{3} + d\sqrt{6}$ is also a ring isomorphism. It can be shown that these four are all the ring isomorphism from $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$ to itself.)

Notes

Remarks on Example 2.10 Every nonzero element in $\mathbb{Q}$ has a multiplicative inverse, that is, given any $q \in \mathbb{Q}$ with $q \neq 0$, we can find a rational number q' such that $qq' = 1$. The same cannot be said for the integers: not every nonzero integer has a multiplicative inverse within the integers. For example, there is no integer a such that $2a = 1$, so 2 does not have a multiplicative inverse.

Remarks on Example 2.12 The sum and product of any two elements $a + b\sqrt{2}$ and $c + d\sqrt{2}$ of $\mathbb{Q}[\sqrt{2}]$ are (respectively) $(a + c) + (b + d)\sqrt{2}$ and $(ac + 2bd) + (ad + bc)\sqrt{2}$. Since $a + c$, $b + d$, $ac + 2bd$ and $ad + bc$ are all rational numbers, the sum and product also lie in $\mathbb{Q}[\sqrt{2}]$. Thus, the standard method of adding and multiplying two real numbers of the form $x + y\sqrt{2}$ with x and y in $\mathbb{Q}$ indeed gives us binary operations on $\mathbb{Q}[\sqrt{2}]$. (In the language of the next section, $\mathbb{Q}[\sqrt{2}]$ is closed under addition and multiplication.) Now suppose you were trying to prove that, say, addition in $\mathbb{Q}[\sqrt{2}]$ is associative, that is, for any u, v, and w in $\mathbb{Q}[\sqrt{2}]$, $(u + v) + w = u + (v + w)$. Notice that in addition to being in $\mathbb{Q}[\sqrt{2}]$, u, v, and w are also *real numbers.* Since associativity holds in the reals, we find upon viewing u, v, and w as real numbers that $(u + v) + w = u + (v + w)$. Now viewing u, v, and w in this equation back again as elements of $\mathbb{Q}[\sqrt{2}]$, we find that associativity holds in $\mathbb{Q}[\sqrt{2}]$! This same argument holds for associativity of multiplication and distributivity of multiplication over addition. To prove that $a + b\sqrt{2} = 0$ iff $a = 0$ and $b = 0$, proceed as follows: If b is not zero, $a+b\sqrt{2} = 0$ yields $\sqrt{2} = -a/b$. Since $-a/b$ is a rational number, this contradicts Chapter 1, Exercise 1.42, so b must be zero. But if $b = 0$, $a + b\sqrt{2} = 0$ reads $a = 0$, so we find that both a and b are zero.

Remarks on Example 2.14 Assume that $a + \imath b = 0$. If b is not zero, we can write $\imath = a/b$, and squaring both sides, we find $-1 = a^2/b^2$. The right hand side is positive, since both a^2 and b^2 are positive (they are squares). But the left hand side is negative. Because of this contradiction, b must be zero. As before, we find that a is also zero.

Remarks on Examples 2.15 Given two elements a and b in $\mathbb{Z}_{(2)}$, write a as x/y where $\gcd(x, y) = 1$ and y is odd. (Why can you do this?) Write b as u/v where $\gcd(u, v) = 1$ and v is odd. Then $a + b = (xv + yu)/yv$. This fraction may not be reduced, but notice that yv, being a product of two odd integers, is odd. After you cancel all common factors from $(xv + yu)$ and yv, the resultant fraction will still have an odd denominator (why?). Hence $a + b$ will be in $\mathbb{Z}_{(2)}$. In a similar way, show that ab (gotten by the usual multiplication of two rational numbers) will also be in $\mathbb{Z}_{(2)}$. Now that you have two binary operations on $\mathbb{Z}_{(2)}$, you can check that the ring axioms hold. As with previous examples, associativity and distributivity follow from the fact that they hold for the rationals. Notice that the fact that the product of two odd integers is odd was essential in showing that both $a + b$ and ab lie in $\mathbb{Z}_{(2)}$. How could we generalize this? Rewrite this property in the contrapositive form, yv is even implies either y or v is even, that is, $2|yv$ implies $2|y$ or $2|v$. If we could find another integer n that has the property that

$n|yv$ implies $n|y$ or $n|v$, we could use the same arguments to show that $\mathbb{Z}_{(n)}$ is also a ring. (Assuming you found such an integer n, how would you define $\mathbb{Z}_{(n)}$?) Can you think of other integers that have this property? (Hint: You have come across such integers in the previous chapter.)

Remarks on Example 2.16 For $n = 1$, $M_n(\mathbb{R})$ is just $\mathbb{R}$, so it is commutative. For all other n, $M_n(\mathbb{R})$ is noncommutative. Given A in $M_n(\mathbb{R})$, write A as $(a_{i,j})$. (Recall what this notation means: you are referring to the (i,j)-th entry as "$a_{i,j}$.") Similarly, write B as $(b_{i,j})$ and C as $(c_{i,j})$. Consider $(A+B)+C$. What is the (i,j)-th entry of this resultant matrix? It is $(a_{i,j}+b_{i,j})+c_{i,j}$. On the other hand, what is the (i,j)-th entry of $A+(B+C)$? It is $a_{i,j}+(b_{i,j}+c_{i,j})$. Are the two (i,j)-th entries equal on both sides? Yes! Why? Because $a_{i,j}$, $b_{i,j}$, and $c_{i,j}$ are just real numbers, and *since addition is associative in* $\mathbb{R}$, $(a_{i,j} + b_{i,j}) + c_{i,j} = a_{i,j} + (b_{i,j} + c_{i,j})$! Since this is true for every pair (i,j), we find that $(A + B) + C = A + (B + C)$. (Notice how the associativity of addition in $M_n(\mathbb{R})$ *depends* on the associativity of addition in $\mathbb{R}$.) In a similar manner, try to prove the distributive property of multiplication over addition for $M_n(\mathbb{R})$; your proof should invoke the fact that distributivity holds in $\mathbb{R}$. Actually, if R is any ring, $M_n(R)$ is also a ring. It is noncommutative if $n \geq 2$. When $n = 1$, $M_n(R)$ is just R, so for $n = 1$, $M_n(R)$ is commutative if and only if R is commutative.

Remarks on Example 2.17 For any ring R, we can consider the set of polynomials with coefficients in R with the usual definition of addition and multiplication of polynomials. This will be a ring, with additive identity the constant polynomial 0 and multiplicative identity the constant polynomial 1. If R is commutative, $R[x]$ will also be commutative. (Why? Play with two general polynomials $f = \sum_{i=0}^{n} f_i x^i$ and $g = \sum_{j=0}^{m} g_j x^j$ and study fg and gf.) If R is not commutative, $R[x]$ will also not be commutative. To see this last assertion, suppose a and b in R are such that $ab \neq ba$. Then viewing a and b as constant polynomials in $R[x]$, we find that we get two different products of the "polynomials" a and b depending on the order in which we multiply them!

Here is something strange that can happen with polynomials with coefficients in an arbitrary ring R. First, the *degree* and *highest coefficient* of polynomials in $R[x]$ (where R is arbitrary) are defined exactly as for polynomials with coefficients in the reals. Now over $\mathbb{R}[x]$, if $f(x)$ and $g(x)$ are two nonzero polynomials, then

$deg(f(x)g(x)) = deg(f(x)) + deg(g(x)$. But for an arbitrary ring R, the degree of $f(x)g(x)$ *can be less than* $deg(f(x)) + deg(g(x))$*!*

To see why this is, suppose $f(x) = f_n x^n +$ lower-degree terms (with $f_n \neq 0$), and suppose $g(x) = g_m x^m +$ lower-degree terms (with $g_m \neq 0$). On multiplying out $f(x)$ and $g(x)$, the highest power of x that will show up in the product is x^{n+m}, and its coefficient will be $f_n g_m$. If we are working in $\mathbb{R}$, then $f_n \neq 0$ and $g_m \neq 0$ will force $f_n g_m$ to be nonzero, so the degree of $f(x)g(x)$ will be exactly $n+m$. But over arbitrary rings, it is quite possible for $f_n g_m$ to be zero even though f_n and g_m are themselves nonzero. (You have already seen examples of this in matrix rings. Elements a and b in a ring R such that $a \neq 0$ and $b \neq 0$ but $ab = 0$ will be referred to later in the chapter as zero-divisors.) When this happens, the highest *nonzero* term in $f(x)g(x)$ will be something lower than the x^{n+m} term, so the degree of $f(x)g(x)$ will be less than $n + m$!

Clearly, this phenomenon will not occur if the coefficient ring R does not have any zero-divisors. As will be explained further along in the chapter, *fields* do not have any zero-divisors (i.e., they are *integral domains*.) Hence if F is a field and $f(x)$ and $g(x)$ are two nonzero polynomials in $F[x]$, then $deg(f(x)g(x)) = deg(f(x)) + deg(g(x))$. (In particular, this shows that if F is any field, $F[x]$ also does not have zero-divisors—why?)

Remarks on Example 2.22 The additive identity is $(0, 0)$ and the multiplicative identity is $(1, 1)$. What is the product of $(1, 0)$ and $(0, 1)$? Of $(2, 0)$ and $(0, 2)$?

Remarks on some properties of rings deducible from the axioms Here is a hint for some of these properties listed in Remark 2.24:

1. *Uniqueness of additive identity:* Suppose 0 and $0'$ are two additive identities in a ring R. Consider the expression $0 + 0'$. First view 0 as the identity, and then view $0'$ as the identity. What do you find?

2. *Additive cancellation:* Given that $a + b = a + c$, what happens if you add the additive inverse of a to both sides of the equation, and use associativity?

3. $a \cdot 0 = 0 \cdot a = 0$. What happens if you invoke the fact that $0 = 0 + 0$ and multiply both sides by a?

4. $(-1)\cdot(-1) = 1$. You would by now already have proved that $a \cdot 0 = 0 \cdot a = a$ for all a in your ring R. Write $(-1)\cdot 0$ as $(-1)\cdot(1+(-1))$ and play with this!

Remarks on Definition 2.27 The requirement that 1 be in S arises from a rather nasty technical point that can be ignored during a first reading. If you are curious, recall first that "1" is merely notation for the multiplicative identity of R; we could just as easily have referred to it as "e" or something else all along. It turns out that if we defined subrings without the condition that 1 be in S, then it is possible for S to be a subring of R (under this hypothetical definition) with S and R having *different* multiplicative identities! This is a scenario we wish to avoid, and it turns out that insisting that the multiplicative identity of R (namely 1) be in S will take care of this problem. At the same time, it turns out that no such precaution needs to be taken for the *additive* identity—the additive identities of R and S will necessarily be equal. (The proof is simple: write 0_S and 0_R for the two additive identities, so we wish to prove that $0_S = 0_R$. We know that $1 \in S$. So, working in S, we find $0_S + 1 = 1$, by the very definition of 0_S. On the other hand, working in R, we find $0_R + 1 = 1$. Comparing the two expressions for 1 and *working in* R we find $1 = 0_S + 1 = 0_R + 1$. Additive cancellation in R now shows that $0_S = 0_R$.) This is of course all too pedantic for a first go around—we would do best by just accepting the definition above and getting on with our lives!

Remarks on Examples 2.33 Since every integer a can be written as $a/1$, and since 1 of course is 2^0, $\mathbb{Z} \subseteq \mathbb{Z}[1/2]$. Since 2 does not divide 1, every integer a $(= a/1)$ is also in $\mathbb{Z}_{(2)}$. Hence, $\mathbb{Z}[1/2] \cap \mathbb{Z}_{(2)}$ certainly contains $\mathbb{Z}$. Now let x be any rational number in $\mathbb{Z}[1/2] \cap \mathbb{Z}_{(2)}$. Since $x \in \mathbb{Z}[1/2]$, x can be written in the reduced form $a/2^n$, for some integer a and some $n \geq 0$. If $n > 0$, then x cannot be in $\mathbb{Z}_{(2)}$. Hence $n = 0$, that is, $x \in \mathbb{Z}$. It follows that $\mathbb{Z}[1/2] \cap \mathbb{Z}_{(2)}$ is precisely $\mathbb{Z}$.

Remarks on the notation $\mathbb{Q}[\sqrt{2}]$: Subring Generated by an Element

We have used notation like $\mathbb{Q}[\sqrt{2}]$, $\mathbb{Q}[\imath]$, $\mathbb{Z}[1/2]$, to denote various rings that we have studied. There is a reason for this notation: these are all examples of rings generated by a subring and an element. We consider this notion here.

We will consider only commutative rings, even though the notion exists for noncommutative rings as well. Accordingly, let R be a commutative ring, and let S be a subring. (Must S be commutative as well?) Let a be any element in R. (For instance, let R be the reals, let S be the rationals, and let a be the real number $1+\sqrt{2}$.) In general, $S \cup \{a\}$ will not be a subring of R, since this new set may not be closed under addition and multiplication. (In our example, the square of $1+\sqrt{2}$, which is $3+2\sqrt{2}$, is not in $\mathbb{Q} \cup \{1+\sqrt{2}\}$. Similarly, the sum of, say 2 and $1+\sqrt{2}$, which is $3+\sqrt{2}$ is not in $\mathbb{Q} \cup \{1+\sqrt{2}\}$.) One could then ask: If in general $S \cup \{a\}$ is not a subring of R, what are the elements of R that you should adjoin to the set $S \cup \{a\}$ to get a set that is actually a subring of R?

To get a subring of R that contains both S and a, it is clear that we need to be able to multiply a with itself any number of times, since our desired set must be closed under multiplication. Hence, we need to adjoin all the elements a^2, a^3,... Next, once all powers a^i are adjoined, we need to be able to multiply any power of a with any element of S, so we need to adjoin all products of the form sa^i, where s is an arbitrary element of S and a^i is an arbitrary power of a. (The assumption that R is commutative is being used here somewhere. Where exactly do you think it is used?) Once we have such products, we need to be able to add such products together if we are to have a ring (remember, our target set must be closed under addition), so we need to have all elements of the form $s_0 + s_1a + s_2a^2 + \cdots + s_na^n$, where the s_i are arbitrary elements of S, and $n \geq 0$. Is this enough? It turns out it is!

Definition 2.139. Let R be a commutative ring, S a subring, and a an element of R. An expression such as $s_0 + s_1a + s_2a^2 + \cdots + s_na^n$ is called a *polynomial expression in a with coefficients in S*. Let $S[a]$ denote the set of all polynomial expressions in a with coefficients in S, that is, the set of all elements of R that can be written in the form $s_0 + s_1a + s_2a^2 + \cdots + s_na^n$, for some $n \geq 0$, and some elements s_0, s_1,..., s_n in S. $S[a]$ is known as the *subring of R generated by S and a*. (If it is clear that we are working inside a fixed ring R, we often refer to $S[a]$ merely as the *ring generated by S and a*.)

Of course, we have blithely referred to $S[a]$ as a ring in the definition above, but we have yet to prove that $S[a]$ is actually a ring! We will do so in a moment.

Lemma 2.140. *Let R be a commutative ring, and let S be a subring of R. Let a be an element of R. The set $S[a]$ defined above is a subring of R.*

Proof. Since $S \subset S[a]$, and since $1 \in S$, 1 is in $S[a]$. Every element in $S[a]$ is of the form $s_0 + s_1a + s_2a^2 + \cdots + s_na^n$ for some $n \geq 0$ and some elements s_0, s_1,..., s_n in S. The negative of such an element is $(-s_0) + (-s_1)a + (-s_2)a^2 + \cdots + (-s_n)a^n$, which is also a polynomial expression in a with coefficients in S, and is hence in

$S[a]$. By Lemma 2.2.1, we only need to show that $S[a]$ is closed under addition and multiplication. You should be able to do this yourselves: show that the sum and product of two polynomial expressions in a with coefficients in S are also polynomial expressions in a with coefficients in S. $\square$

Notice that $S[a]$ includes both S and a. Our arguments preceding the lemma above show that any subring of R that contains both S and a *must* contain all polynomial expressions in a with coefficients in S, that is, it *must* contain $S[a]$. $S[a]$ should thus be thought of as the *smallest* subring of R that contains both S and a.

Here is an exercise: In the setup above, if two polynomial expressions $s_0+s_1a+s_2a^2+\cdots+s_na^n$ and $s'_0+s'_1a+s'_2a^2+\cdots+s'_ma^m$ are equal (as elements of R), can you conclude that $n=m$ and $s_i=s'_i$ for $i=0,\ldots,n$? (Hint: See the examples below.)

Now let us consider some examples:

Example 2.141. What, according to our definition above, is the subring of the reals generated by $\mathbb{Q}$ and $\sqrt{2}$? It is the set of all polynomial expressions in $\sqrt{2}$ with coefficients in $\mathbb{Q}$, that is, the set of all expressions of the form $q_0+q_1\sqrt{2}+q_2(\sqrt{2})^2+\cdots+q_n(\sqrt{2})^n$. Now let us look at these expressions more closely. Since $(\sqrt{2})^2=2$, $q_2(\sqrt{2})^2$ is just $2q_2$, $q_4(\sqrt{2})^4$ is just $4q_4$, etc. Similarly, $q_3(\sqrt{2})^3$ is just $2q_3\sqrt{2}$, $q_5(\sqrt{2})^5$ is just $4q_5\sqrt{2}$, etc. By collecting terms together, it follows that every polynomial expression in $\sqrt{2}$ with coefficients in $\mathbb{Q}$ can be rewritten as $a+b\sqrt{2}$ for suitable rational numbers a and b. (For example, $1+2\sqrt{2}+(1/2)(\sqrt{2})^2+(1/4)(\sqrt{2})^3$ can be rewritten as $2+(5/2)\sqrt{2}$.) Hence, the subring of the reals generated by the rationals and $\sqrt{2}$ is the set of all real numbers of the form $a+b\sqrt{2}$. It is for this reason that we denoted this ring $\mathbb{Q}[\sqrt{2}]$ as far back as Example 2.12.

Example 2.142. Similarly, the subring of $\mathbb{Q}[\sqrt{2}]$ generated by $\mathbb{Z}$ and $\sqrt{2}$ is the set of all real numbers of the form $a+b\sqrt{2}$, where a and b are integers. This is why we denoted this ring $\mathbb{Z}[\sqrt{2}]$ in Example 2.31.

Example 2.143. Using the fact that $i^2=-1$, show that the subring of $\mathbb{C}$ generated by $\mathbb{Q}$ and i is the set of all complex numbers of the form $a+bi$, where a and b are rational numbers. This explains the notation $\mathbb{Q}[\imath]$ for the ring in Example 2.14.

Example 2.144. Similarly, the subring of $\mathbb{Q}[\imath]$ generated by $\mathbb{Z}$ and i is is the set of all complex numbers of the form $a + bi$, where a and b are integers. Hence the notation $\mathbb{Z}[\imath]$ in Example 2.32.

Example 2.145. Show that the subring of $\mathbb{Q}$ generated by $\mathbb{Z}$ and $1/2$ is the set of all rational numbers that have the property that, when written in the reduced form a/b with $\gcd(a, b) = 1$, the denominator b is a power of 2. This explains the notation $\mathbb{Z}[1/2]$ in Example 2.33.

Example 2.146. Prove that the subring of $\mathbb{R}$ generated by $\mathbb{Q}[\sqrt{2}]$ and $\sqrt{3}$ is precisely the ring of Example 2.34. Thus, this ring should be denoted $\mathbb{Q}[\sqrt{2}][\sqrt{3}]$. We will often avoid using the second pair of brackets and simply refer to this ring as $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$.

Here is a quick exercise: In Lemma 2.140, suppose a is actually in S. Can you prove that the ring generated by S and a is just S?

Remarks on Definition 2.46 Most textbooks define a field to be a commutative ring in which every nonzero a is invertible. In other words, the extra condition that we have imposed, namely that the ring in question first be an integral domain, is omitted by most textbooks. This is because this extra condition is not really required—one can show easily that any commutative ring in which every nonzero element a is invertible must *necessarily* be an integral domain. (If there were to exist a pair of nonzero elements a and b such that $ab = 0$, then multiplying both sides by a^{-1}, which exists by hypothesis, we would find $b = 0$, a contradiction. Hence there can be no pair of nonzero elements that multiply out to zero.) The reason we have chosen to define a field as an *integral domain* in which every nonzero element is invertible is to highlight the hierarchical nature of the objects that we have been considering: rings are fairly general objects, commutative rings are special rings that are nicer to deal with, integral domains are special commutative rings that are even nicer, and finally, fields are special integral domains that are nicest of all!

Chapter 3

Vector Spaces

3.1 Vector Spaces: Definition and Examples

Recall from elementary linear algebra the notation $\mathbb{R}^2$ for 2-dimensional xy space and $\mathbb{R}^3$ for 3-dimensional xyz space. A *vector* in $\mathbb{R}^2$ (respectively $\mathbb{R}^3$) is an arrow with its base at the origin and its tip at some point in $\mathbb{R}^2$ (respectively $\mathbb{R}^3$). If v and w are vectors, then we *add* v and w using the parallelogram law. We know that this process of addition is commutative, that is, $v + w = w + v$ for all vectors v and w. Vector addition is also associative, that is, $v+(w+u) = (v+w)+u$ for all vectors v, w, and u. The vector whose base *and* tip are at the origin is denoted 0 (suggestively), and satisfies $v+0 = 0+v$ for all vectors v. Finally, for every vector v, the vector we get by inverting v about the origin is denoted $-v$ (also suggestively), and satisfies $v + (-v) = (-v) + v = 0$.

Focusing just on $\mathbb{R}^2$ for convenience, let us stop thinking of $\mathbb{R}^2$ as a geometric object. Instead, since every point of $\mathbb{R}^2$ corresponds to a vector whose tip is at the given point, let us consider $\mathbb{R}^2$ as a set consisting of abstract objects called *vectors*. This set has a binary operation defined on it—addition, where $v + w$ is defined as the vector we get by temporarily reverting to the geometric interpretation of $\mathbb{R}^2$ as a plane and considering the vector obtained as the diagonal of the parallelogram formed by v and w. What do you notice about this set of vectors with this binary operation? The binary operation satisfies all the axioms for an abelian group! Thus, in addition to being a geometric object (the plane), $\mathbb{R}^2$, when considered as a set with a binary operation, has an algebraic structure—it is an abelian group!

But there is more. Let us go back to the interpretation of $\mathbb{R}^2$ as 2-dimensional xy space, and let us recall the notion of scalar multiplication. A *scalar* is any real number, and given a scalar r and a vector v, we multiply r and v according to the following definition—if $r \geq 0$, then $r \cdot v$ is the vector in the same direction as v but whose length is r times the length of v, and if $r < 0$, then $r \cdot v$ is the vector in the opposite direction as v but whose length is $|r|$ times the length of v. What are the properties of scalar multiplication? If r and s are any two scalars, and if v and w are any two vectors, we have the following: $r \cdot (v + w) = r \cdot v + r \cdot w$, $(r + s) \cdot v = r \cdot v + s \cdot v$, $(rs) \cdot v = r \cdot (s \cdot v)$, and $1 \cdot v = v$.

Observe that the set of scalars, namely the real numbers, is a *field.* Now, let us attempt to generalize all this. In the case of $\mathbb{R}^2$ above, we have seen that the geometric interpretation of $\mathbb{R}^2$ as 2-dimensional xy space furnishes us with the notion of vector addition and scalar multiplication, but once these definitions have been furnished, $\mathbb{R}^2$ seems to have an *algebraic* life of its own. For instance, $(\mathbb{R}^2, +)$ is an abelian group, while scalar multiplication has the (algebraic) properties listed above. Could similar sets of objects called *vectors* and *scalars* not arise in different circumstances, with the same properties as the ones listed above, but with the vector addition and scalar multiplication perhaps defined by some process other than a geometric one? The answer is *yes*, and in fact, they arise in vastly different situations. As with the other concepts that we have seen (groups, rings, fields, etc.), it is worth isolating this phenomenon and studying it in its own right.

Definition 3.1. Let F be a field. A *vector space over* F (also called an F–*vector space*) is an abelian group V together with a function $F \times V \to V$ called *scalar multiplication* and denoted $\cdot$ such that for all r and s in F and v and w in V,

1. $r \cdot (v + w) = r \cdot v + r \cdot w$,
2. $(r + s) \cdot v = r \cdot v + s \cdot v$,
3. $(rs) \cdot v = r \cdot (s \cdot v)$, and
4. $1 \cdot v = v$.

The elements of V are called *vectors* and the elements of F are called *scalars*.

Thus, $\mathbb{R}^2$ and $\mathbb{R}^3$ are both vector spaces over $\mathbb{R}$. Let us look at several examples of vector spaces that arise from other than geometric considerations:

Example 3.2. We have looked at $\mathbb{R}^2$ and $\mathbb{R}^3$, why not generalize these, and consider $\mathbb{R}^4$, $\mathbb{R}^5$, etc.? These would of course correspond to higher-dimensional "worlds." It is certainly hard to visualize such spaces, but there is no problem considering them in a purely algebraic manner. Recall that every vector in $\mathbb{R}^2$ can be described *uniquely* by the *pair* (a, b), consisting of the x and y components of the vector. ("Uniquely" means that the vector (a, b) equals the vector (a', b') if and only if $a = a'$ and $b = b'$.) Similarly, every vector in $\mathbb{R}^3$ can be described uniquely by the *triple* (a, b, c), consisting of the x, y, and z components of the vector. Thus, $\mathbb{R}^2$ and $\mathbb{R}^3$ can be described respectively as the set of all pairs (a, b) and the set of all triples (a, b, c), where a, b, and c are arbitrary real numbers. Proceeding analogously, for any positive integer n, we will let $\mathbb{R}^n$ denote the set of n-*tuples* $(a_1, a_2, \ldots, a_n)$, where the a_i are arbitrary real numbers. (As with $\mathbb{R}^2$ and $\mathbb{R}^3$, the understanding here is that two n-tuples $(a_1, a_2, \ldots, a_n)$ and $(a'_1, a'_2, \ldots, a'_n)$ are equal if and only if their respective components are equal, that is, $a_1 = a'_1$, $a_2 = a'_2$, $\ldots$, and $a_n = a'_n$.) These n-tuples will be our vectors; how should we add them? Recall that in $\mathbb{R}^2$ we add the vectors (a, b) and (a', b') by adding a and a' together and b and b' together, that is, by adding *componentwise*.

Exercise 3.2.1. Deduce from the parallelogram law of addition of vectors in $\mathbb{R}^2$ that the sum of (a, b) and (a', b') is $(a + a', b + b')$.

We will do the same with $\mathbb{R}^n$—we will decree that $(a_1, a_2, \ldots, a_n) + (a'_1, a'_2, \ldots, a'_n) = (a_1 + a'_1, a_2 + a'_2, \ldots, a_n + a'_n)$.

Exercise 3.2.2. Check that with this definition of addition, $(\mathbb{R}^n, +)$ is an abelian group.

What should our scalars be? Just as in $\mathbb{R}^2$ and $\mathbb{R}^3$, let us take our scalars to be the field $\mathbb{R}$. How about scalar multiplication? In $\mathbb{R}^2$, the product of the scalar r and the vector (a, b) is (ra, rb), that is, we multiply each component of the vector (a, b) by the real number r. (Is that so? Check!) We will multiply scalars and vectors in $\mathbb{R}^n$ in the same way: we will decree that the product of the real number r and the n-tuple $(a_1, a_2, \ldots, a_n)$ is $(ra_1, ra_2, \ldots, ra_n)$.

Exercise 3.2.3. Check that this definition satisfies the axioms of scalar multiplication in Definition 3.1.

Thus, $\mathbb{R}^n$ is a vector space over $\mathbb{R}$.

Example 3.3. Now, why restrict the examples above to n-tuples of $\mathbb{R}$? For *any* field F, let F^n stand for the set of n-tuples $(a_1, a_2, \ldots, a_n)$, where the a_i are arbitrary elements of F. Add two such n-tuples componentwise, that is, define addition via the rule $(a_1, a_2, \ldots, a_n) + (a'_1, a'_2, \ldots, a'_n) = (a_1 + a'_1, a_2 + a'_2, \ldots, a_n + a'_n)$. Take the field F to be the field of scalars, and define scalar multiplication just as in $\mathbb{R}^n$: given an arbitrary $f \in F$, and an arbitrary n-tuple $(a_1, a_2, \ldots, a_n)$, define their scalar product to be the n-tuple $(fa_1, fa_2, \ldots, fa_n)$.

Exercise 3.3.1. Check that these definitions of vector addition and scalar multiplication make F^n a vector space over F.

Taking $F = \mathbb{C}$ and $n = 2$ for instance, we get *complex 2-space*, which, for example, is a natural arena in which to study plane curves.

Example 3.4. Similarly, for any field F, let $\prod_0^\infty F$ denote the set of all *infinite*-tuples $(a_0, a_1, a_2, \dots)$, where the a_i are in F. (It is convenient in certain applications to index the components from 0 rather than 1, but if this bothers you, it is harmless to think of the tuples as $(a_1, a_2, a_3, \dots)$.) Addition and scalar multiplication are defined just as in F^n, except that we now have infinitely many components. With these definitions, $\prod_0^\infty F$ becomes an F–vector space. (This example is known as the *direct product* of (countably infinite) copies of F.)

Example 3.5. Consider the ring $M_n(\mathbb{R})$. Focusing just on the addition operation on $M_n(\mathbb{R})$, recall that $(M_n(\mathbb{R}), +)$ is an abelian group. (Remember, for any ring R, $(R, +)$ is always an abelian group.) We will treat the reals as scalars. Given any real number r and any matrix $(a_{i,j})$ in $M_n(\mathbb{R})$, we will define their product to be the matrix $(ra_{i,j})$. (See the notes on page 153 for a comment on this product.) Verify that with this definition, $M_n(\mathbb{R})$ is a vector space over $\mathbb{R}$. In a similar manner, if F is any field, $M_n(F)$ will be a vector space over F.

Example 3.6. Consider the field $\mathbb{Q}[\sqrt{2}]$. Then $(\mathbb{Q}[\sqrt{2}], +)$ is an abelian group (why?). Think of the rationals as scalars. There is a very natural way of multiplying a rational number q with an element $a+b\sqrt{2}$ of $\mathbb{Q}[\sqrt{2}]$, namely, $q\cdot(a+b\sqrt{2}) = qa+qb\sqrt{2}$. With this definition of scalar multiplication, check that $\mathbb{Q}[\sqrt{2}]$ becomes a vector space over the rationals.

If you probe this example a little harder, you may come up with an apparent anomaly. What exactly is the role of the rationals here? True, we want to think of the rationals as scalars. However, $\mathbb{Q} \subseteq \mathbb{Q}[\sqrt{2}]$, so every rational number is also an element of $\mathbb{Q}[\sqrt{2}]$, and is therefore also a vector! How do we resolve this conflict? As it turns out, there really is nothing to resolve, we merely accept the fact that the rationals have a dual role in this example! When we see a rational number "q" by itself, we want to think of it as $q + 0\sqrt{2}$, that is, we want to think of q as an element of $\mathbb{Q}[\sqrt{2}]$, or in other words, we want to think of q as a vector. However, when we see q

in an expression like $q(a + b\sqrt{2})$, we want to think of q as a scalar, that is, something we multiply vectors by!

Example 3.7. Let us generalize Example 3.6. What we needed above were that

1. $\mathbb{Q}[\sqrt{2}]$ is a field, so $(\mathbb{Q}[\sqrt{2}], +)$ is automatically an abelian group, and
2. $\mathbb{Q} \subseteq \mathbb{Q}[\sqrt{2}]$, so that we could use the natural multiplication inside $\mathbb{Q}[\sqrt{2}]$ to multiply any $q \in \mathbb{Q}$ with any $a + b\sqrt{2} \in \mathbb{Q}[\sqrt{2}]$.

These two facts together gave us a $\mathbb{Q}$–vector space structure on $\mathbb{Q}[\sqrt{2}]$. Now let K/F be *any* field extension. Since K is a field, $(K, +)$ is an abelian group. Next, let us consider multiplication. Given any two elements k and l of K, we know we can multiply the two elements together. However, let us ignore this fact temporarily, and just consider the fact that given any element f of F and any element k of K, we can multiply f and k. (Notice that we have restricted the first element to be from F. However, we have placed no restriction on the second element, it can be any element of K. This is just like considering the multiplication of any $q \in \mathbb{Q}$ and any $a+b\sqrt{2} \in \mathbb{Q}[\sqrt{2}]$ in Example 3.6 above.) Now note the following properties of this (restricted) multiplication, which are just consequences of the properties of the (unrestricted) multiplication in K: If f and g are any two elements of F, and k and l are any two elements of K, then 1) $f \cdot (k+l) = f \cdot k + f \cdot l$, 2) $(f + g) \cdot k = f \cdot k + g \cdot k$, 3) $(fg) \cdot k = f \cdot (g \cdot k)$, and 4) $1 \cdot k = k$. (In this last property, we consider 1 as an element of F.) What do we notice? If we take the field F as our scalars, $(K, +)$ as our vectors, and the multiplication operation between elements of F and elements of K (that arises from the multiplication operation on K) as scalar multiplication, then, just as in Example 3.6 above, K becomes an F–vector space!

Also, exactly as in Example 3.6 above, the elements of F have a dual role, both as scalars and as vectors. When we see an element $f \in F$ by itself, f is playing the role of a vector. But when we see an element $f \in F$

in an expression like $f \cdot k$, f is playing the role of a scalar that is multiplying the vector k!

Example 3.8. Now let us generalize Example 3.6 even further, by once again considering the two conditions at the beginning of Example 3.7. Do we really need the full force of the fact that $\mathbb{Q}[\sqrt{2}]$ is a *field?* No, all we need is the fact that $\mathbb{Q}[\sqrt{2}]$ is a *ring* that contains the field $\mathbb{Q}$; this is enough to provide an abelian group structure on $(\mathbb{Q}[\sqrt{2}], +)$ and to furnish a scalar product between elements of $\mathbb{Q}$ and elements of $\mathbb{Q}[\sqrt{2}]$. Now let R be any ring that contains a field F. Then just as in Example 3.7 above, $(R, +)$ is an abelian group, and we can use the multiplication in R to define the scalar product between any element f of F and any element r of R. This multiplication clearly satisfies the scalar product axioms in Definition 3.1, so R becomes an F–vector space. Just as in Example 3.7 above, the elements of F have a dual role, both as scalars and as vectors.

Here is a familiar instance of this phenomenon. Consider the polynomial ring $\mathbb{R}[x]$. This ring contains $\mathbb{R}$ (since every real number r lives inside $\mathbb{R}[x]$ as the constant polynomial $r + 0x + 0x^2 + \cdots$). Thus, $\mathbb{R}[x]$ is a vector space over $\mathbb{R}$. Explicitly, the scalar product of any real number r and any polynomial $f = \sum_{i=0}^{n} a_i x^i$ (where the a_i are real numbers and n is some nonnegative integer) is the polynomial $\sum_{i=0}^{n} r a_i x^i$. The real numbers have a dual role here: when we see a real number r by itself, we want to think of it as a vector, and when we see it in an expression $r \cdot f$, we want to think of it as a scalar multiplying the vector f.

In the same vein, $F[x]$ is an F–vector space for any field F.

Example 3.9. Here is an example related to $F[x]$. For any field F and any nonnegative integer n, write $F_n[x]$ for the set of all polynomials in x with coefficients in F whose degrees are *at most* n. Then $F_n[x]$ is an F–vector space.

Question 3.9.1. Why?

Example 3.10. Now think about this: Suppose V is a vector space over a field K. Suppose F is a subfield of K. Then V is also a vector space over F!

> **Question 3.10.1.** Why? What do you think the scalar multiplication ought to be? (See the notes on page 153 for some remarks on this.)

As an example of this phenomenon, $\mathbb{R}[x]$, besides being an $\mathbb{R}$–vector space, is also a $\mathbb{Q}$–vector space. Vector addition is the usual addition of polynomials. As for scalar multiplication, when we consider $\mathbb{R}[x]$ as a $\mathbb{Q}$–vector space, we only allow multiplication of polynomials by *rational* numbers—we ignore the fact that we can multiply polynomials by arbitrary real numbers.

Similarly, $M_2(\mathbb{Q}[\sqrt{2}]$, besides being a $\mathbb{Q}[\sqrt{2}]$–vector space, is also a $\mathbb{Q}$–vector space.

Example 3.11. Here is an example that may seem pathological at first, but is not really so! Consider the *trivial abelian group* V*:* this consists of a single element, namely, the identity element 0_V. The only addition rule here is $0_V + 0_V = 0_V$, and it is easy to check that the set $\{0_V\}$ with the addition rule above is indeed an abelian group. Now let F be any field. Then V is a vector space over F with the product rule $f \cdot 0_V = 0_V$. There is only vector in this space, namely 0_V, although, there are lots of scalars! This vector space is known as the *trivial* vector space or the *zero* vector space over F, and shows up quite naturally as kernels of injective linear transformations (see Lemma 3.87 ahead, for instance).

Remark 3.12. Now observe that all these examples of vector spaces have the following properties:

1. For any scalar f, f times the zero vector is just the zero vector.
2. For any vector v, the *scalar* 0 times v is the zero *vector.*
3. For any scalar f and any vector v, $(-f) \cdot v = -(f \cdot v)$.

4. If v is a nonzero vector, then $f \cdot v = 0$ for some scalar f implies $f = 0$.

These properties somehow seem very natural, and one would expect them to hold for *all* vector spaces. Just as in Remark 2.24, where we considered a similar set of properties for rings, we would like these properties to be deducible from the vector space axioms themselves. This would, among other things, convince us that our vector space axioms are the "correct" ones, that is, they yield objects that behave more or less like the examples above instead of objects that are rather pathological. As it turns out, our expectations are not misguided: these properties *are* deducible from the vector space axioms, and therefore *do* hold in all vector spaces. We will leave the verification of this to the exercises (see Exercise 3.97).

3.2 Linear Independence, Bases, Dimension

Now, given a field F and an F–vector space V, it is natural to wonder about the *size* of V. To measure this size, we need to consider the concept of the *dimension* of a vector space.

Let us contrast $\mathbb{R}^2$ with $\mathbb{R}^3$. We all share the intuition that $\mathbb{R}^3$ is somehow *bigger* than $\mathbb{R}^2$. But what precisely is it about $\mathbb{R}^2$ and $\mathbb{R}^3$ that makes us feel that one is bigger than the other? If we examine our intuition a little more closely, we discover that the reason that $\mathbb{R}^3$ seems bigger than $\mathbb{R}^2$ is that $\mathbb{R}^3$ has *three* coordinate axes, while $\mathbb{R}^2$ has only two. *Hidden in this fact is the concept of the dimension of a vector space.* And in fact, without necessarily having paused to think through the notion of dimension or make it precise, most of us have already absorbed this concept and integrated it into our lives—we readily describe $\mathbb{R}^2$ as a *2-dimensional space* and $\mathbb{R}^3$ as a *3-dimensional space.*

With this in mind, what should we take to be the dimension of a vector space? The number of coordinate axes it contains? As it turns out, this is indeed correct, but we have some work to do first. Remember, a vector space is an *algebraic* object. It is defined as an abelian group $(V, +)$ along with a scalar multiplication $F \times V \to V$ with the properties that we have described above. Thus, while the term "coordinate axes" has an obvious meaning in the geometric examples of $\mathbb{R}^2$ and $\mathbb{R}^3$, it is not clear what meaning it should have in a general vector space. So our first task is to convert the geometric notion of coordinate axes into an *algebraic* notion. Next, we need to worry about the possibility that an arbitrary vector space defined purely algebraically may not have any coordinate axes at all, as well as the possibility that different sets of coordinate axes of the same vector space may have different numbers of axes in each set! If either of these possibilities were to occur, we would not have a unique number that we could assign as the dimension of the vector space. As it turns out, neither of these can happen, and our second task is to consider the impossibility of these two scenarios.

Let us turn to the first task. Focusing on $\mathbb{R}^2$ for convenience, let us denote the vector with tip at the point $(1, 0)$ by $\mathbf{i}$, and the one with the tip at the point $(0, 1)$ by $\mathbf{j}$. From vector calculus, we know that if we take an arbitrary vector in $\mathbb{R}^2$, say $\mathbf{u}$, with its tip at (a, b), then the projection of $\mathbf{u}$ onto the x-axis is just a times the vector $\mathbf{i}$ and the projection on the y-axis is just b times the vector $\mathbf{j}$. The parallelogram law then shows that $\mathbf{u}$ is the sum of $a \cdot \mathbf{i}$ and $b \cdot \mathbf{j}$, that is, $\mathbf{u} = a \cdot \mathbf{i} + b \cdot \mathbf{j}$. Since $\mathbf{u}$ was an arbitrary vector in this discussion, we find that *every* vector in $\mathbb{R}^2$ can be written as a scalar times $\mathbf{i}$ added to another scalar times $\mathbf{j}$. This example motivates two definitions.

Definition 3.13. Let V be a vector space over a field F. A linear combination of vectors $v_1, \cdots, v_n$ (or, an F-*linear* combination of vectors $v_1, \cdots, v_n$, if we wish to emphasize the field over which the vector space is defined) is any vector in V that can be written as $f_1 \cdot v_1 + \cdots + f_n \cdot v_n$ for suitable scalars $f_1, \cdots, f_n$.

Thus, what we found above is that every vector in $\mathbb{R}^2$ can be written as a $\mathbb{R}$-linear combination of the vectors $\mathbf{i}$ and $\mathbf{j}$. (To give you more examples, the vectors $\mathbf{i}+\mathbf{j}$, $\sqrt{2}\mathbf{i}-3\mathbf{j} = \sqrt{2}\mathbf{i}+(-3)\mathbf{j}$, and $\pi\mathbf{i}+3\pi^2\mathbf{j}$ are all linear combinations of $\mathbf{i}$ and $\mathbf{j}$.)

The other definition motivated by the example of the vectors $\mathbf{i}$ and $\mathbf{j}$ in $\mathbb{R}^2$ is the following:

Definition 3.14. Let V be a vector space over a field F. A subset S of V is said to *span* V (or S is said to be a *spanning set for* V) if every vector $v \in V$ can be written as $\sum_{i=1}^{n} f_i \cdot v_i$ for some integer $n \geq 1$, some choice of vectors v_1, $\dots$, v_n from S, and some choice of scalars f_1, $\dots$, f_n. (In the language of Definition 3.13 above, S is a spanning set for V if every vector in V is expressible as a *linear combination* of some elements of S.

The discussion before Definition 3.13 showed that the set $S = \{\mathbf{i}, \mathbf{j}\}$ is a spanning set for $\mathbb{R}^2$. Here are more examples:

Example 3.15. We have seen in Example 3.6 that $\mathbb{Q}[\sqrt{2}]$ is a $\mathbb{Q}$–vector space. Note that every element of $\mathbb{Q}[\sqrt{2}]$ is of the form $a + b\sqrt{2}$ for suitable a and $b \in \mathbb{Q}$. Thinking of "a" as "$a \cdot 1$," this tells us that every element of $\mathbb{Q}[\sqrt{2}]$ is expressible as a $\mathbb{Q}$-linear combination of 1 and $\sqrt{2}$. (We are thinking of 1 as a vector in this last statement. Recall the discussion of the dual role of $\mathbb{Q}$ in Example 3.6.) Hence, $S = \{1, \sqrt{2}\}$ is a spanning set for the $\mathbb{Q}$-vector space $\mathbb{Q}[\sqrt{2}]$.

Example 3.16. The set $\{1, x, x^2, \dots\}$ is a spanning set for the polynomial ring $\mathbb{R}[x]$ considered as a vector space over $\mathbb{R}$ (see Example 3.8 above). This is clear since every polynomial in $\mathbb{R}[x]$ is of the form $r_0 + r_1 x + \cdots + r_n x^n$ for some integer $n \geq 0$ and suitable real numbers r_0, r_1, $\dots$, r_n. Put differently, every polynomial can be expressed as a $\mathbb{R}$-linear combination of 1, x, $\dots$, x^n for some integer $n \geq 0$. Since different polynomials have different degrees, we need to use all powers x^i $(i = 1, 2, \dots)$ to get a spanning set for $\mathbb{R}[x]$.

Remark 3.17. By convention, the empty set is taken as a spanning set for the zero vector space. Moreover, by convention, the trivial space is the only space spanned by the empty set. This convention will be useful later, when defining the dimension of the zero vector space.

So, returning to our study of dimension, should we take the algebraic analog of coordinate axes to be any set S of vectors that span V? No, not yet! There could be *redundancy* in this set! It may turn out, for example, that the smaller set $S - \{v\}$ obtained by deleting a particular vector v from the set already spans V! (If so, why bother using this vector v as one of coordinate axes?!?)

Let us formulate this as a definition:

Definition 3.18. Given a vector space V over a field F, a vector v in a spanning set S is said to be *redundant* if the subset $S - \{v\}$ obtained by removing v is itself a spanning set for V. (Put differently, v is redundant in S if every vector in V can already be expressed as a linear combination of elements in $S - \{v\}$, so the vector v is not needed at all.) We will say that there is redundancy in the spanning set S if any one of the vectors in this set is redundant.

Example 3.19. For an example of a spanning set with redundancy in it, we do not have to look very far: Going back to $\mathbb{R}^2$, let us write $\mathbf{w}$ for the vector with tip at $(1/\sqrt{2}, 1/\sqrt{2})$. Then $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{w}$ also span $\mathbb{R}^2$.

> **Question 3.19.1.** This is of course very trivial to see—the vector with tip at (a, b) can be written as the sum $a \cdot \mathbf{i} + b \cdot \mathbf{j} + 0 \cdot \mathbf{w}$. More interestingly, can you show that it can also be written as $(a - 1/\sqrt{2}) \cdot \mathbf{i} + (b - 1/\sqrt{2}) \cdot \mathbf{j} + \mathbf{w}$?

Since $\mathbf{i}$ and $\mathbf{j}$ already span $\mathbb{R}^2$, there is clearly redundancy in the set $\{\mathbf{i}, \mathbf{j}, \mathbf{w}\}$.

To push this example a bit further, note that $\mathbf{i}$ and $\mathbf{w}$ also form a spanning set for $\mathbb{R}^2$. To see this, note that $\mathbf{j} = -\mathbf{i} + \sqrt{2}\mathbf{w}$. Thus, any vector $a \cdot \mathbf{i} + b \cdot \mathbf{j}$ in $\mathbb{R}^2$, can be written as $(a - b) \cdot \mathbf{i} + (\sqrt{2}b) \cdot \mathbf{w}$ by simply substituting $-\mathbf{i} + \sqrt{2}\mathbf{w}$ for $\mathbf{j}$. This shows that $\mathbf{i}$ and $\mathbf{w}$ also span $\mathbb{R}^2$.

Notice that there is no redundancy in the set $\{\mathbf{i}, \mathbf{w}\}$, because if you remove, say $\mathbf{w}$, then the remaining vector $\mathbf{i}$ alone will not span $\mathbb{R}^2$: the various "linear combinations of $\mathbf{i}$" are the vectors of the form $r\mathbf{i}$, where r is an arbitrary real number, and these are all aligned with the vector $\mathbf{i}$ and will therefore not give all of $\mathbb{R}^2$. (Similarly, if you remove $\mathbf{i}$, the linear combinations of remaining vector $\mathbf{w}$ will all be aligned with $\mathbf{w}$ and will not give all of $\mathbb{R}^2$.)

Question 3.19.2. Similarly, can you show that $\mathbf{j}$ and $\mathbf{w}$ also span $\mathbb{R}^2$.

In this example, we would of course take the set $\{\mathbf{i}, \mathbf{j}\}$ as coordinate axes for $\mathbb{R}^2$, as is the usual practice, but we could just as easily take the set $\{\mathbf{i}, \mathbf{w}\}$ or the set $\{\mathbf{j}, \mathbf{w}\}$ as coordinate axes.

Example 3.20. Let S be a spanning set for a vector space V. If v is any vector in V that is *not* in S, then $S \cup \{v\}$ is also a spanning set for V in which there is redundancy. More generally, if T is any nonempty subset of V that is disjoint from S, then $S \cup T$ is also a spanning set for V in which there is redundancy.

Exercise 3.20.1. Convince yourself of this!

For instance, we have seen in Example 3.16 above that the set $\{1, x, x^2, \dots\}$ is a spanning set for the polynomial ring $\mathbb{R}[x]$ considered as a vector space over $\mathbb{R}$. Taking $T = \{1 + x, 1 + x + x^2, 1 + x + x^2 + x^3, \dots\}$, it follows that the set $U = \{1, x, 1 + x, x^2, 1 + x + x^2, x^3, 1 + x + x^2 + x^3, \dots\}$ is a spanning set for $\mathbb{R}[x]$ in where there is redundancy.

Exercise 3.20.2. Continuing with the example of the polynomial ring $\mathbb{R}[x]$ considered as a vector space over $\mathbb{R}$, show that there is no redundancy in the spanning set $\{1, x, x^2, \dots\}$.

Remember, we are trying to formulate an algebraic definition of coordinate axes. Our intuition from Example 3.19, as well as Example 3.16 and Exercise 3.20.2 above, would suggest that a set of coordinate axes, first, should span our vector space, and next, should not have more vectors than are needed to span the space, that is, should not have redundancy in it.

It would be very useful to have alternative characterizations of redundancy. We have the following:

Lemma 3.21. *Let V be a vector space over a field F, and let S be a spanning set for V. Then, the following are equivalent:*

1. *There is redundancy in S,*

2. *Some vector v in S is expressible as a linear combination of some vectors from the set $S - \{v\}$, and*

3. *There exist a positive integer m and scalars $f_1, \ldots, f_m$, not all zero, such that for some vectors $v_1, \ldots, v_m$ from S, we have the relation $f_1 \cdot v_1 + \cdots + f_m \cdot v_m = 0$.*

Proof. Let us prove the implications (1) $\Rightarrow$ (2) $\Rightarrow$ (3) $\Rightarrow$ (1).
(1) $\Rightarrow$ (2): Given that there is redundancy in S, this means that there is some $v \in S$ such that $S - \{v\}$ is already a spanning set for V. Thus, by definition of what it means to span V, there are some vectors $v_1, \ldots, v_n$ (for some integer $n \geq 1$) in $S - \{v\}$ such that v is expressible as a linear combination of $v_1, \ldots, v_n$. Thus, (1) $\Rightarrow$ (2).
(2) $\Rightarrow$ (3): Given that v in S is expressible as a linear combination of some vectors from the set $S - \{v\}$, this means that there exist vectors $v_1, \ldots, v_n$ (for some integer $n \geq 1$) in $S - \{v\}$, and some scalars $f_1, \ldots, f_n$, such that $v = f_1 v_1 + \cdots + f_n v_n$. We may rewrite this as $f_1 v_1 + \cdots + f_n v_n + (-1)v = 0$. Thus, taking "$m$" to be $n+1$, v_m to be v, f_m to be -1, we find $f_1 v_1 + \cdots + f_n v_n + f_m v_m = 0$. Notice here that $f_m \neq 0$, and that $m \geq 2 > 1$. Thus, we have proved that (2) $\Rightarrow$ (3).
(3) $\Rightarrow$ (1): Given a dependence relation $f_1 v_1 + \cdots + f_n v_n + f_m v_m = 0$ where some f_i is nonzero, assume for convenience that $f_m \neq 0$. If $m = 1$, then the relation $f_1 v_1 = 0$ yields $v_1 = 0$. Clearly, S is redundant, since the 0-vector is not needed for any spanning relation, so $S - \{0\}$ will continue to span V. Now assume that $m > 1$. Then, dividing by f_m and moving v_m to the other side, we find

$$v_m = (-f_1/f_m)v_1 + \cdots + (-f_{m-1}/f_m)v_{m-1}$$

Write v for v_m. We claim now that the set $S - \{v\}$ is already a spanning set for V. For, given a vector w, we know that it is expressible as a linear combination $g_1u_1 + \cdots + g_nu_n$ of some elements $u_1, \ldots, u_n$ from S (here the g_i are scalars). If v is not one of these vectors $u_1, \ldots, u_n$, then $u_1, \ldots, u_n$ are all in $S - \{v\}$, so w is already expressible as a linear combination of vectors from $S - \{v\}$. So assume that v is one of these vectors, say (for simplicity), $v = u_n$. Then, invoking our earlier expression for $v = v_m$, we find

$$\begin{aligned} w &= g_1u_1 + \cdots + g_nu_n \\ &= g_1u_1 + \cdots + g_n\left((-f_1/f_m)v_1 + \cdots + (-f_{m-1}/f_m)v_{m-1}\right) \\ &= g_1u_1 + \cdots + (-g_nf_1/f_m)v_1 + \cdots + (-g_nf_{m-1}/f_m)v_{m-1} \end{aligned}$$

Notice that w is now expressed as a linear combination of the vectors $u_1, \ldots, u_{n-1}, v_1, \ldots, v_{m-1}$—*all of which are in* $S - \{v\}$. It follows therefore that every vector in V is expressible as a linear combination of vectors in $S - \{v\}$. In other words, $S - \{v\}$ already spans V, so there is redundancy in S. Thus, we have proved that (3) $\Rightarrow$ (1).

□

With this lemma in mind, we make the following definition:

Definition 3.22. Let F be a field and V an F–vector space. Let $v_1, \ldots, v_n$ be elements of v. Then $v_1, \ldots, v_n$ are said to be *linearly dependent over F*, or *F–linearly dependent* if there exist scalars $a_1, \ldots, a_n$, not all zero, such that $a_1 \cdot v_1 + \cdots + a_n \cdot v_n = 0$. If no such scalars exist, then $v_1, \ldots, v_n$ are said to be *linearly independent over F*, or *F–linearly independent.* (If there is no ambiguity about the field F, the vectors are merely referred to as *linearly dependent* or *linearly independent.* Also, if $v_1, \ldots, v_n$ are linearly independent, respectively linearly dependent, vectors, then the *set* $\{v_1, \ldots, v_n\}$ is said to be a linearly independent, respectively a linearly dependent, set.) An arbitrary subset S of V is said to be *linearly independent* if every finite subset of S is linearly independent. Similarly, an arbitrary subset S of V is said to be *linearly dependent* if some finite subset of S is linearly dependent.

Thus, the implications $1 \Leftrightarrow 3$ of Lemma 3.21 can be stated in this new language as follows: There is redundancy in S *if and only if* S is linearly dependent.

Before proceeding further, here are a few quick exercises:

Exercise 3.22.1. Show that if v is a *nonzero* vector, then the set $\{v\}$ must be linearly independent. See Property (4) in Remark 3.12.

Exercise 3.22.2. Show that two vectors are linearly dependent if and only if one is a scalar multiple of the other.

Exercise 3.22.3. Are the following subsets of the given vector spaces linearly independent? (Very little computation, if any, is necessary.)

1. In $\mathbb{R}^3$: $\{(1,1,1),(10,20,30),(23,43,63)\}$
2. In $\mathbb{R}^3$: $\{(1,0,0),(2,2,0),(3,3,3)\}$
3. In $\mathbb{R}[x]$: $\{(x+1)^3, x^2+x, x^3+1\}$

Exercise 3.22.4. We know that $\mathbb{C}^2$ is a vector-space over both $\mathbb{C}$ (Example 3.3) and over $\mathbb{R}$ (Example 3.10). Show that $v=(1+\imath, 2\imath)$ and $w=(1,1+\imath)$ are linearly dependent when $\mathbb{C}^2$ is considered as a $\mathbb{C}$ vector space, but linearly independent when considered as a $\mathbb{R}$ vector space.

Also, let us illustrate the meaning of the last two sentences of the Definition 3.22 above. Let us consider the following:

Example 3.23. Consider the subset $S=\{1,x,x^2,x^3,\dots\}$ of $\mathbb{R}[x]$, with $\mathbb{R}[x]$ viewed as a vector space over $\mathbb{R}$ (we have already considered this set in Examples 3.16 and 3.20 above). This is, of course, an *infinite* set. Consider any nonempty finite subset of S, for instance, the subset $\{x,x^5,x^{17}\}$, or the subset $\{1,x,x^2,x^{20}\}$, or the subset $\{1,x^3,x^{99},x^{100},x^{1001},x^{1004}\}$. In general, a nonempty finite subset of S would contain n elements (for some $n\ge 1$), and these elements would be various powers of x—say x^{i_1}, x^{i_2}, …, x^{i_n}. These elements are definitely linearly independent, since if $a_1x^{i_1}+\cdots+a_nx^{i_n}$ is the zero polynomial, then by the definition of the zero polynomial, each a_i must be zero. *This is true regardless of which finite subset of S we take*—all that would be different in different finite subsets is the number of elements (the integer n) and the particular powers of x (the integers i_1 through i_n) chosen. Thus, according to our definition, the set S is linearly independent.

On the other hand, consider the subset $S' = S \cup \{1 + x\}$. Any finite subset of S' that does not contain all three vectors 1, x and $1 + x$ will be linearly independent (check!). However, this alone is not enough for you to conclude that S' is a linearly independent set. For the subset $\{1, x, 1 + x\}$ of T is linearly dependent: $1 \cdot 1 + 1 \cdot x + (-1) \cdot (1 + x) = 0$. By the definition above, T is a linearly *dependent* set.

Remark 3.24. Note that the zero vector is linearly dependent: for example, the nonzero scalar 1 multiplied by 0_V gives 0_V. Thus, if V is the zero vector space, then $\{0_V\}$ is a linearly dependent spanning set, so by Lemma 3.21, this set has to have redundancy. Hence, some subset of $\{0_V\}$ must already span the trivial space. But the only subset of $\{0_V\}$ is the empty set, hence this lemma tells us that the empty set must span $\{0_V\}$. This is indeed consistent with the convention adopted in Remark 3.17 above.

We are now ready to construct the algebraic analog of coordinate axes. We will choose as our candidate any set of vectors that spans our vector space and in which there is no redundancy. Moreover, instead of using the term coordinate axes (which is inspired by the geometric examples of $\mathbb{R}^2$ and $\mathbb{R}^3$), we will coin a new term—the algebraic analog of coordinate axes will be called a *basis* of our vector space. Since redundancy is equivalent to linear dependence (Lemma 3.21), *lack* of redundancy is equivalent to linear *in*dependence. We hence have the following definition:

Definition 3.25. Let F be a field and V an F–vector space. A subset S of V is said to be a *basis* of V if S spans V and there is no redundancy in S. Alternatively, since lack of redundancy is equivalent to linear independence, S is said to be a basis of V if S spans V and is linearly independent. The individual vectors that belong to S are referred to as *basis vectors*. Sometimes, when we wish to emphasize the field of scalars, we refer to S as an F-*basis* of V.

Here are some examples of bases of vector spaces:

Example 3.26. The set consisting of the vectors $\mathbf{i}$ and $\mathbf{j}$ is a basis for $\mathbb{R}^2$. We have already seen in the text that $\mathbf{i}$ and $\mathbf{j}$ span $\mathbb{R}^2$.

Exercise 3.26.1. Argue carefully why there is no redundancy in the set $\{\mathbf{i}, \mathbf{j}\}$. Alternatively, argue why the set $\{\mathbf{i}, \mathbf{j}\}$ is linearly independent.

Exercise 3.26.2. Show that the set consisting of the vectors $\mathbf{i}$ and $\mathbf{w} = (1/\sqrt{2}, 1/\sqrt{2})$ also forms a basis. (We have already done this, in Example 3.19!)

Example 3.27. Recall the definition of the vector space $\mathbb{R}^n$ in Example 3.2. Let e_i stand for the vector whose components are all zero except in the i-th slot, where the component is 1. (For example, in $\mathbb{R}^4$, $e_1 = (1, 0, 0, 0)$, $e_3 = (0, 0, 1, 0)$, etc.). Then the e_i form a basis for $\mathbb{R}^n$ as a $\mathbb{R}$-vector space. They clearly span $\mathbb{R}^n$ since any n-tuple $(r_1, \ldots, r_n) \in \mathbb{R}^n$ is just $r_1e_1 + \cdots + r_ne_n$. As for the linear independence, assume that $r_1e_1 + \cdots + r_ne_n = 0$ for some scalars $r_1, \ldots, r_n$. Since the sum $r_1e_1 + \cdots + r_ne_n$ is just the vector $(r_1, \ldots, r_n)$, we find $(r_1, \ldots, r_n) = (0, \ldots, 0)$, so each r_i must be zero.

This basis is known as the *standard basis* for $\mathbb{R}^n$. Of course, in $\mathbb{R}^2$, e_1 and e_2 are more commonly written as $\mathbf{i}$ and $\mathbf{j}$, and in $\mathbb{R}^3$, e_1, e_2, and e_3 are more commonly written as $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$.

Exercise 3.27.1. Show that the vectors e_1, $e_2 - e_1$, $e_3 - e_2$, $\ldots$, $e_n - e_{n-1}$ also form a basis for $\mathbb{R}^n$.

Example 3.28. The set consisting of the elements 1 and $\sqrt{2}$ forms a basis for $\mathbb{Q}[\sqrt{2}]$ as a vector space over $\mathbb{Q}$. (We have seen in Example 3.15 above that 1 and $\sqrt{2}$ span $\mathbb{Q}[\sqrt{2}]$. As for the $\mathbb{Q}$–linear independence of 1 and $\sqrt{2}$, you were asked to prove this in Exercise 2.12.4 in Chapter 2!)

Example 3.29. The set $\{1, x, x^2, \ldots\}$ forms a basis for $\mathbb{R}[x]$ as a vector space over $\mathbb{R}$. We have seen in Example 3.16 that this set spans $\mathbb{R}[x]$. As for the linear independence, see the argument in Example 3.23 above.

Exercise 3.29.1. Prove that the set $\mathcal{B} = \{1, 1+x, 1+x+x^2, 1+x+x^2+x^3 \dots\}$ is also a basis for $\mathbb{R}[x]$ as a vector space over $\mathbb{R}$. (Hint: Writing $v_0 = 1$, $v_1 = 1+x$, $v_2 = 1+x+x^2$, etc., note that for $i = 1, 2, \dots$, $x^i = v_i - v_{i-1}$. It follows that all powers of x (including x^0) are expressible as linear combinations of the v_i. Why does it follow from this that the v_i span $\mathbb{R}[x]$? As for linear independence, suppose that for some finite collection $v_{i_1}, \dots, v_{i_k}$ (with $i_1 < i_2 < \dots < i_k$), there exist scalars $r_1, \dots, r_k$ such that $r_1 v_{i_1} + \dots + r_k v_{i_k} = 0$. What is the highest power of x in this expression? In how many of the elements $v_{i_1}, \dots, v_{i_k}$ does it show up? What is its coefficient? So?)

Example 3.30. Consider $F_n[x]$ as an F–vector space (see Example 3.9 above). You should easily be able to describe a basis for this space and prove that your candidate is indeed a basis.

Example 3.31. The set $\{1, \sqrt{2}, \sqrt{3}, \sqrt{6}\}$ forms a basis for $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$ as a vector space over $\mathbb{Q}$. You have seen in Example 2.34 that, by our very definition of the ring, every element of $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$ is of the form $a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6}$, where a, b, c, and d are all rational numbers. This simply says that the set $\{1, \sqrt{2}, \sqrt{3}, \sqrt{6}\}$ spans $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$ as a vector space over $\mathbb{Q}$. As for the linear independence of this set, this was precisely the point of Exercise 2.118 in Chapter 2!)

Example 3.32. The n^2 matrices $e_{i,j}$ (see Exercise 2.16.4 of Chapter 2 for this notation) are a basis for $M_n(\mathbb{R})$.

Exercise 3.32.1. Prove this! To start you off, here is a hint: In $M_2(\mathbb{R})$, for example, a matrix such as

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

can be written as the linear combination $e_{1,1} + 2e_{1,2} + 3e_{2,1} + 4e_{2,2}$.

Example 3.33. Certain linear combinations of basis vectors also give us a basis:

Exercise 3.33.1. Going back to $\mathbb{Q}[\sqrt{2}]$, show that the vectors 1 and $1 + \sqrt{2}$ also form a basis. (Hint: Any vector $a + b\sqrt{2}$ can be rewritten as $(a - b) + b(1 + \sqrt{2})$. So?)

Exercise 3.33.2. Now show that if V is any vector space over any field with basis $\{v_1, v_2\}$, then the vectors $v_1, v_1 + v_2$ also form a basis. How would you generalize this pattern to a vector space that has a basis consisting of n elements $\{v_1, v_2, \dots, v_n\}$? Prove that your candidate forms a basis.

Exercise 3.33.3. Let V be a vector space with basis $\{v_1, \dots, v_n\}$. Study Exercise 3.27.1 and come up with a linear combination of the v_i, similar to that exhibited in that exercise, that also forms a basis for V. Prove that your candidate forms a basis.

Example 3.34. Consider the vector space $\prod_0^\infty F$ of Example 3.4 above. You may find it hard to describe explicitly a basis for this space. However, let e_i (for $i = 0, 1, \dots$) be the infinite-tuple with 1 in the position indexed by i and zeros elsewhere. (Thus, $e_0 = (1, 0, 0, \dots)$, $e_1 = (0, 1, 0, \dots)$, etc.)

Exercise 3.34.1. Why is the set $S = \{e_0, e_1, e_2, \dots\}$ *not* a basis for $\prod_0^\infty F$? Is S at least linearly independent? (See the notes on page 154 for some comments on this example.)

Example 3.35. The empty set is a basis for the trivial vector space. This follows from Remark 3.17 (see also Remark 3.24), since the empty set spans the trivial space, and since the empty set is *vacuously* linearly independent.

Here is a result that describes a useful property of bases and is very easy to prove.

Proposition 3.36. *Let V be a vector space over a field F, and let S be a basis. Then in any expression of a vector $v \in V$ as $v = f_1 b_1 + \cdots + f_n b_n$ for suitable vectors $b_i \in S$ and nonzero scalars f_i, the b_i and the f_i are uniquely determined.*

Proof. What we need to show is that if v is expressible as $f_1 b_1 + \cdots + f_n b_n$ for suitable vectors $b_i \in S$ and nonzero scalars f_i, and is also expressible as $g_1 c_1 + \cdots + g_m c_m$ for suitable vectors $c_i \in S$ and nonzero scalars g_i, then $n = m$, and after relabelling if necessary, each $b_i = c_i$ and each $f_i = g_i$ ($i = 1, \dots, n$). To do this, assume, after relabelling if necessary, that $b_1 = c_1$,

$\ldots$, $b_t = c_t$ (for some $t \leq \min(m, n)$), and that the sets $\{b_{t+1}, \ldots, b_n\}$ and $\{c_{t+1}, \ldots, c_m\}$ are disjoint. Then, bringing all terms to one side, we may rewrite our equality as

$$\begin{aligned}(f_1 - g_1)b_1 + \cdots + (f_t - g_t)b_t \quad &+ \quad f_{t+1}b_{t+1} + \cdots + f_n b_n \\ &- \quad g_{t+1}b_{t+1} - \cdots - g_m b_m = 0\end{aligned}$$

By the linear independence of the subset $\{b_1, \ldots, b_t, b_{t+1}, \cdots, b_n, c_{t+1}, \cdots, c_m\}$ of S, we find that $f_1 = g_1$, $\ldots$, $f_t = g_t$, $f_{t+1} = \cdots = f_n = 0$, $g_{t+1} = \cdots = g_m = 0$. But since the scalars were assumed to be nonzero, "$f_{t+1} = 0$" and "$g_{t+1} = 0$" are impossible, so, to begin with, there must have been no "f_{t+1}" or "g_{t+1}" to speak of! Thus, t must have equaled n, and similarly, t must have equaled m. From this, we get $n = m$ $(= t)$, and then, by our very definition of t, we find that $b_1 = c_1$, $\ldots$, $b_n = c_n$. Coupled with our derivation that $f_1 = g_1$, $\ldots$, $f_t = g_t$, we have our desired result. □

Now that we have arrived at the algebraic analog of coordinate axes, we turn our attention to the next step in our program—we need to show that every vector space has a basis, and that different bases of the same vector space have the same number of elements in them.

The first of these two tasks, namely, showing that every vector space has a basis, is a little tricky to do: to do full justice to this task, we need to invoke *Zorn's Lemma*, an extremely useful tool of logic. (Zorn's Lemma, in spite of its name, is really not a lemma, but an *axiom* of logic. See Chapter B in the Appendix.) For a first introduction to abstract algebra, any usage of Zorn's Lemma can seem dense and somewhat foreboding (what else will the Gods of Logic hurl at us?), so we will relegate the full proof to the same Chapter B in the Appendix (see Theorem B.7 there). However, to help build a more concrete feel for the existence of bases, we will also give a proof of the existence of a basis in the special case when we *know* that the vector space in question has a finite spanning set.

We will assume that our vector space is not the trivial space, since we already know that the trivial space has a basis (see Example 3.35 above).

Proposition 3.37. *Let V be a vector space over a field F. Let S be a spanning set for V, and assume that S is a finite set. Then some subset of S is a basis of V. In particular, every vector space with a finite spanning set has a basis.*

Proof. Note that S is nonempty, since V has been assumed to not be the trivial space (see Remark 3.17). If the zero vector appears in S, then the set $S' = S - \{0\}$ that we get by throwing out the zero vector will still span V (why?) and will still be finite. Any subset of S' will also be a subset of S, so if we can show that some subset of S' must be a basis of V, then we would have proved our theorem. Hence, we may assume that we are given a spanning set S for V that is not only finite, but one in which none of the vectors is zero.

Let $S = \{v_1, v_2, \ldots, v_n\}$ for some $n \geq 1$. If there is no redundancy in S, then there is nothing to prove: S would be a basis by the very definition of a basis. So assume that there is redundancy in S. By relabelling if necessary, we may assume that v_n is redundant. Thus, $S_1 = \{v_1, v_2, \ldots, v_{n-1}\}$ is itself a spanning set for V. Once again, if there is no redundancy in S_1, then we would be done; S_1 would be a basis. So assume that there is redundancy in S_1. Repeating the arguments above and shrinking our set further and further, we find that this process must stop somewhere, since at worst, we would shrink our spanning set down to one vector, say $S_{n-1} = \{v_1\}$, and a set containing just one nonzero vector must be linearly independent (Exercise 3.22.1), so S_{n-1} would form a basis. (Note that this is only the worst case; in actuality, this process may stop well before we shrink our spanning set down to just one vector.) When this process stops, we would have a subset of S that would be a basis of V. $\square$

Remark 3.38. Notice that to prove that bases exist (in the special case where V has a finite spanning set) what we really did was to show that every finite spanning set of V can be shrunk down to a basis of V. This result is true more generally: Given *any* spanning set S of a vector space V (in other words, not just a finite spanning set S), there exists a subset S' of S that forms a basis of V. See the notes on page 223 Chapter B in the Appendix.

Having proved that every vector space has a basis, we now need to show that different bases of a vector space have the same number of elements in them. (Remember our original program. We wish to measure the size of a vector space, and based on our examples of $\mathbb{R}^2$ and $\mathbb{R}^3$, we think that a good measure of the size would be the number of coordinate axes, or basis elements, that a vector space has. However, for this to make sense, we need to be guaranteed that every vector space has a basis—we just convinced ourselves of this—and that different bases of a vector space have the same number of elements in them.) In preparation, we will prove an important lemma. Our desired results will fall out as corollaries.

We continue to assume that our vector space is not the trivial space.

Lemma 3.39 (Exchange Lemma)**.** *Let V be a vector space over a field F, and let $B = \{v_1, \ldots, v_n\}$ $(n \geq 1)$ be a spanning set for V. Let $C = \{w_1, \ldots, w_m\}$ be a linearly independent set. Then $m \leq n$.*

Proof. The basic idea behind the proof is to replace vectors in the spanning set B one after another with vectors in C, and observing at the end that if m were greater than n, then there would not be enough replacements of elements of B to guarantee linear independence of the set C.

We begin as follows: Since B spans V, every vector in V is expressible as a linear combination of elements of B. In particular, we may write w_1 as a linear combination of elements of B, that is, $w_1 = c_1v_1 + +c_2v_2 + \cdots + c_nv_n$ for suitable scalars c_i, not all zero. Since one of these scalars is nonzero, we may assume for convenience (by relabelling the vectors of B if necessary), that $c_1 \neq 0$. As usual, we may write $v_1 = (-1/c_1)w_1 + (-c_2/c_1)v_2 + (-c_3/c_1)v_3 +$

$\cdots + (-c_n/c_1)v_n$. Now go back and study how we proved (2) $\Rightarrow$ (3) in Lemma 3.21. We are going to use the same sort of an argument here: we will prove that the set $\{w_1, v_2, v_3, \ldots, v_n\}$ spans V. For given any vector v in V, it can be written as a linear combination $v = f_1v_1 + f_2v_2 + \cdots + f_nv_n$ for suitable scalars f_i (why?). Now, in this expression, substitute $(-1/c_1)w_1 + (-c_2/c_1)v_2 + (-c_3/c_1)v_3 + \cdots + (-c_n/c_1)v_n$ for v_1, and what do you find?—v is expressible as a linear combination of w_1, v_2, v_3, $\ldots$, v_n! Thus, the set $\{w_1, v_2, v_3, \ldots, v_n\}$ spans V as claimed.

Now observe what we have done: we have replaced v_1 with w_1. Let us take this to the next step. Since the set $\{w_1, v_2, v_3, \ldots, v_n\}$ spans V, we can write w_2 as a linear combination of elements of this set. Thus, $w_2 = g_1w_1 + g_2v_2 + g_3v_3 + \cdots + g_nv_n$ for suitable scalars g_i, not all zero. Now the scalars g_2, g_3, $\ldots$, g_n cannot all be zero, since g_1 would then have to be nonzero (why?) and this relation would then read $w_2 = g_1w_1$—a contradiction, as the set C is linearly independent. Hence, one of the scalars g_2, g_3, $\ldots$, g_n must be nonzero. For convenience, we may assume (by relabelling the vectors v_2, v_3, $\ldots$, v_n if necessary) that $g_2 \neq 0$. Dividing by g_2 and moving all terms but v_2 to one side, we can write v_2 as a linear combination of the vectors w_1, w_2, v_3, $\ldots$, v_n. Exactly as in the last paragraph, we find that since the set $\{w_1, v_2, v_3, \ldots, v_n\}$ spans V, the set $\{w_1, w_2, v_3, \ldots, v_n\}$ also spans V.

So far, we have succeeded in replacing v_1 with w_1 and v_2 with w_2, and the resultant set $\{w_1, w_2, v_3, \ldots, v_n\}$ still spans V. Now continue this process, and consider what would happen if we were to assume that m is greater than n. Well, we would replace v_3 by w_3, v_4 by w_4, etc., and then v_n by w_n. (We know that we would be able to replace all the v's with w's because by assumption, there are more w's than v's.) At each stage of the replacement, we would be left with a set that spans V. In particular, the set we would be left with after replacing v_n by w_n, namely $\{w_1, w_2, \ldots, w_n\}$, would span V. But since we assumed that m is greater than n, there would be at least one

"w" left, namely w_{n+1}. Since $\{w_1, w_2, \ldots, w_n\}$ would span V, we would be able to write w_{n+1} as a linear combination of the vectors w_1, w_2, $\ldots$, w_n. This is a contradiction, since the set C is linearly independent! Hence m cannot be greater than n, that is, $m \leq n$! □

We are now ready to prove that different bases of a given vector space have the same number of elements. We will distinguish between two cases: vector spaces having bases with finitely many elements, and those having bases with infinitely many elements. We will take care of the infinite case first.

Corollary 3.40. *If a vector space V has one basis with an infinite number of elements, then every other basis of the vector space also has an infinite number of elements.*

Proof. Let S be the basis of V with an infinite number of elements (that exists by hypothesis), and let T be any other basis. Assume that T has only finitely many elements, say m. Since S has infinitely many elements, we can certainly pick $m+1$ vectors from it. So pick any $m+1$ vectors from S and denote this selected set of vectors by S'. Since the vectors in S' are part of the basis S, they are certainly linearly independent. We may think of the set T as the set "B" of Lemma 3.39 (after all, T being a basis, will span V), and we may think of the set S' as the set "C" of the same lemma (after all, S' is linearly independent). The lemma then shows that $m+1 \leq m$, a clear contradiction. Hence T must also be infinite! □

We settle the finite case now. Recall that we are assuming that our vector space is not the trivial space. The trivial space has only one basis anyway, the empty set (see Remark 3.24).

Corollary 3.41. *If a vector space V has one basis with a finite number of elements n, then every basis of V contains n elements.*

Proof. Let $S = \{x_1, \dots, x_n\}$ be the given basis of V with n elements, and let T be any other basis. If T were infinite, Lemma 3.40 above says that S must also be infinite. Since this is not true, we find that T must have a finite number of elements. So, assume that T has m elements, say $T = \{y_1, \dots, y_m\}$. We wish to show that $m = n$. We may think of S as the set "B" of Lemma 3.39, since it clearly spans V. Also, we may think of the set T as the set "C" of the lemma, since T, being a basis, is certainly linearly independent. Then the lemma says that m must be less than or equal to n. Now let us reverse this situation: let us think of T as the set "B," and let us think of S as the set "C". (Why can we do this?) Then the lemma says that n must be less than or equal to m. Thus, we have $m \leq n$ and $n \leq m$, so we find that $n = m$. □

We are finally ready to make the notion of the size of a vector space precise!

Definition 3.42. A (nontrivial) vector space V over a field F is said to be *finite-dimensional* (or *finite-dimensional over F*) if it has a basis with a finite number of elements in it; otherwise, it is said to be *infinite-dimensional* (or *infinite-dimensional over F*). If V is finite-dimensional, the *dimension* of V is defined to be the number of elements in any basis. If V is infinite-dimensional, the *dimension* of V is defined to be infinite. If V has dimension n, then V is also referred to as an *n-dimensional* space (or as being *n-dimensional over F*); this is often written as $\dim_F(V) = n$.

Remark 3.43. By convention, the dimension of the trivial space is taken to be zero. This is consistent with the fact that it has as basis the empty set, which has zero elements.

Let us consider the dimensions of some of the vector spaces in the examples on page 97 (see also the examples on page 111, where we consider bases of these vector spaces). $\mathbb{R}^2$ and $\mathbb{R}^3$ have dimensions 2 and 3 (respectively) as vector spaces over $\mathbb{R}$.

Question 3.44. What is the dimension of $\mathbb{R}^n$?

$\mathbb{Q}[\sqrt{2}]$ is 2-dimensional over $\mathbb{Q}$. $\mathbb{R}[x]$ is infinite-dimensional over $\mathbb{R}$, while $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$ is 4-dimensional over $\mathbb{Q}$. Similarly, $M_n(\mathbb{R})$ is n^2-dimensional over $\mathbb{R}$.

Question 3.45. What is the dimension of $F_n[x]$ over F? (Warning! It is *not* n.)

Question 3.46. By Example 3.7, $\mathbb{C}$ is a vector space over $\mathbb{C}$, and as well, over $\mathbb{R}$. What is the dimension of $\mathbb{C}$ as a $\mathbb{C}$-vector space? As a $\mathbb{R}$-vector space?

With the definition of dimension under our belt, the following is another corollary to the Exchange Lemma (Lemma 3.39):

Corollary 3.47. *Let V be an n-dimensional vector space. Then every subset S of V consisting of more than n elements is linearly dependent. (Alternatively, if S is a linearly independent subset of V then S has at most n elements.)*

Proof. Assume, to the contrary, that V contains a linearly independent subset S that contains more than n elements. Therefore, we can find $n + 1$ distinct elements $v_1, v_2, \ldots, v_{n+1}$ in S. Write C for the set $\{v_1, v_2, \ldots, v_{n+1}\}$ and let B be any basis. By the very definition of dimension, B must have n elements. Now apply Lemma 3.39 to the sets B and C—we find that $n + 1 \leq n$, which is a contradiction. Hence every subset of V consisting of more than n elements must be linearly dependent, or, what is the same, any linearly independent subset of V must have at most n elements. $\square$

Similarly, with the definition of dimension under our belt, the following is an easy corollary of Proposition 3.37:

Corollary 3.48. *Let V be an n-dimensional vector space. Then any spanning set for V has at least n elements.*

Proof. Let S be a spanning set, and assume that $|S| = t < n$. By Proposition 3.37, some subset of S is a basis of V. Since this subset can have at most t

elements, it follows that the dimension of V, which is the size of this basis, is at most t. This contradicts the fact that the dimension of V is n. $\square$

Putting together Corollary 3.48 and Proposition 3.37, we find that if V is an n-dimensional vector space, then any spanning set for V must have at least n elements, and this set can then be shrunk to a basis of V (consisting of exactly n elements). There is a corresponding result for linearly independent elements in V. Corollary 3.47 shows that any linearly independent subset of V must have at most n elements. What we will see in Proposition 3.49 below is that any linearly independent subset of V can be expanded to a basis of V (which will then have exactly n elements).

Proposition 3.49 below holds even when V is not assumed to be finite-dimensional, but a full proof requires the use of Zorn's Lemma. The proof of the general case is sketched in the remarks on page 224 in Chapter B in the Appendix.

Proposition 3.49. *Let V be a finite-dimensional vector space, and let C be a linearly indepenent set. Then C can be expanded to a basis of V, i.e., there exists a basis B of V such that $C \subseteq B$.*

Proof. Let n be the dimension of V. Then by Corollary 3.47 C has at most n elements in it. Assume that $C = \{v_1, v_2, \ldots, v_t\}$ for some $t \leq n$. If C already spans V, then C would be a basis and we would be done. (And if this happens, you know that t must equal n by Corollary 3.41!) So assume that C does not span V. By the very definition of what it means to span a vector space, there must be a vector in V, call it v_{t+1}, that is not expressible as a linear combination of the elements in C. We claim that the set $C_1 = \{v_1, v_2, \ldots, v_t, v_{t+1}\}$ must be linearly independent. For suppose $f_1v_1 + \cdots + f_tv_t + f_{t+1}v_{t+1} = 0$ for some scalars f_i, not all of which are zero. Then f_{t+1} cannot be zero, since otherwise our relation would read $f_1v_1 + \cdots + f_tv_t = 0$ for nonzero scalars f_i, and this would violate the linear independence of C. Therefore, we may divide our original relation by f_{t+1}

to find $v_{t+1} = (-f_1/f_{t+1})v_1 + \cdots + (-f_t/f_{t+1})v_t$, contradicting the fact that v_{t+1} is not expressible as a linear combination of elements of C. Thus, C_1 is indeed linearly independent as claimed.

Note that the set C_1 has $t+1$ elements. If C_1 spans V, then C_1 would be a basis of V containing C, and we would be done. Otherwise, we could expand C_1 to a linearly independent set C_2 and repeat our arguments ....

Notice that in the process above, we start with our set C with t elements, and at each stage, we come up with a set that has one more element than the set at the previous stage. When we reach a set with exactly n elements, this set *must* span V, for if not, the set we would get at the *next* stage would contain $n+1$ elements and would be linearly independent, contradicting Corollary 3.47 above. This set with exactly n elements would therefore be a basis of V containing C. □

Example 3.50. For example, in $\mathbb{R}^2$, consider the linear independent set $\{\mathbf{i}\}$. The contention of the theorem above is that one can adjoin one other vector to this to get a basis for $\mathbb{R}^2$: for instance the set $\{\mathbf{i}, \mathbf{j}\}$ is a basis, and so, for that matter, is the set $\{\mathbf{i}, \mathbf{w}\}$. (Here, just as earlier in the chapter, $\mathbf{i} = (1,0)$, $\mathbf{j} = (0,1)$, and $\mathbf{w} = (1/\sqrt{2}, 1/\sqrt{2})$.)

We end this section with two more easy results concerning spanning sets and linearly independent sets: the proofs simply consist of combining earlier results!

Proposition 3.51. *Let V be an n-dimensional vector space and S a subset of V. Then:*

1. *If S is a spanning set for V (so $|S| \geq n$ by Corollary 3.48), and if moreover $|S| = n$, then S is a basis for V.*

2. *If S is a linearly independent set (so $|S| \leq n$ by Corollary 3.47), and if moreover $|S| = n$, then S is a basis for V.*

Proof. As promised, the proof simply consists of combining previous results:

1. Given S a spanning set with n elements, Proposition 3.37 shows that some subset S' of S is a basis. Hence, as V is n-dimensional, $|S'| = n$. Since $|S| = n$ as well, we find $S' = S$, i.e., S is already a basis for V.

2. Given S a linearly independent set with n elements, Proposition 3.49 shows that S can be expanded to a basis S'. Hence, as V is n-dimensional, $|S'| = n$. Since $|S| = n$ as well, we find $S' = S$, i.e., S is already a basis for V.

□

Remark 3.52. We have proved quite a few results in this section concerning spanning sets, linearly independent set, and bases. It would be helpful to summarize these results here. In what follows, V is, as usual, a vector space over a field F:

1. A basis for V is a subset of V that *spans* V and in which there is *no redundancy.* Alternatively, a basis is a subset that spans V and is *linearly independent.*

2. Bases always exist.

3. If one basis for V has an infinite number of elements in it, then every other basis for V must also have an infinite number of elements. When this occurs, we say V is *infinite dimensional.*

4. If one basis for V has a finite number of elements "n" in it, then every other basis must also have n elements. When this occurs, we say V if *finite-dimensional* and we define the *dimension* of V to be n.

5. Assume that V is of finite dimension n:

   (a) Any spanning set S for V must contain at least n elements.

   (b) Any spanning set S can be shrunk to a basis for V.

(c) If a spanning set S has exactly n elements, then it is already a basis for V.

(d) Any linearly independent set S must contain at most n elements.

(e) Any linearly independent set S can be expanded to a basis for V.

(f) If a linearly independent set S has exactly n elements in it, then it is already a basis for V.

Of course, the statements in both (5b) and (5e) above hold even when V is infinite-dimensional.

3.3 Subspaces and Quotient Spaces

The idea behind subspaces is very similar to the idea behind subrings, while the idea behind quotient spaces is very similar to the idea behind quotient rings. (There is one key difference: quotient rings are obtained by modding out rings by *ideals,* modding out by subrings will not work. However, quotient spaces can be made by modding out by subspaces. We will see this later in the chapter.)

We will consider subspaces first:

Definition 3.53. Given a vector space V over a field F, a *subspace* of V is a nonempty subset W of V that is closed with respect to vector addition and scalar multiplication, such that with respect to this addition and scalar multiplication, W is itself a vector space (that is, W satisfies all the axioms of a vector space).

Now, we saw in the context of rings (Exercise 2.28 in Chapter 2) that one could have a subset S of a ring R such that S is closed with respect to addition and multiplication, and yet S is *not* a subring of R. It turns out that in the case of vector spaces, it is *enough* for a (nonempty) subset W of a vector space V to be closed with respect to vector addition and scalar multiplication—W will then automatically satisfy all the axioms of a vector space. This is the content of Theorem 3.55 below.

But first, a quick exercise, which is really a special case of Exercise 4.22 in Chapter 4 ahead:

> **Exercise 3.54.** Let W be a subspace of the vector space V. Thus, by definition $(W, +)$ is an abelian group. Let 0_W denote the identity element of this group, and let 0_V denote the usual "0" of V. Show that $0_W = 0_V$. (See also Exercise 2.29 in Chapter 2.)

Theorem 3.55. *Let V be a vector space over a field F, and let W be a nonempty subset of V that is closed with respect to vector addition and scalar multiplication. Then W is a subspace of V.*

Proof. We need to check that all the axioms of a vector space hold. Let us first check that $(W, +)$ is an abelian group. Vector addition in W is both commutative and associative, since for any $v_1, v_2, v_3 \in W$, we may consider v_1, v_2 and v_3 to be elements of V, and in V, the relations $v_1 + v_2 = v_2 + v_1$, and $(v_1 + v_2) + v_3 = v_1 + (v_2 + v_3)$ certainly hold. Next, given any $v \in W$, let us show that $-v$ is also in W. For this we invoke that fact that W is closed with respect to scalar multiplication—since $v \in W$, $-1 \cdot v$ is also in W, and $-1 \cdot v$ is, of course, just $-v$ (see Remark 3.12 above). Now let us show that 0 is in W. Observe that so far, we have not used the hypothesis that W is nonempty. (The proofs that we have given for the fact that addition in W is associative and that every element in W has its additive inverse in W hold *vacuously* even in the case where W is empty. For instance, the chain of arguments $v \in W \Rightarrow -1 \cdot v \in W$ (as W is closed with respect to scalar multiplication) $\Rightarrow -v \in W$ is correct even when there is no vector v in W to begin with!) Now let us use the fact that W is nonempty. Since W is nonempty, it contains at least one vector, call it v. Then, by what we proved above, $-v$ is also in W. Since W is closed under vector addition, $v + (-v)$ is in W, and so 0 is in W. We have thus shown that $(W, +)$ is an abelian group.

It remains to be shown that the four axioms of scalar multiplication also hold for W. But for any r and s in F and v and w in W, we may consider v and w to be elements of V, and as elements of V, we certainly have the relations $r \cdot (v + w) = r \cdot v + r \cdot w$, $(r + s) \cdot v = r \cdot v + s \cdot v$, $(rs) \cdot v = r \cdot (s \cdot v)$, and $1 \cdot v = v$. Hence, the axioms of scalar multiplication hold for W.

This proves that W is a subspace of V. □

We have the following, which captures both closure conditions of the test in Theorem 3.55 above:

Corollary 3.56. *Let V be a vector space over a field F, and let W be a nonempty subset of V that is* closed under linear combinations*, i.e., for all w_1, w_2 in W and all f_1, f_2 in F, the element $f_1w_1 + f_2w_2$ is also in W. Then W is a subspace of V Conversely, if W is a subspace, then W is closed under linear combinations.*

Proof. Assume that W is closed under linear combinations. Taking $f_1 = f_2 = 1$, we find that $w_1 + w_2$ is in W for all w_1, w_2 in W, i.e., W is closed under addition. Taking $f_2 = 0$ we find f_1w_1 is in W for all w_1 in W and all f_1 in F, i.e., W is closed under scalar multiplication. Thus, by Theorem 3.55, W is a subspace. Conversely, if W is a subspace, then for w_1, w_2 in W and all f_1, f_2 in F, f_1w_1 and f_2w_2 are both in W because W is closed under scalar multiplication, and then, $f_1w_1 + f_2w_2$ is in W because W is closed under vector addition. Hence, W is closed under linear combinations. □

Here are some examples of subspaces. In each case, check that the conditions of Theorem 3.55 apply.

Example 3.57. The set consisting of just the element 0 is a subspace.

Question 3.57.1. Why?

We refer to this as the *zero subspace*.

Example 3.58. If you think of $\mathbb{R}^2$ as the vectors lying along the xy plane of 3-dimensional xyz space, then $\mathbb{R}^2$ becomes a subspace of $\mathbb{R}^3$.

Example 3.59. For any nonnegative integers n and m with $n < m$, $F_n[x]$ is a subspace of $F_m[x]$. Also, $F_n[x]$ and $F_m[x]$ are both subspaces of $F[x]$.

Example 3.60. $U_n(\mathbb{R})$ (the set of upper triangular $n \times n$ matrices with entries in $\mathbb{R}$) is a subspace of the $\mathbb{R}$–vector space $M_n(\mathbb{R})$.

Example 3.61. $\mathbb{Q}[\sqrt{2}]$ is a subspace of the $\mathbb{Q}$–vector space $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$. Of course, we know very well by now that since $\mathbb{Q} \subseteq \mathbb{Q}[\sqrt{2}]$, $\mathbb{Q}[\sqrt{2}]$ is directly a $\mathbb{Q}$–vector space. Both $\mathbb{Q}$–vector space structures on $\mathbb{Q}[\sqrt{2}]$ are the same, that is, in both ways of looking at $\mathbb{Q}[\sqrt{2}]$ as a $\mathbb{Q}$–vector space, the rules for vector addition and scalar multplication are the same. In the first way (viewing $\mathbb{Q}[\sqrt{2}]$ as a subspace of $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$), we first think of any element $a+b\sqrt{2}$ of $\mathbb{Q}[\sqrt{2}]$ as the element $a+b\sqrt{2}+0\sqrt{3}+0\sqrt{6}$ of $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$. Doing so, the vector sum of $a+b\sqrt{2}+0\sqrt{3}+0\sqrt{6}$ $(= a+b\sqrt{2})$ and $a'+b'\sqrt{2}+0\sqrt{3}+0\sqrt{6}$ $(= a'+b'\sqrt{2})$ is $(a+a')+(b+b')\sqrt{2}+0\sqrt{3}+0\sqrt{6}$ $(= (a+a')+(b+b')\sqrt{2})$. On the other hand, viewing $\mathbb{Q}[\sqrt{2}]$ directly as a $\mathbb{Q}$–vector space, the vector sum of $a + b\sqrt{2}$ and $a' + b'\sqrt{2}$ is also $(a + a') + (b + b')\sqrt{2}$. In a similar manner, you can see that the rules for scalar multiplication are also identical.

Example 3.62. The example above generalizes as follows: Suppose $F \subseteq K \subseteq L$ are fields. The field extension L/F makes L an F–vector space. Since K is closed with respect to vector addition and scalar multiplication, K becomes a subspace of L. But the field extension K/F exhibits K directly as an F–vector space. The two F–vector space structures on K, one that we get from viewing K as a subspace of the F–vector space L and the other that we get directly from the field extension K/F, are the same.

Example 3.63. In Example 3.4 let $\bigoplus_0^\infty F$ denote the set of all infinite tuples $(a_0, a_1, \dots)$ in which *only finitely many of the a_i are nonzero.* Then $\bigoplus_0^\infty F$ is a subspace of $\prod_0^\infty F$.

> **Exercise 3.63.1.** Prove this!
>
> **Exercise 3.63.2.** Show that the set $S = \{e_0, e_1, e_2, \dots\}$ *is* a basis for $\bigoplus_0^\infty F$? (Contrast this with Exercise 3.34.1 above.)

This example is known as the *direct sum* of (countably infinite) copies of F.

Example 3.64. For any field F, $F[x^2]$ (that is, the set of all polynomials of the form $\sum_{i=0}^{n} f_i x^{2i}$, $n \geq 0$) is a subspace of $F[x]$.

> **Question 3.64.1.** What is the dimension of this subspace? Can you discover a basis for this subspace?

Example 3.65. Let V be a vector space over a field F, and let S be any nonempty subset of V.

> **Definition 3.65.1.** The *linear span of* S is defined as the set of all linear combinations of elements of S, that is, the set of all vectors in V that can be written as $c_1 s_1 + c_2 s_2 + \cdots + c_k s_k$ for some integer $k \geq 1$, some scalars c_i, and some vectors $s_i \in S$.

> **Exercise 3.65.1.** Show that the linear span of S is a subspace of V.

For instance, in $\mathbb{R}^3$, if we take $S = \{\mathbf{i}, \mathbf{j}\}$, then the linear span of S is the set of all vectors in $\mathbb{R}^3$ that are of the form $a\mathbf{i} + b\mathbf{j}$ for suitable scalars a and b, in other words, the xy-plane. As we saw in Example 3.58 above, the xy-plane is a subspace of $\mathbb{R}^3$!

You should be able to do the following:

> **Question 3.66.** Which of the following are subspaces of $\mathbb{R}^3$?
>
> 1. $\{(a, b, c) \mid a + 3b = c\}$
> 2. $\{(a, b, c) \mid a = b^2\}$
> 3. $\{(a, b, c) \mid ab = 0\}$

We turn our attention now to quotient spaces. Recall how we constructed the quotient ring R/I given a ring R and an ideal I: we first defined an equivalence relation on R by $a \sim b$ if and only if $a - b \in I$ (see page 57 in Chapter 2). We found that the equivalence class of an element a is precisely the coset $a + I$ (Lemma 2.78 in that chapter). We then defined the ring R/I

to be the set of equivalence class of R under the naturally induced definitions $[a]+[b]=[a+b]$ and $[a][b]=[ab]$ (see Definition 2.79 in that chapter). Of course, we had to check that our operations were well-defined and that we indeed obtained a ring by this process (see Lemma 2.80 and Theorem 2.82 in that chapter). We will follow the same approach here.

So, given a vector space V over a field F, and a subspace W, we define an equivalence relation on W by $v \sim w$ if and only if $v - w \in W$. Exactly as on page 57, we can see that this is indeed an equivalence relation. We define the coset $a+W$ to be the set of all elements of the vector space of the form $a+w$ as w varies in W, and we call this the *coset of W with respect to a* We have the following, whose proof is exactly as in Lemma 2.78 of Chapter 2 and is therefore omitted:

Lemma 3.67. *The equivalence class $[a]$ is precisely the coset $a+W$.*

As with quotient rings, we will denote the set of equivalence classes of V by V/W, whose members we will denote as both $[a]$ and $a+W$. We define an addition operation on V/W and a scalar multiplication $F \times V/W \to V/W$ by the following:

Definition 3.68. $[u]+[v]=[u+v]$ and $f \cdot [u] = [f \cdot u]$ for all $[u]$ and $[v]$ in V/W and all f in F. (In coset notation, this would read $(u+W)+(v+W) = (u+v)+W$, and $f(u+W) = fu+W$.) As always, if the context is clear, we will often omit the "$\cdot$" sign and write $r[b]$ for $r \cdot [b]$.

The following should now be easy, after your experience with quotient rings (see Lemma 2.80 in Chapter 2):

> **Exercise 3.69.** Show that the operations of addition and scalar multiplication on V/W described above in Definition 3.68 are well-defined. Show that the addition operation is commutative.

We now have the following:

Theorem 3.70. *$(V/W, +, \cdot)$ is a vector space over F.*

Proof. As in Theorem 2.82 of Chapter 2, the proof involves checking that all the vector space axioms of Definition 3.1 hold. The proof that $(V/W, +)$ is an abelian group is in fact identical to the proof that $(R/I, +)$ is an abelian group, and we will not do it here (see the remarks on page 154 on where the similarity comes from). As for the axioms for scalar multiplication, let us go through them one-by-one:

1. For all $r \in F$ and $[v],\ [w] \in V/W$, we have $r([v] + [w]) = r[v + w] = [r(v+w)] = [rv+rw]$, where the first and second equalities are because of the way operations are defined on V/W and the last equality is because $r(v + w) = rv + rw$ is a property that holds in the original vector space V. On the other hand, $r[v]+r[w] = [rv]+[rw] = [rv+rw]$, where the equalities are because of the way operations are defined on V/W. Thus, both sides equal $[rv + rw]$, so indeed $r([v] + [w]) = r[v] + r[w]$.

2. For all $r,\ s \in F$ and $[v] \in V/W$, $(r + s)[v] = [(r + s)v] = [rv + sv]$, where the last equality is because of properties of the original vector space V. On the other hand, $r[v] + s[v] = [rv] + [sv] = [rv + sv]$. It follows that $(r + s)[v] = r[v] + s[v]$.

3. For all $r,\ s \in F$ and $[v] \in V/W$, $(rs)[v] = [(rs)v] = [r(sv)]$, where the last equality is because of properties of the original vector space V, while $r(s[v]) = r[sv] = [r(sv)]$. It follows that $(rs)[v] = r(s[v])$.

4. For all $[v] \in V/W$, $1[v] = [1\cdot v] = [v]$, where the last equality is because $1 \cdot v = v$ holds in V

$\square$

Definition 3.71. $(V/W, +, \cdot)$ is called the *quotient space* of V by the subspace W.

As with the case of quotient rings, the intuition behind V/W is that it is a space formed by setting all elements of W to zero. More colloquially, the construction "kills" all elements in W, or "divides out" all elements in W. This last description explains the term "quotient space," and pushing the analogy one step further, V/W can then be thought of as the set of all "remainders" after dividing out by W, endowed with the natural "quotient" binary operation and scalar multiplication of Definition 3.71.

For example, take $V = \mathbb{R}^3$ and W to be the subspace consisting of all vectors lying on the xy plane (Example 3.58 above). What sense do we make of V/W? Every vector v in $\mathbb{R}^3$ can be written as $a\mathbf{i} + b\mathbf{j} + c\mathbf{k}$ for unique real numbers a, b, and c (see Example 3.27 above). Notice that both $a\mathbf{i}$ and $b\mathbf{j}$ are in W. If we "set these to zero" we are left simply with $c\mathbf{k}$ which is a vector lying on the z-axis. Moreover every vector $c\mathbf{k}$ lying on the z-axis arises this way (why?) so we find that V/W is precisely the z-axis. As in the case of rings, this is more than just an equality of sets: this identification of V/W with the z-axis preserves the vector space structure as well, which we will make more precise in the next section.

The following lemma will be useful ahead. We will state the result only for finite-dimensional vector spaces, although, the result (suitably phrased) is true for infinite-dimensional spaces as well (see Exercise 3.109):

Lemma 3.72. *Let V be a finite-dimensional vector space over a field F and let W be a subspace. Let $\{b_1, \ldots, b_m\}$ be a basis for W. Expand this to a basis $\{b_1, \ldots, b_m, b_{m+1}, \ldots, b_n\}$ of V (see Theorem 3.49). Then the set (of equivalence classes of vectors) $\{b_{m+1} + W, \ldots, b_n + W\}$ is a basis for the quotient space V/W.*

Proof. Given any $v + W \in V/W$, we may write $v = r_1 b_1 + \cdots + r_m b_m + r_{m+1} b_{m+1} + \cdots + r_n b_n$ for suitable scalars r_1, ..., r_n. Since the vectors b_1, ..., b_m are in W, so is the vector $r_1 b_1 + \cdots + r_m b_m$. Thus, $v - (r_{m+1} b_{m+1} + \cdots + r_n b_n) \in W$. But this just says that $v + W = (r_{m+1} b_{m+1} + \cdots + r_n b_n) + W$. Recalling how vector addition and scalar multiplication are defined in V/W, we find $v + W = (r_{m+1} b_{m+1} + \cdots + r_n b_n) + W = r_{m+1}(b_{m+1} + W) + \cdots + r_n(b_n + W)$. This shows that the set $\{b_{m+1} + W, \ldots, b_n + W\}$ spans V/W.

As for the linear independence, assume that $r_{m+1}(b_{m+1} + W) + \cdots + r_n(b_n + W) = 0_{V/W}$ for some scalars $r_{m+1}, \ldots, r_n$. Since $0_{V/W}$ is the class of W, we find $r_{m+1}b_{m+1} + \cdots + r_nb_n = w$ for some $w \in W$. But the set $\{b_1, \ldots, b_m\}$ is a basis for W, so we may write $w = r_1b_1 + \cdots + r_mb_m$ for suitable scalars $r_1, \ldots, r_m$. Putting this together, we find $r_1b_1 + \cdots + r_mb_m + (-r_{m+1})b_{m+1} + \cdots + (-r_n)b_n = 0$. Since the set $\{b_1, \ldots, b_m, b_{m+1}, \ldots, b_n\}$ is a basis of V, each r_i $(i = 1, \ldots, n)$ must be zero. In particular, $r_{m+1}, \ldots, r_n$ must all be zero, proving the linear independence of $\{b_{m+1}+W, \ldots, b_n+W\}$. □

We get an easy corollary from this:

Corollary 3.73. *Let V be a finite-dimensional vector space over a field F and let W be a subspace. Then $dim(V) = dim(W) + dim(V/W)$.*

Proof. This is clear from the statement of the lemma above. □

3.4 Vector Space Homomorphisms: Linear Transformations

The ideas in this section parallel the development of ring homomorphisms in Chapter 2. As in the passage from R to R/I, we notice some preservation of structure when passing from V to V/W: the operations in V/W are "essentially the same" as the operations in V except that the elements of V have all been divided out by W. What this means is analogous to the situation with R and R/I: let us denote by f the function $f : V \to V/W$ that "pushes" $u \in V$ "down" to $u+W$. Since $u+W = f(u)$, $v+W = f(v)$, and $(u+v)+W = f(u+v)$, we find $f(u)+f(v) = f(u+v)$. The function f that sends u to $u+W$, along with the property $f(u) + f(v) = f(u + v)$ for all u and v in V, precisely captures the notion that addition in V/W and V are "essentially the same."

Similarly, the definition of scalar multiplication in V/W: $r(u + W) = ru + W$ (here r is in F) gives the feeling that scalar multiplication in V/W is "the same" as the scalar multiplication in V except for dividing out by W: once again this intuition is captured by the function f above along with the property $rf(u) = f(ru)$ for all $r \in F$ and u in V.

Just as with rings we will turn this situation around. Suppose one has a function f from one vector space V over F to another vector space X over F (note that the set of scalars F is the same for both spaces) which has the two properties described above, then one similarly gets the sense that the vector space operations in the two F vector spaces V and X are "essentially the same" except perhaps for dividing out by some subspace. In analogy with rings, we should call this a vector space homomorphism, but traditionally, such a function has been called a linear transformation:

Definition 3.74. Let V and X be two vector spaces over a field F, and let $f\colon V \to X$ be a function. Suppose that f has the following properties:

1. $f(u) + f(v) = f(u + v)$ for all u, v, in V,
2. $rf(u) = f(ru)$ for all r in F and u in V.

Then f is said to be a *linear tranformation* from V to X.

Remark 3.75. As with ring homomorphisms, there are some features of this definition that are worth noting:

1. In the equation $f(u)+f(v) = f(u+v)$, note that the operation on the left side represents vector addition in the vector space X, while the operation on the right side represents addition in the vector space V.

2. Similarly for the equation $rf(u) = f(ru)$: the operation on the left side represents scalar multiplication in X, while the operation on the right side represents scalar multiplication in V.

3. By the very definition of a function, f is defined on all of V, however, the image of V under f need not be all of X i.e, f need not be

surjective (see Example 3.83 or Example 3.84 for instance, although, such examples are really very easy to write down). However, the image of V under f is not an arbitrary subset of X, the definition of a linear tranformation ensures that the image of V under f is actually a *subspace* of X (see Lemma 3.88 later in this section).

4. Note that it is not necessary to stipulate that $f(0_V) = 0_X$ since the property holds automatically, see Lemma 3.77 below.

5. The condition (1) of the definition simply says that f should be a *group* homomorphism from the group $(V, +)$ to the group $(X, +)$ (see Definition 4.57 in Chapter 4 ahead), while the second condition (2) says that the group homomorphism should, in addition, be F-linear.

The following lemma combines the two conditions in the definition of a linear transformation into one:

Lemma 3.76. *Let V and X be two F-vector spaces, and let $f : V \to X$ be a function that satisfies the property that $f(r_1v_1 + r_2v_2) = r_1f(v_1) + r_2f(v_2)$ for all v_1, v_2 in V and all r_1, r_2 in F. Then f is a linear transformation. Conversely, if f is a linear transformation, then $f(r_1v_1 + r_2v_2) = r_1f(v_1) + r_2f(v_2)$ for all v_1, v_2 in V and all r_1, r_2 in F.*

Proof. Assume that f satisfies the property that $f(r_1v_1 + r_2v_2) = r_1f(v_1) + r_2f(v_2)$ for all v_1, v_2 in V and all r_1, r_2 in F. Taking $r_1 = r_2 = 1$, we see that $f(v_1 + v_2) = f(v_1) + f(v_2)$, and taking $r_2 = 0$, we see that $f(r_1v_1) = r_1f(v_1)$. Thus, f is a linear transformation. As for the converse, if f is a linear transformation, then for all v_1, v_2 in V and all r_1, r_2 in F, $f(r_1v_1 + r_2v_2) = f(r_1v_1) + f(r_2v_2) = r_1f(v_1) + r_2f(v_2)$, as desired.

□

The following lemma is analogous to Lemma 2.90 in Chapter 2:

Lemma 3.77. *Let V and X be two F-vector spaces, and let $f : V \to X$ be a linear tranformation. Then $f(0_V) = 0_X$.*

Proof. This proof is identical to the proof of the corresponding Lemma 2.90 in Chapter 2, (since, ultimately, these are both proofs that a *group homomorphism* from a group G to a group H maps the identity in G to the identity in H—see Lemma 4.59 in Chapter 4 ahead). We start with the fact that $f(0_V) = f(0_V + 0_V) = f(0_V) + f(0_V)$. We now have an equality in X: $f(0_V) = f(0_V) + f(0_V)$. Since $(X, +)$ is an abelian group, every element of X has an additive inverse, so there is an element, denoted $-f(0_V)$ with the property that $f(0_V) + (-f(0_V)) = (-f(0_V)) + f(0_V) = 0_X$. Adding $-f(0_V)$ to both sides of $f(0_V) = f(0_V) + f(0_V)$, we get $-f(0_V) + f(0_V) = -f(0_V) + (f(0_V) + f(0_V))$. The left side is just 0_X, while by associativity, the right side is $(-f(0_V) + f(0_V)) + f(0_V) = 0_X + f(0_V)$. But by the definition of 0_X, $0_X + f(0_V)$ is just $f(0_V)$. We thus find $0_X = f(0_V)$, thereby proving the lemma. □

Remark 3.78. Here is another way to prove the statement of the lemma above: Pick any $v \in V$. Then, $0_V = 0_F v$, so $f(0_V) = f(0_F v) = 0_F f(v) = 0_X$. (Here, the first equality is due to Remark 3.12.2, and the last but one equality is because $f(rv) = rf(v)$ for any scalar r since f is a linear transformation.)

Before proceeding to examples of linear transformations, let us consider one remaining object, analogous to the kernel of a ring homomorphism. The concept of a linear transformation was introduced to capture the notion of operations on two F-vector spaces being "the same" except for dividing out by some subspace. Just as with ring homomorphisms, the natural candidate for this subspace is the following:

Definition 3.79. Given a linear transformation $f : V \to X$ between two F-vector spaces, the *kernel* of f is the set $\{u \in V \mid f(u) = 0_X\}$. It is denoted $ker(f)$.

As in the case of kernels of ring homomorphisms, the following statement should come as no surprise:

Proposition 3.80. *Let V and X be vector spaces over a field F. The kernel of a linear tranformation $f : V \to X$ is a subspace of V.*

Proof. By Corollary 3.56, it is sufficient to check that $ker(f)$ is a nonempty subset of V that is closed under linear combinations. Since $0_V \in ker(f)$ (Lemma 3.77), $ker(f)$ is nonempty. Now, for any w_1, w_2 in $ker(f)$ and any r_1, r_2 in F, we find $f(r_1w_1+r_2w_2) = r_1f(w_1)+r_2f(w_2) = r_1 \cdot 0_X + r_2 \cdot 0_X = 0_X$. Hence $r_1w_1+r_2w_2$ is indeed in the kernel of f, so $ker(f)$ is closed under linear combinations.

□

Remark 3.81. As in the case of ring homomorphisms, for any linear transformation $f : V \to X$ between two F-vector spaces, we will have $f(-v) = -f(v)$. One proof is exactly the same as in Remark 2.91 in Chapter 2, and this is not surprising: this is really a proof that in any *group homomorphism* f from a group G to a group H, $f(g^{-1})$ will equal $(f(g))^{-1}$ for all $g \in G$ (see Corollary 4.60 in Chapter 4). Another proof, of course, is to invoke scalar multiplication and Remark 3.12.3: $f(-v) = f(-1 \cdot v) = -1f(v) = -f(v)$.

We are now ready to study examples of linear transformations. The first example is really the *master-example:* it provides an algorithm for constructing linear transformations and leads to matrix representations of linear transformations that are useful for computations:

Example 3.82. *Master-Example of Linear Transformation:* Let V be an F-vector space that is (for simplicity) finite-dimensional, and let $\{b_1, \dots, b_n\}$ be a basis for V. Let X be an F-vector space, and let $w_1, \dots, w_n$ be *arbitrary* vectors in X. Then we have the following:

Lemma 3.82.1. *The function $f : V \to X$ that sends each basis element b_i to the vector w_i ($i = 1, \dots, n$) and a general linear combination $r_1b_1 + \cdots + r_nb_n$ ($r_i \in F$) to the vector $r_1w_1 + \cdots + r_nw_n$ is a (well-defined)* *linear transformation. Conversely, any linear transformation* $f : V \to X$ *is determined fully by where* f *sends each basis vector* b_i *to: if* $f(b_i) = w_i$*, then* f *must be defined on all of* V *by the formula* $f(r_1b_1 + \cdots + r_nb_n) = r_1w_1 + \cdots + r_nw_n$.

Proof. That f is well-defined comes from the fact that the b_i form a *basis* for V, so each element $u \in V$ is expressible as $r_1b_1 + \cdots + rf_nb_n$ for a *unique* choice of scalars r_i. Hence, defining what f does to the element u in terms of the scalars r_i poses no problem as the r_i are uniquely determined by u.

It is now trivial to check that f is a linear transformation: Given $u = r_1b_1+\cdots+r_nb_n$ and $v = s_1b_1+\cdots+s_nb_n$ (here, the r_i and the s_j are scalars), we find $u + v = (r_1 + s_1)b_1 + \cdots + (r_n + s_n)b_n$, so $f(u + v) = (r_1 + s_1)w_1 + \cdots + (r_n + s_n)w_n = (r_1w_1 + \cdots + r_nw_n) + (s_1w_1 + \cdots + s_nw_n) = f(u) + f(v)$.

Similarly, given any scalar $r \in F$, $rv = r(r_1b_1 + \cdots + r_nb_n) = (rr_1)b_1 + \cdots + (rr_n)b_n$, so $f(rv) = (rr_1)w_1 + \cdots + (rr_n)w_n = r(r_1w_1 + \cdots + r_nw_n) = rf(v)$.

Exercise 3.82.1. Which vector space axioms were used in the two chains of equalities in the proof above that $f(u+v) = f(u)+f(v)$ and $f(rv) = rf(v)$?

Exercise 3.82.2. Would the proof be any more complicated if V were not assumed to be finite-dimensional? (Work it out!)

Conversely, if f is any linear transformation from V to X and if $f(b_i) = w_i$ $(i = 1, \ldots, n)$, then, since f is a linear transformation, $f(r_1b_1 + \cdots + r_nb_n) = r_1f(b_1) + \cdots + r_nf(b_n) = r_1w_1 + \cdots + r_nb_n$. Since any vector in V is a linear combination of the vectors $b_1, \ldots, b_n$, this formula completely determines what f sends each vector in V to. $\square$

Now let us carry this one step further. Let $f : V \to X$ be a linear transformation, and suppose (for simplicity) that X is also finite-dimensional, with some basis $\{c_1, \ldots, c_m\}$. Thus, every vector $w \in X$ can be uniquely expressed as $s_1c_1 + \cdots + s_mc_m$ for suitably scalars c_i. In particular, each of the vectors w_i $(= f(b_i))$ therefore can be expressed as a linear combination of the c_j as follows:

$$\begin{aligned} w_1 &= p_{1,1}c_1 + \cdots + p_{1,m}c_m \\ w_2 &= p_{2,1}c_1 + \cdots + p_{2,m}c_m \\ \vdots &= \vdots \\ w_n &= p_{n,1}c_1 + \cdots + p_{n,m}c_m \end{aligned}$$

(The $p_{i,j}$ are scalars. Note how they are indexed: $p_{i,j}$ stands for the coefficient of c_j in the expression of w_i as a linear combination of the various c's. Thus, across each row of this equation, it is the *second* index in $p_{i,j}$ that varies.) Now consider an arbitrary $u \in V$, expressed as a linear combination $u = r_1 b_1 + \cdots + r_n b_n$ for suitable scalars r_i. Then

$$\begin{aligned} f(u) &= r_1 w_1 + \cdots + r_n w_n \\ &= r_1\,(p_{1,1}c_1 + \cdots + p_{1,m}c_m) \\ &+ r_2\,(p_{2,1}c_1 + \cdots + p_{2,m}c_m) \\ &\vdots\ \ \vdots \\ &+ r_n\,(p_{n,1}c_1 + \cdots + p_{n,m}c_m) \end{aligned}$$

Now let us regroup the right side so that all the scalars that are attached to the basis vector c_1 are together, all scalars attached to the basis vector c_2 are together, etc. Doing so, we find

$$\begin{aligned} f(u) &= (p_{1,1}r_1 + p_{2,1}r_2 + \cdots + p_{n,1}r_n)\,c_1 \\ &= (p_{1,2}r_1 + p_{2,2}r_2 + \cdots + p_{n,2}r_n)\,c_2 \\ &= \vdots \\ &= (p_{1,m}r_1 + p_{2,m}r_2 + \cdots + p_{n,m}r_n)\,c_m \end{aligned}$$

(Study this relation carefully: note how the indices of the $p_{i,j}$ behave: $p_{i,j}$ multiplies r_i and is attached to c_j. Notice that across each row of this equation, it is the *first* index of $p_{i,j}$ that changes: this is in contrast to the behavior of the indices in the previous equations. There, it was the second index of $p_{i,j}$ that changed in each row.)

Now suppose that we adopt the convention that we will write any vector $u \in V$, $u = r_1 b_1 + \cdots + r_n b_n$ as the column vector

$$u = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix}$$

and any vector $w \in X$, $w = s_1 c_1 + \cdots + s_m c_m$ as the column vector

$$w = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_m \end{pmatrix}$$

Let us rewrite our equation for $f(u)$ above in the form $f(u) = s_1 c_1 + \cdots + s_m c_m$ for suitable scalars s_i. Since the coefficient of c_1 in $f(u)$ is $p_{1,1} r_1 + p_{2,1} r_2 + \cdots + p_{n,1} r_n$ (see the equation above), we find $s_1 = p_{1,1} r_1 + p_{2,1} r_2 + \cdots + p_{n,1} r_n$. Similarly, since the coefficient of c_2 in $f(u)$ is $p_{1,2} r_1 + p_{2,2} r_2 + \cdots + p_{n,2} r_n$, we find $s_2 = p_{1,2} r_1 + p_{2,2} r_2 + \cdots + p_{n,2} r_n$. Proceeding thus, we find that the vectors u and $f(u)$ are related by the matrix equation

$$\begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_m \end{pmatrix} = \begin{pmatrix} p_{1,1} & p_{2,1} & \cdots & p_{n,1} \\ p_{1,2} & p_{2,2} & \cdots & p_{n,2} \\ \vdots & \vdots & \ldots & \vdots \\ p_{1,m} & p_{2,m} & \cdots & p_{n,m} \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix} \tag{3.1}$$

There is an easy way to remember how this matrix is constructed: notice that the first column of the matrix is precisely the vector w_1 written in terms of its coefficients in the basis $\{c_1, \ldots, c_m\}$, the second column is precisely the vector w_2 written in terms of its coefficients in the basis $\{c_1, \ldots, c_m\}$, and so on, until the last column is precisely the vector w_n written in terms of its coefficients in the basis $\{c_1, \ldots, c_m\}$.

Notice something else about this matrix: it depends *vitally* on the choice of the basis $\{b_1, \ldots, b_n\}$ for V and on the choice of the basis $\{c_1, \ldots, c_m\}$ of X. For, our entire derivation of Equation 3.1 depended on our writing u as a linear combination of the vectors $\{b_1, \ldots, b_n\}$ and the v_i as a linear combination of the vectors $\{c_1, \ldots, c_m\}$. A different choice of basis for V or a different choice of basis for X would have led to different linear combinations, hence to a different matrix in Equation 3.1 above.

We refer to this matrix as the *matrix of the linear transformation f in the bases $\{b_1, \ldots, b_n\}$ for V and $\{c_1, \ldots, c_m\}$ for X*. (If $V = X$ and we use the same basis to describe both vectors and their images under f, then we simply refer to this matrix as *the matrix of f in the basis $\{b_1, \ldots, b_n\}$*.) Since each linear transformation is uniquely determined by the $w_i = f(b_i)$, and since the w_i can be written uniquely as a linear combination of the basis vectors $\{c_1, \ldots, c_m\}$, and since these unique coefficients of the c_j then become the i-th column in the matrix, we find that each linear transformation uniquely determines an $m \times n$ matrix with coefficients in F, by this procedure.

But more is true. Since an arbitrary $m \times n$ matrix with coefficients in F determines a collection of vectors $w_1, \ldots, w_n$ from X (with the i-th column representing w_i), and since a linear transformation f can be constructed from these vectors $w_1, \ldots, w_n$ by defining $f(r_1 b_1 + \cdots + r_n b_n) = r_1 w_1 + \cdots + r_n w_n$, we find that an arbitrary $m \times n$ matrix with coefficients in F leads to a linear transformation $f : V \to X$. Thus, the set of linear transformations $f : V \to X$ is in one-to-one correspondence with $m \times n$ matrices with coefficients in F.

This is all very pretty!

Question 3.82.1. For practice with a concrete example, think about the following:

1. If you are given that a linear transformation $f : \mathbb{R}^2 \to \mathbb{R}^3$ sends the vector $\mathbf{i} = (1, 0)$ to the vector $(1, 2, 0)$ and the vector $\mathbf{j} = (0, 1)$ to the vector $(2, 1, 3)$, what does f do to an arbitrary vector (a, b) in $\mathbb{R}^2$?

2. What is the matrix of f with respect to the basis $\{\mathbf{i}, \mathbf{j}\}$ of $\mathbb{R}^2$ and the basis $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$ of $\mathbb{R}^3$?

3. What is the matrix of f with respect to the basis $\{\mathbf{i}, \mathbf{w} = (1/\sqrt{2}, 1/\sqrt{2})\}$ of $\mathbb{R}^2$ (see Example 3.26) and the basis $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$ of $\mathbb{R}^3$? (Hint: What does f do to $\mathbf{w}$?)

Question 3.82.2. What are the coordinates, in the standard basis for $\mathbb{R}^3$ (see Example 3.27), of the vector $x\mathbf{i} + y\mathbf{j}$, after it undergoes the linear transformation $f : \mathbb{R}^2 \to \mathbb{R}^3$ given by the matrix

$$\begin{pmatrix} a & b \\ c & d \\ e & f \end{pmatrix}$$

where the matrix is written with respect to the basis $\{\mathbf{i}, \mathbf{w} = (1/\sqrt{2}, 1/\sqrt{2})\}$ of $\mathbb{R}^2$ and the basis $\{(1, 0, 0), (-1, 1, 0), (0, -1, 1)\}$ of $\mathbb{R}^3$? (See Exercise 3.27.1 for why $\{(1, 0, 0), (-1, 1, 0), (0, -1, 1)\}$ is a basis of $\mathbb{R}^3$.)

Question 3.82.3. How will the treatment in this example change if either V or X (or both) were to be infinite-dimensional F-vector spaces? (See the remarks on page 155 in the notes for some hints.)

Example 3.83. Let V be an F-vector space. The map $f : V \to V$ that sends any $v \in V$ to 0 is a linear transformation.

Question 3.83.1. If V is n-dimensional with basis $\{b_1, \ldots, b_n\}$, what is the matrix of f with respect to this basis?

Example 3.84. Let V be an F-vector space, and let W be a subspace. The map $f : W \to V$ defined by $f(w) = w$ is a linear transformation.

Question 3.84.1. Assume that W is m-dimensional and V is n-dimensional. Pick a basis $B = \{b_1, \ldots, b_m\}$ of W and expand to a basis $C = \{b_1, \ldots, b_m, b_{m+1}, \ldots, b_n\}$ of V. What is the matrix of f with respect to the basis B of W and the basis C of V?

Example 3.85. Let F be a field, and view $M_n(F)$ as a vector space over F (see Example 3.5). Now view F as an F-vector space (see Example 3.7: note that F is trivially an extension field of F). Then the function $f : M_n(F) \to F$ that sends a matrix to its trace is a linear tranformation. (Recall that the trace of a matrix is the sum of its diagonal entries.)

To prove this, note that this is really a function that sends basis vectors of the form $e_{i,i}$ to 1 and $e_{i,j}$ $(i \neq 0, j \neq 0)$ to 0, and an arbitrary matrix $\sum_{i,j} m_{i,j}e_{i,j}$ to $m_{1,1} \cdot 1 + \cdots + m_{n,n} \cdot 1$. Now apply Lemma 3.82.1 to conclude that f must be a linear transformation.

See Exercise 3.100 at the end of the chapter.

Example 3.86. Let V be a vector space over a field F and let W be a subspace. Assume that V is finite-dimensional (for simplicity). Let $\dim_F(V) = n$, and $\dim_F(W) = m$. Let $\{b_1, \ldots, b_m\}$ be a basis for W, and let us expand this to a basis $\{b_1, \ldots, b_m, b_{m+1}, \ldots, b_n\}$ (see Theorem 3.49). Given any $v \in V$, we may therefore write $v = f_1b_1 + \cdots + f_mb_m + f_{m+1}b_{m+1} + \cdots f_nb_n\}$ for unique scalars $f_i \in F$.

Exercise 3.86.1. Show that the function $\pi : V \to W$ that sends any v expressed as above to the vector $f_1b_1 + \cdots + f_mb_m$ in W is a linear transformation from V to W.

Exercise 3.86.2. Is π surjective? Describe a basis for $ker(\pi)$.

Exercise 3.86.3. The basis $\{b_1, \ldots, b_m\}$ of W can be expanded to a basis $\{b_1, \ldots, b_m, b_{m+1}, \ldots, b_n\}$ of V in many different ways (see Example 3.50). The definition of π above depends on which choice of $\{b_{m+1}, \ldots, b_n\}$ we make. For example, take $V = \mathbb{R}^2$ and W the subspace represented by the x-axis. Take the vector $b_1 = \mathbf{i}$ $(= (1, 0))$ as a basis for W. Show that the definition of π depends crucially on the choice of vector b_2 used to expand $\{b_1\}$ to a basis for $\mathbb{R}^2$ as follows: Select b_2 in two different ways and show that for suitable $v \in \mathbb{R}^2$, $\pi(v)$ defined with one choice of b_2 will be different from $\pi(v)$ defined by the other choice of b_2.

We now come to isomorphisms between vector spaces. In analogy with ring isomorphisms, vector space isomorphisms capture the notion that the vector space structures in two spaces are "essentially the same" *without even having to divide out by any subspace.* As with rings, we need a couple of lemmas first:

Lemma 3.87. *Let V and X be two vector spaces over a field F and let $f : V \to X$ be a linear transformation. Then f is an injective function if and only if $ker(f)$ is the zero subspace.*

> **Exercise 3.87.1.** The proof of this is very similar to the proof of the corresponding Lemma 2.102 in Chapter 2: study that proof and write down a careful proof of Lemma 3.87 above.

Our next lemma is analogous to Lemma 2.103 of Chapter 2:

Lemma 3.88. *Let V and X be two vector spaces over a field F and let $f : V \to X$ be a linear transformation. Write $f(V)$ for the image of V under f. Then $f(V)$ is a subspace of X.*

Proof. We will apply Corollary 3.56 to $f(V)$. By Lemma 3.77, $0_X \in f(V)$, so $f(V)$ is nonempty. Now take any w_1, w_2 in $f(V)$, and any r_1, r_2 in F. We wish to show that $r_1w_1 + r_2w_2$ is also in $f(V)$. Since $w_1 \in f(V)$, $w_1 = f(v_1)$ for some $v_1 \in V$. Similarly, $w_2 = f(v_2)$ for some $v_2 \in V$. Then, $f(r_1v_1 + r_2v_2) = f(r_1v_1) + f(r_2v_2) = r_1f(v_1) + r_2f(v_2) = r_1w_1 + r_2w_2$, as desired.

□

Lemma 3.88 above allows us to make a quick definition:

Definition 3.89. Let V and X be two vector spaces over a field F and let $f : V \to X$ be a linear transformation. The *rank* of f is defined to be the dimension of $f(V)$ as an F-vector space.

It is tempting to prove the following two easy results before proceeding to vector space isomorphisms:

Lemma 3.90. *Let V and X be vector spaces over a field F and let $f : V \to X$ be a linear transformation. Let B be a basis for V. Then the vectors $\{f(b) \mid b \in B\}$ span $f(V)$.*

Proof. Any vector in $f(V)$ is of the form $f(v)$ for some $v \in V$. Since B is a basis for V, $v = r_1b_1 + \cdots r_nb_n$ for some scalars $r_1, \ldots, r_n$ and some vectors $b_1, \ldots, b_n$ from B. Then, $f(v) = r_1f(b_1) + \cdots + r_nf(b_n)$, showing that every vector in $f(V)$ is expressible as a linear combination of the vectors $\{f(b) \mid b \in B\}$, as desired. □

Lemma 3.91. *Continuing with the notation of Lemma 3.90, assume further that f is injective. Then the vectors $\{f(b) \mid b \in B\}$ form a basis for $f(V)$.*

Proof. With the additional assumption that f is injective, we need to show that the vectors $\{f(b) \mid b \in B\}$ are linearly independent, since we already know from Lemma 3.90 that they span $f(V)$. Assume that $r_1f(b_1) + \cdots + r_nf(b_n) = 0_X$ for some scalars $r_1, \ldots, r_n$ and some vectors $b_1, \ldots, b_n$ from B. Since f is a linear transformation, the left side is just $f(r_1b_1 + \cdots + r_nb_n)$. By the injectivity of f, we find $r_1b_1 + \cdots + r_nb_n = 0_V$. But since the b_i are linearly independent in V, $r_1, \ldots, r_n$ must all be zero, showing that the vectors $\{f(b) \mid b \in B\}$ are indeed linearly independent. □

We now have the following, completely in analogy with rings:

Definition 3.92. Let V and X be vector spaces over a field F, and let $f : V \to X$ be a linear transformation. If f is both injective and surjective, then f is said to be an *isomorphism* between V and X. Two vector spaces V and X are said to be *isomorphic* (written $V \cong X$) if there is some function $f : V \to X$ that is an isomorphism between V and X.

Example 3.93. Any two vector spaces over the same field F of the same dimension n are isomorphic. For, if V and W are two vector spaces over F both of dimension n, and if, say, $\{v_1, v_2, \ldots, v_n\}$ is a basis for V and $\{w_1, w_2, \ldots, w_n\}$ is a basis for W, then the function $f : V \to W$ defined by

$f(v_1) = w_1$, $f(v_2) = w_2$, ..., $f(v_n) = w_n$, and $f(r_1v_1 + r_2v_2 + \cdots + r_nv_n) = r_1w_1 + r_2w_2 + \cdots + r_nw_n$ is an F-linear transformation, by Lemma 3.82.1. This map is injective: if $v = r_1v_1+r_2v_2+\cdots+r_nv_n$ is such that $f(v) = 0$, then this means $r_1w_1+r_2w_2+\cdots+r_nw_n = 0$, and since the w_i form a basis for W, each r_i must be zero, so v must be zero. Also, f is surjective: clearly, given any $w = r_1w_1+r_2w_2+\cdots+r_nw_n$ in W, the vector $v = r_1v_1+r_2v_2+\cdots+r_nv_n$ maps to w under f. Thus, f is an isomorphism between V and W.

Remark 3.93.1. If $f : V \to W$ is an isomorphism between two vector spaces V and W over a field F then, since f provides a bijection between V and W, we may define $f^{-1} : W \to V$ by $f^{-1}(w)$ equals that unique $v \in V$ such that $f(v) = w$. Clearly, the composite function $V \xrightarrow{f} W \xrightarrow{f^{-1}} V$ is just the identity map on V, and similarly, the composite function $W \xrightarrow{f^{-1}} V \xrightarrow{f} W$ is just the identity map on W. But more: the map f^{-1} is a linear transformation from W to V.

> **Exercise 3.93.1.** If $f : V \to W$ is an isomorphism, show that the map f^{-1} of Remark 3.93.1 above is a linear transformation from W to V.

The following is analogous to Theorem 2.110 of Chapter 2:

Theorem 3.94. *(Fundamental Theorem of Linear Transformations of Vector Spaces.) Let V and X be vector spaces over a field F. Let $f : V \to X$ be a linear transformation, and write $f(V)$ for the image of V under f. Then the function $\tilde{f} : V/ker(f) \to f(V)$ defined by $\tilde{f}(v + ker(f)) = f(v)$ is well-defined, and provides an isomorphism between $V/ker(f)$ and $f(V)$.*

Proof. The proof is similar to the proof of 2.110 of Chapter 2. We first check that $\tilde{f}$ is well-defined. Suppose $u + ker(f) = v + ker(f)$. Then $u - v \in ker(f)$, so $f(u - v) = f(u) - f(v) = 0_X$, so $f(u) = f(v)$. Thus, $\tilde{f}(u + ker(f)) = \tilde{f}(v + ker(f))$, i.e., $\tilde{f}$ is well-defined.

Now let us apply Lemma 3.76: We have $\tilde{f}\,(r_1(v_1 + ker(f)) + r_2(v_2 + ker(f))) = \tilde{f}\,((r_1v_1 + ker(f)) + (r_2v_2 + ker(f))) = \tilde{f}\,((r_1v_1 + r_2v_2) + ker(f)) = f(r_1v_1+ r_2v_2) = r_1f(v_1) + r_2f(v_2) = r_1\tilde{f}(v_1 + ker(f)) + r_2\tilde{f}(v_2 + ker(f))$. Hence $\tilde{f}$ is a linear transformation.

> **Exercise 3.94.1.** Justify all the equalities above.

We check that $\tilde{f}$ is surjective as a function from $V/ker(f)$ to $f(V)$. Note that any element of $f(V)$ is, by definition, of the form $f(v)$ for some $v \in V$. But then, by the way we have defined $\tilde{f}$, we find $f(v) = \tilde{f}(v + ker(f))$, so indeed $\tilde{f}$ is surjective.

Finally, we check that $\tilde{f}$ is injective. Suppose that $v + ker(f)$ is in $ker(\tilde{f})$. Thus, $\tilde{f}(v + ker(f)) = 0_X$. Since $\tilde{f}(v + ker(f)) = f(v)$, we find $f(v) = 0_X$. Hence $v \in ker(f)$. But this means that the coset $v + ker(f)$ equals the coset $ker(f)$ (why?), so $v + ker(f)$ is the zero element of $V/ker(f)$. Thus $\tilde{f}$ is injective.

Putting this together, we find that $\tilde{f}$ provides an isomorphism between $V/ker(f)$ and $f(V)$.

□

We now study the relation between the dimensions of V, $ker(f)$ and $f(V)$ in the case where V is finite-dimensional. But first, let us state a consequence of Lemmas 3.90 and 3.91:

Corollary 3.95. *Let V and X be vector spaces over a field F and let $f : V \to X$ be a linear transformation. If f is an isomorphism between V and X, then f sends any basis of V to a basis of X.*

> **Exercise 3.95.1.** Convince yourselves that this follows from Lemmas 3.90 and 3.91!

We are now ready to prove:

Theorem 3.96. *Let V and X be vector spaces over a field F and let $f : V \to X$ be a linear transformation. Assume that V is finite-dimensional. Then $dim_F(V) = dim_F(f(V)) + dim_F(ker(f))$.*

Proof. The proof is a combination of Theorem 3.94, Lemma 3.72, and Corollary 3.95. Start with a basis $\{b_1, \ldots, b_m\}$ of $ker(f)$, and expand this to a basis $\{b_1, \ldots, b_m, b_{m+1}, \ldots, b_n\}$ of V. (Thus, $\dim_F(ker(f)) = m$ and

$\dim_F(V) = n$.) Then, according to that lemma, the set $\{b_{m+1}+ker(f), \ldots, b_n+ker(f)\}$ is a basis for $V/ker(f)$. By Theorem 3.94, the function $\tilde{f} : V/ker(f) \to f(V)$ defined by $\tilde{f}(v + ker(f)) = f(v)$ is an isomorphism, so by Corollary 3.95 the set of vectors $\{\tilde{f}(b_{m+1} + ker(f)), \ldots, \tilde{f}(b_n + ker(f))\}$ forms a basis for $f(V)$. In particular, the dimension of $f(V)$ must be the size of this set, which is $n - m$. It follows that $\dim_F(V) = \dim_F(f(V)) + \dim_F(ker(f))$.

□

3.5 Further Exercises

Exercise 3.97. Starting from the vector space axioms, prove that the properties listed in Remark 3.12 hold for all vector spaces. (Hint: You should get ideas from the solutions to the corresponding Exercise 2.114 of Chapter 2: the proofs of the first three properties are quite similar in spirit. As for the last property, look to f^{-1} for help!)

Exercise 3.98. Prove that the polynomials 1, $1 + x$, $(1 + x)^2$, $(1 + x)^3$, $\ldots$ also form a basis for $\mathbb{R}[x]$ as a $\mathbb{R}$–vector space. (Hint: To show that these polynomials span $\mathbb{R}[x]$, it is sufficient to show that the polynomials 1, x, x^2, $\ldots$ are in the linear span (see Example 3.65 above) of 1, $1 + x$, $(1 + x)^2$, $(1 + x)^3$, $\ldots$ (Why?) The vector 1 is of course in the linear span. Assuming inductively that the vectors 1, x, $\ldots$, and x^{n-1} are in the linear span, show that x^n is also in the linear span by considering the binomial expansion of $(1 + x)^n$. As for linear independence, suppose that $\sum_{i=0}^{n} d_i(1 + x)^i = 0$. You may assume that $d_n \neq 0$ (why?) Now expand each term $(1 + x)^i$ above and consider the coefficient of x^n. What do you find?)

If you find the hint too computational, you can also establish this result by invoking Exercise 3.106 ahead and Exercise 2.109.2 in Chapter 2. (However, note that Exercise 2.109.2 in turn is computational, so this merely shifts all the computations to a different place!)

Exercise 3.99. Show that the matrices $e_{i,j}$ and $\sqrt{2}e_{i,j}$ $(1 \leq i, j \leq 2)$ form a basis for $M_2(\mathbb{Q}[\sqrt{2}])$ considered as a $\mathbb{Q}$–vector space. ($\sqrt{2}e_{i,j}$ is the 2×2 matrix with $\sqrt{2}$ in the (i, j) slot, and zeros in the remaining slots.) Now discover a basis for $M_2(\mathbb{C})$ considered as a vector space over $\mathbb{R}$.

Exercise 3.100. Show that the set of all matrices in $M_n(\mathbb{R})$ whose trace is zero is a subspace of $M_n(\mathbb{R})$ by exhibiting this space as the kernel of a suitable homomorphism that we have considered in the text. Use Theorem 3.96 to prove that this subspace has dimension $n^2 - 1$. Discover a basis for this subspace.

Exercise 3.101. Let V be an F-vector space. So far, we have considered individual linear tranformations of the form $f : V \to V$; this exercise deals with the collection of *all* such F-linear transformations. Let $End_F(V)$ denote the set of all F-linear transformations from V to V. ("End" is short for the word "endomorphism," which is another word for a homomorphism from one (abelian) group to itself, while the subscript F indicates that we are considering those (abelian) group homomorphisms that are in addition F-linear—see (5) in Remark 3.75 earlier in this chapter.)

1. Let f and g be two elements in $End_F(V)$. Consider the function, suggestively denoted "$f+g$" that is obtained by defining $(f+g)(v) = f(v)+g(v)$. Show that $f+g$ is also an F-linear transformation, and hence is an element of $End_F(V)$.

2. Show that $End_F(V)$, with this definition of addition of two linear transformations, is an abelian group. What is the identity element in this group? How do you define the inverse with respect to addition of any $f \in End_F(V)$?

3. Let $f \circ g$ denote the usual composition of functions on V, defined by $(f \circ g)(v) = f(g(v))$. Show that $f \circ g$ is also an F-linear transformation, and hence is an element of $End_F(V)$.

4. Show that by thinking of function composition "$\circ$" as a multiplication operation on $End_F(V)$, the set $(End_F(V), +, \circ)$ becomes a ring. What is the multiplicative identity in this ring? Is this ring commutative? (What if the dimension of V is 1?)

Exercise 3.102. Prove that an element $f \in End_F(V)$ (see Exercise 3.101 above) is invertible if and only if f is an isomorphism. (Hint: For one direction of this problem, Remark 3.93.1 and Exercise 3.93.1 may be helpful.)

Exercise 3.103. Now that you have shown that $End_F(V)$ is a ring in Exercise 3.101 above, here is an example that shows $ab = 1$ doesn't imply $ba = 1$ in an arbitrary ring! (See Definition 2.44 in Chapter 2.)

Let V be a vector space with a countably infinite basis v_i, $i \in \mathbb{Z}$. (For example, see Exercise 3.63.2 earlier in this chapter.) Let T be the F-linear transformation that sends v_i to v_{i+1} for $i = 1, 2, \ldots$, and let S be the linear transformation that sends v_i to v_{i-1} for $i = 1, 2, \ldots$ with the understanding that v_0 means the zero vector. (Why are these linear transformations? See the remarks on page 155 on how to define linear transformations between infinite-dimensional spaces.) Show that in the ring $End_F(V)$, the product $ST = 1$ but the product TS sends v_1 to zero and hence is not 1.

Exercise 3.104. Let V be an F-vector space of dimension n with basis $\{b_1, \ldots, b_n\}$. Recall from Example 3.82 how one can assign to each F-linear transformation T on V the $(n \times n)$ matrix of T with respect to the basis $\{b_1, \ldots, b_n\}$. Write M_T for the matrix in $M_n(F)$ that corresponds to T under this assignment. Study the addition and multiplication operations on $End_F(V)$ in Exercise 3.101 above, and prove that the map $M : End_F(V) \to M_n(F)$ that sends T to M_T provides a *ring isomorphism* between $End_F(V)$ and $M_n(F)$.

Exercise 3.105. Let K/F be a field extension. By Example 3.7, K may be viewed as an F-vector space. Assume that the dimension of K as an F-vector space is n. This exercise shows how K may be realized as a subring of $M_n(F)$, thus generalizing Example 2.108 in Chapter 2.

1. For each $k \in K$, write l_k for the map from K to K that sends any $x \in K$ to kx. Show that l_k is an F-linear transformation from K to K.

2. Recall from Exercise 3.101 that $End_F(V)$, the set of all F-linear transformations of an F-vector space V, is a ring, under the operation of composition of functions. In particular, viewing K as an F-vector space, $End_F(K)$ is a ring, and the linear transform l_k of Part (1) above is an element of this ring. Show that the map $l : K \to End_F(K)$ that sends $k \in K$ to the linear transform l_k is an *injective* ring homomorphism from K to $End_F(K)$.

3. Let $\{b_1, \ldots, b_n\} \subseteq K$ be an F-basis of K. The linear transformation l_k corresponds to a matrix M_{l_k} with respect to the basis $\{b_1, \ldots, b_n\} \subseteq K$ (as in Example 3.82). Show that the map from K to $M_n(F)$ that sends k to M_{l_k} is a *ring homomorphism.*

   (Hint: By Exercise 3.104 above, $End_F(K)$ is isomorphic to $M_n(F)$ via the map M that sends a linear transform T to its matrix M_T written in the basis $\{b_1, \ldots, b_n\}$. Compose the map $l : K \to End_F(K)$ with the map $M : End_F(K) \to M_n(F)$.)

4. Show that this ring homomorphism in (3) above is injective. Conclude that K is isomorphic to a subring of $M_n(F)$ using Lemma 2.103 and Theorem 2.110 of Chapter 2.

The image of K under the homomorphism in (3) above is called the *regular representation of K in $M_n(F)$.*

Exercise 3.106. Let R be a ring containing a field F, so R is an F-vector space (see Example 3.8 earlier in this chapter). Let $f\colon R \to R$ be a ring isomorphism that *acts as the identity* on F (i.e., $f(r) = r$ for all $r \in F$). Show that if $B \subset R$ is an F-basis of R, then the set $f(B) = \{f(b) \mid b \in B\}$ is also an F-basis of R.

Exercise 3.107. Recall from Exercise 2.123 in Chapter 2 that the set S of all functions from $\mathbb{R}$ to $\mathbb{R}$ is a ring under the operation of pointwise addition and multiplication of functions. Since, by that same exercise, the set of constant functions is a subring of S that is isomorphic to $\mathbb{R}$, S carries the natural structure of a $\mathbb{R}$-vector space. (Explicitly, the vector space structure is given by the map $\mathbb{R} \times S \to S$ that sends (r, f) to the function $s_r \cdot f$, where s_r is as in Exercise 2.123. More simply, however, the product of the real number r and the function $f(x)$ is the function, suggestively denoted $r \cdot f$, defined by $(r \cdot f)(x) = rf(x)$.)

1. Which of the following are subspaces of S?
   (a) $\{f \in R \mid f(1) = 0\}$
   (b) $\{f \in R \mid f(0) = 1\}$
   (c) The set of all constant functions.
   (d) $\{f \in R \mid f(x) \geq 0$ for all $x \in \mathbb{R}\}$
2. Show that the set $\{1, \sin^2(x), \cos^2(x)\}$ is linearly dependent.
3. Is the set $\{e^x, 1, x, x^2, x^3, \dots\}$ linearly dependent or independent?

Exercise 3.108. Prove Proposition 3.49 without the assumption that V is finite-dimensional. (See the notes on page 224 in Chapter B in the Appendix for hints.)

Exercise 3.109. This exercise shows that Lemma 3.72 holds even for infinite-dimensional spaces. Let V be a vector space over a field F and let W be a subspace. Let B be a basis for W. Expand this to a basis S of V (see Proposition 3.49, as well as the remarks on page 224 in Chapter B in the Appendix). Write T for $S - B$ (so S is the disjoint union of B and T). Prove that the set (of equivalence classes of vectors) $\{t + W \mid t \in T\}$ is a basis for the quotient space V/W.

Exercise 3.110. If V is a finite-dimensional vector space and if W is a subspace of V, prove that the dimension of W is no bigger than the dimension of V. Now prove that if the dimension of W and V are equal, then $W = V$.

Exercise 3.111. Let V be a vector space over a field F, and let U and W be two subspaces.

1. Show that $U \cap W$ is a subspace of V. (Is $U \cup W$ a subspace of V?)

2. Denote by $U + W$ the set $\{u + w \mid u \in U \text{ and } w \in W\}$. Show that $U + W$ is a subspace of V.

3. Now assume that V is *finite-dimensional*. The aim of this part is to establish the following:

$$\dim(U + W) = \dim(U) + \dim(W) - \dim(U \cap W)$$

   (a) Let $\{v_1, \dots, v_p\}$ be a basis for $U \cap W$ (so $\dim(U \cap W) = p$). Expand this to a basis $\{v_1, \dots, v_p, u_1, \dots, u_q\}$ of U, and also to a basis $\{v_1, \dots, v_p, w_1, \dots, w_r\}$ of W (so $\dim(U) = p + q$ and $\dim(W) = p + r$). Show that the set $B = \{v_1, \dots, v_p, u_1, \dots, u_q, w_1, \dots, w_r\}$ spans $U + W$.

   (b) Show that the set B is linearly independent. (Hint: Assume that we have the relation $f_1 v_1 + \cdots + f_p v_p + g_1 u_1 + \cdots + g_q u_q + h_1 w_1 + \cdots + h_r w_r = 0$. Rewrite this as $g_1 u_1 + \cdots + g_q u_q = -(f_1 v_1 + \cdots + f_p v_p + h_1 w_1 + \cdots + h_r w_r)$. Observe that the left side is in U while the right is in W, so $g_1 u_1 + \cdots + g_q u_q$ must be in $U \cap W$. Hence, $g_1 u_1 + \cdots + g_q u_q = j_1 v_1 + \cdots + j_p v_p$ for some scalars $j_1, \dots, j_p$. Why does this show that the g_i must be zero? Now proceed to show that the f_i and the h_i must also be zero.)

   (c) Conclude that $\dim(U + W) = \dim(U) + \dim(W) - \dim(U \cap W)$.

   (d) Prove that any two 2-dimensional spaces of $\mathbb{R}^3$ must intersect in a space of dimension at least 1.

Exercise 3.112. Show that the $nth\ Bernstein\ Poylnomials$ $B_i^{(n)}(x) = \binom{n}{i} x^i (1-x)^{n-i}$, $(i = 0, 1, \dots, n)$ form a basis for $\mathbb{R}_n[x]$ $(n \geq 1)$ as follows:

1. Show that $1 = \sum_{i=0}^{n} B_i^{(n)}$.

2. The equation in part 1 above continues to hold if we replace n by $n - 1$ everywhere. (Why?) Make this replacement, multiply throughout by x, and derive the relation $x = \sum_{i=0}^{n} (i/n) B_i^{(n)}$. (Hint: you will need to use the relation $\binom{n-1}{i-1} = (i/n)\binom{n}{i}$. Why does this last relation hold?)

3. Similarly, for $k = 2, \ldots, n-1$, show that $x^k = \sum_{i=0}^{n}(i(i-1)\cdots(i-k+1)/n(n-1)\cdots(n-k+1))B_i^{(n)}$.

4. Now conclude that the $B_i^{(n)}$ span $\mathbb{R}_n[x]$.

5. Use Proposition 3.51 above to conclude that the $B_i^{(n)}$ form a basis.

These Bernstein polynomials find applications in diverse areas of mathematics, as well as in various applied fields, such as computer graphics! For instance, in advanced calculus, they are useful in showing that any continuous function on an interval $[a, b]$ can be approximated arbitrarily closely by a *polynomial* function. (This is known as the *Weierstrass Approximation Theorem.*) In computer graphics, they are used to fit, through a given set of points, a curve that is smooth and has minimal "wiggle," and as well, to provide convenient handles by which the user can then control the shape of this curve.

Notes

Remarks on Example 3.5 It is worth remarking that our definition of scalar multiplication is a very natural one. First, observe that we can consider $\mathbb{R}$ to be a subring of $M_n(\mathbb{R})$ in the following way: the set of matrices of the form $\text{diag}(r)$, as r ranges through $\mathbb{R}$, is essentially the same as $\mathbb{R}$ (see Example 2.106 in Chapter 2). (Observe that this makes the set of diagonal matrices of the form $\text{diag}(r)$ a field in its own right!) Under this identification of $r \in \mathbb{R}$ with $\text{diag}(r)$, what is the most natural way to multiply a scalar r and a vector $(a_{i,j})$? Well, we think of r as $\text{diag}(r)$, and then define $r \cdot (a_{i,j})$ as just the usual product of the two matrices $\text{diag}(r)$ and $(a_{i,j})$. But, as you can check easily, the product of $\text{diag}(r)$ and $(a_{i,j})$ is just $(ra_{i,j})$! It is in this sense that our definition of scalar multiplication is natural—it arises from the rules of matrix multiplication itself. Notice that once $\mathbb{R}$ has been identified with the subring of $M_n(\mathbb{R})$ consisting of the set of matrices of the form $\text{diag}(r)$, this example is just another special case of Example 3.8.

Remarks on Example 3.10 $(V, +)$ remains an abelian group. This does not change when we restrict our attention to the subfield F. So we only need to worry about what the new scalar multiplication ought to be. But there is a natural way to multiply any element f of F with any element v of V: simply consider f as an element of K, and use the multiplication already defined between elements of K

and elements of V! The scalar multiplication axioms clearly hold: for any f and g in F and any v and w in V, we may first think of f and g as elements of K, and since the scalar multiplication axioms hold for V viewed as a vector space over K, we certainly have $f\cdot(v+w) = f\cdot v+f\cdot w$, $(f+g)\cdot v = f\cdot v+g\cdot v$, $(fg)\cdot v = f\cdot(g\cdot v)$, and $1\cdot v = v$.

Remarks on Example 3.34 This example is a bit tricky. Why are the e_i not a basis? They are certainly linearly independent, since if $\sum_{i=0}^{n} c_i e_i = 0$ for some scalars $c_i \in F$, then the tuple $(c_0, c_1, \ldots, c_n, 0, 0, \ldots)$ must be zero, but a tuple is zero if and only if each of its components is zero. Thus, each of c_0, c_1, …, c_n must be zero, proving linear independence. However, the e_i *do not span* $\prod_0^\infty F$, contrary to what one might expect. To understand this, let us look at something that has been implicit all along in the definition of linear combination. The e_i would span $\prod_0^\infty F$ if every vector in $\prod_0^\infty F$ could be written as a linear combination of elements of the set $\{e_0, e_1, e_2, \ldots\}$. Now notice that whenever we consider linear combinations, we only consider sums of a *finite* number of terms. Hence, a linear combination of elements of the set $\{e_0, e_1, e_2, \ldots\}$ looks like $c_{i_1}e_{i_1} + c_{i_2}e_{i_2} + \cdots + c_{i_n}e_{i_n}$ for some *finite* n. It is clear that any vector that is expressible in such a manner will have only finitely many components that are nonzero. (These will be at most the ones at the slots i_1, i_2, …, i_n; all other components will be zero.) Consequently, the vectors in $\prod_0^\infty F$ in which infinitely many components are nonzero (for example, the vector $(1, 1, 1, \ldots)$), cannot be expressed as linear combinations of the e_i.

On the other hand, see Exercise 3.63.2.

It is worth pointing out that infinite sums have no algebraic meaning. Addition is, to begin with, a binary operation, that is, it is a rule that assigns to a_1 and a_2 the element $a_1 + a_2$. This can be extended inductively to a finite number of a_i: for instance, the sum $a_1 + a_2 + a_3 + a_4 + a_5$ is defined as $(((a_1 + a_2) + a_3) + a_4) + a_5$. (In other words, we first determine $a_1 + a_2$, then we add a_3 to this, then a_4 to what we get from adding a_3, and then finally a_5 to what we got at the previous step.) While this inductive definition makes sense for a *finite* number of terms, it makes no sense for an infinite number of terms. To interpret infinite sums of elements, we really need to have a notion of convergence (such as the ones you may have seen in a course on real analysis). Such notions may not exist for arbitrary fields.

Remarks on the proof Theorem 3.70 The reason why the proofs that $(V/W,+)$ and $(R/I,+)$ are abelian groups are so similar is that what we are essentially proving in both is that if $(G,+)$ is an abelian group and if H is a subgroup, then the set of equivalence classes of G under the relation $g_1 \sim g_2$ if and only if $g_1 - g_2 \in H$ with the operation $[g_1]+[g_2]=[g_1+g_2]$ is indeed an abelian group in its own right! We will take this up in Chapter 4 ahead.

Remarks on linear transformations $f : V \to X$ when V or X are not necessarily finite-dimensional Similar considerations will apply: we let $S=\{b_\beta \mid \beta \in B\}$ be a basis for V, where B is some index set. Let $\{w_\beta \mid \beta \in B\}$ be arbitrary vectors in X. Every vector in V can be uniquely written as $r_1 b_{\beta_1} + \cdots + r_k b_{\beta_k}$, where the r_i are scalars from the field F and $\{b_{\beta_1}, \ldots, b_{\beta_k}\}$ is some *finite* subset of S. Then, just as in the finite-dimensional case, the function $f : V \to X$ that sends $r_1 b_{\beta_1} + \cdots r_k b_{\beta_k}$ to $r_1 w_{\beta_1} + \cdots r_k w_{\beta_k}$ is a linear transformation, and all linear transformations from V to X are given in this way. Let $T=\{c_\gamma \mid \gamma \in C\}$ be a given basis of X (again, C is some index set). The matrix representation of f with respect to the basis S of V and T of X is a $|B| \times |C|$ matrix (where $|B|$ and $|C|$ are the cardinality of the possibly infinite sets B and C), with the rows indexed by the basis vectors in T and the columns indexed by the basis vectors in S. The column with index β represents the image of b_β under f, written as a column vector, whose entry in the row indexed by γ is the coefficient of c_γ (in the expression of $f(b_\beta)$ as a linear combination of the c_γ). Note that since any vector is always expressed as a *finite* linear combination of the basis vectors in C (see the remarks on Example 3.34 on page 154), each column of the matrix will have only *finitely-many nonzero entries.* Conversely, given any $|B| \times |C|$ matrix with entries in F in which each column has only finitely many nonzero entries, one can defined a linear transformation $f : V \to X$ exactly as in Example 3.82, with the column indexed by β corresponding to $f(b_\beta)$.

Chapter 4

Groups

4.1 Groups: Definition and Examples

Of all the algebraic objects that we have considered in this course–groups, rings, fields, and vector spaces–groups are technically the most elementary: they are sets with just one binary operation, and there are just three axioms that govern them: (i) the binary operation should be associative, (ii) there should be an identity for this operation, and (iii) every element should have an inverse with respect to this operation (See Definition 2.2 in Chapter 2). Yet, we have reserved our study of groups to the last and have started with rings instead. The primary reason for this is that even if they are technically more complicated than groups, rings are a much more familiar object to most students who are seeing abstract algebra for the first time: after all, the "number systems" that we have grown up with and are so intimate with, namely the integers, the rationals, the reals, and the complexes, are all examples of rings. Rings are thus, for many, a natural entry point into algebra. In the same vein, examples like $\mathbb{R}^2$ and $\mathbb{R}^3$ make vector spaces also a familiar object, and their study is therefore a natural candidate to follow our study of rings.

However, let neither their elementary definition nor the location of this chapter in this book lull you into underestimating the importance of groups: groups are *vitally* important in mathematics, and they show up in just about every nook and corner of the subject. Although this may not be obvious from the examples that we have seen so far (which have all been groups of the form $(R, +)$, where R is a ring and $+$ is the addition operation on the ring, or of the form $(R^*, \cdot)$, the set of invertible elements of a ring R under multiplication), groups are objects by which one measures *symmetry* in mathematical objects. Of course, what a mathematician means by symmetry is something very abstract, but it is merely a generalization (albeit a vast one) of what people mean by symmetry in day-to-day contexts. Since symmetry is so central to mathematics (one view would have it that all of mathematics is a study of symmetry!), it should come as no surprise that groups are central to mathematics.

Here is what a mathematician would mean by symmetry. Suppose you have a set, and suppose the set has some *structure* on it. To say that the set has some structure is merely to say that it has some specific feature that we are focusing on for the moment: the set could have lots of other features as well, but we will ignore those temporarily. A *symmetry* of a set with a given structure is merely a bijective correspondence from the set to itself which preserves the structure, i.e., a one-to-one onto map from the set to itself which preserves the feature that we are considering. The set of all such one-to-one and onto maps whose inverse also preserves the feature that we are considering constitute a group, which is called the *symmetry group* of the set for the given structure, and both the size and the nature of this group then quantify the kind of symmetry that the set with structure has.

Now this is too advanced for a first reading, so we will postpone consideration of sets with structure and their symmetry groups to the notes at the end of the chapter (see Page 206). But first, let us repeat the definition of groups from Chapter 2, just so to have the definition within this chapter:

Definition 4.1. (Repeat of Definition 2.2) A *group* is a set S with a binary operation "$*$" : $S \times S \to S$ such that

1. $*$ is associative, i.e., $a * (b * c) = (a * b) * c$ for all a, b, and c in S,
2. S has an identity element with respect to $*$, i.e., an element "id" such that $a * id = id * a = a$ for all a in S, and
3. every element of S has an inverse with respect to $*$, i.e., for every element a in S there exists an element "a^{-1}" such that $a * a^{-1} = a^{-1} * a = id$.

To emphasize that there are two ingredients in this definition—the set S and the operation $*$ with these special properties—the group is sometimes written as $(S, *)$, and S is often referred to as *a group with respect to the operation* $*$.

Recall from Chapter 2 (Definition 2.3) that an abelian group is one in which the group operation is commutative, i.e., $a * b = b * a$ for all a and b in the group.

Here are some examples of groups other than those that appear as the additive group of a ring or the multiplicative group of a field:

4.1.1 Symmetric groups

Example 4.2. Consider the set $\Sigma_3 = \{1, 2, 3\}$, and consider one-to-one and onto maps from Σ_3 to itself: in more common language, such maps are known as *permutations* of $\{1, 2, 3\}$. Let us, for example, write $\binom{1\ 2\ 3}{2\ 3\ 1}$ for the permutation that sends 1 to 2, 2 to 3, and 3 to 1 (so we write the image of an element under the element, we will call this the *stack notation*). Then it is easy to see that there are exactly six permutations, and they are listed in the following table (where we have given a name to each permutation):

id	$\binom{1\ 2\ 3}{1\ 2\ 3}$
r_1	$\binom{1\ 2\ 3}{2\ 3\ 1}$
r_2	$\binom{1\ 2\ 3}{3\ 1\ 2}$
f_1	$\binom{1\ 2\ 3}{1\ 3\ 2}$
f_2	$\binom{1\ 2\ 3}{3\ 2\ 1}$
f_3	$\binom{1\ 2\ 3}{2\ 1\ 3}$

Now let us see how these permutations compose. You will observe that $r_1 \circ r_1$ takes 1 to 2 under the first application of r_1 and then 2 to 3 under the second application of r_1. Likewise, $r_1 \circ r_1$ takes 2 to 3 and then to 1, and similarly, 3 to 1 and then to 2. The net result: $r_1 \circ r_1$ is the permutation $\binom{1\ 2\ 3}{3\ 1\ 2}$, that is, $r_1 \circ r_1 = r_2$!

Now play with $r_1 \circ f_1$ and compare it with $f_1 \circ r_1$. When computing $r_1 \circ f_1$, for instance, we observe that 1 goes to 1 under f_1, and then 1 goes to 2 under r_1. Computing fully, we find that $r_1 \circ f_1$ is the permutation $\binom{1\ 2\ 3}{2\ 1\ 3}$, that is, $r_1 \circ f_1 = f_3$. Computing $f_1 \circ r_1$ similarly, we find $f_1 \circ r_1 = f_2$!

Computing fully, we find the following table that describes how the six permutations of S compose:

$\circ$	id	r_1	r_2	f_1	f_2	f_3
id	id	r_1	r_2	f_1	f_2	f_3
r_1	r_1	r_2	id	f_3	f_1	f_2
r_2	r_2	id	r_1	f_2	f_3	f_1
f_1	f_1	f_2	f_3	id	r_1	r_2
f_2	f_2	f_3	f_1	r_2	id	r_1
f_3	f_3	f_1	f_2	r_1	r_2	id

Now observe the following: (i) The composition of two permutations of Σ_3 is another permutation of Σ_3: we computed this out explicitly above, but we already would have known this from an earlier exposure to functions: if $f : S \to S$ and $g : S \to S$ are functions (here S is some set) and if both f and g are bijective, then both compositions $(g \circ f)$ and $(f \circ g)$ from S to S are also bijective, (ii) Composition of functions is an associative operation: this too would be familiar to us from our earlier exposure to functions: if $f : S \to S$, $g : S \to S$, and $h : S \to S$ are three functions on some set S, then for any $s \in S$, $((f \circ g) \circ h)(s) = (f \circ g)(h(s)) = f(g(h(s)))$, while $(f \circ (g \circ h))(s) = f((g \circ h)(s)) = f(g(h(s)))$, so indeed $(f \circ g) \circ h = f \circ (g \circ h)$, (iii) The permutation id acts as the identity element: this is clear from the first row and the first column of the table above, and finally, (iv) Every permutation of S_3 has an inverse: $r_1 \circ r_2 = r_2 \circ r_1 = id$, $id \circ id = id$, $f_1 \circ f_1 = id$, $f_2 \circ f_2 = id$ and $f_3 \circ f_3 = id$. Hence, the set of permutations of Σ_3 forms a group under composition. We denote this group as S_3, and call it the *symmetric group on three elements.* (S_3 can be interpreted as the set of symmetries of Σ_3 with the trivial structure: see Example 4.88 in the notes at the end of the chapter.)

Observe something about this group: it is not a commutative group! For instance, as we observed above, $r_1 \circ f_1 = f_3$ while $f_1 \circ r_1 = f_2$. We say that the group is *nonabelian.*

From now on we will suppress the "$\circ$" symbol, and simply write fg for the composition $f \circ g$. Not only is there less writing involved, but it is notation that we are used to: it is the notation we use for multiplication. Continuing the analogy, we write $f \circ f$ as f^2, and so on, and we sometimes write 1 for the identity (see Remark 4.14 ahead for more on the notation used for the identity and the group operation). In this notation, note that $r_1^3 = r_2^3 = 1$, $f_1^2 = f_2^2 = f_3^2 = 1$.

The table such as the one above that describes how pairs of elements in a group compose under the given binary operation is called the *group table* for the group.

Exercise 4.2.1. Use the group table to show that every element of S_3 can be written as $r_1^i f_1^j$ for *uniquely determined integers* $i \in \{0, 1, 2\}$ and $j \in \{0, 1\}$.

Example 4.3. Just as we considered the set of permutations of the set $\Sigma_3 = \{1, 2, 3\}$ above, we can consider for any integer $n \geq 1$, the permutations of the set $\Sigma_n = \{1, 2, \ldots, n\}$. This set forms a group under composition, just as S_2 and S_3 did above.

Definition 4.3.1. The set of permutations of Σ_n, which forms a group under composition, is denoted S_n and is called the *symmetric group on n elements.*

Exercise 4.3.1. Write down the set of permutations of the set $\Sigma_2 = 1, 2$ and construct the table that describes how the permutations compose. Verify that the set of permutations of Σ_2 forms a group. Is it abelian? This group is denoted S_2, and called the *symmetric group on two elements.*

Exercise 4.3.2. Compare the group table of S_2 that you get in the exercise above with the table for $(\mathbb{Z}/2\mathbb{Z}, +)$ on page 36. What similarities do you see?

Exercise 4.3.3. Prove that S_n has $n!$ elements.

Exercise 4.3.4. Find an element $g \in S_n$ such that $g^n = 1$ but $g^t \neq 1$ (see Remark 4.14 ahead on notation for the identity element) for any positive integer $t < n$.

Here is an alternative notation that is used for a special class of permutations, which we will call the *cycle notation*: Working for the sake of concreteness in Σ_5, consider the permutation that sends 1 to 3, 3 to 4, and 4 back to 1, and acts as the identity on the remaining elements 2 and 5. (This is the permutation we have denoted up to now as $\binom{1\ 2\ 3\ 4\ 5}{3\ 2\ 4\ 1\ 5}$.) Notice the "cyclic" nature of this permutation: it moves 1 to 3 to 4 back to 1, and leaves 2 and 5 untouched. We will use the notation $(1,\ 3,\ 4)$ for this special permutation and call it a "3-cycle." In general, if $a_1,\ \ldots,\ a_d$ are distinct elements of the set Σ_n (so $1 \leq d \leq n$), we will denote by $(a_1,\ a_2, \ldots,\ a_d)$ the permutation that sends a_1 to a_2, a_2 to a_3, $\ldots$, a_{d-1} to a_d, a_d back to a_1, and acts as the identity on all elements of Σ_n other than these a_i. We will refer to $(a_1,\ a_2,\ \ldots,\ a_d)$ as a *d-cycle* or a *cycle of length d*. A 2-cycle $(a_1,\ a_2)$ is known as a transposition, since it only swaps a_1 and a_2 and leaves all other elements unchanged. Of course a 1-cycle (a_1) is really just the identity element since it sends a_1 to a_1 and acts as the identity on all other elements of Σ_n.

Notice something about cycles: the cycle $(1,\ 3,\ 4)$ is the same as $(3,\ 4,\ 1)$, as they both clearly represent the same permutation. More generally, the cycle $(a_1,\ a_2, \ldots,\ a_d)$ is the same as $(a_2,\ a_3,\ \ldots,\ a_d,\ a_1)$, which is the same as

$(a_3,\ a_4,\ \ldots,\ a_d,\ a_1,\ a_2)$, etc. We will refer to these different representations of the same cycle as *internal cyclic rearrangements.*

Since a d-cycle is just a special case of a permutation, it makes perfect sense to compose a d-cycle and an e-cycle: it is just the composition of two (albeit special) permutations. For instance, in any S_n (for $n \geq 3$), we have the relation $(1,\ 3)(1,\ 2) = (1,\ 2,\ 3)$ (check!). (We will see shortly—Corollary 4.3.1 ahead—that every permutation in S_n can be "factored" into transpositions.)

> **Exercise 4.3.5.** Write the 4-cycle $(1,\ 2,\ 3,\ 4)$ of S_n (here n is at least 4) as a product of three transpositions.

> **Exercise 4.3.6.** Show that any k cycle in S_n (here $n \geq k \geq 2$) can be written as the product of $k-1$ transpositions.

Two cycles $(a_1,\ \ldots,\ a_d)$ and $(b_1,\ \ldots,\ b_e)$ are said to be disjoint if none of the integers $a_1,\ \ldots,\ a_d$ appear among the integers $b_1, \ldots, b_e$ and none of the integers $b_1,\ \ldots,\ b_e$ appear among the integers $a_1,\ \ldots,\ a_d$. For example, in S_6, the cycles $s = (1,\ 4,\ 5)$ and $t = (2,\ 3)$ are disjoint. Notice something with this pair of permutations: s and t commute! Let us rewrite s and t in the stack notation and compute:

$$\begin{aligned}
st &= \begin{pmatrix} 1\,2\,3\,4\,5\,6 \\ 4\,2\,3\,5\,1\,6 \end{pmatrix} \begin{pmatrix} 1\,2\,3\,4\,5\,6 \\ 1\,3\,2\,4\,5\,6 \end{pmatrix} = \begin{pmatrix} 1\,2\,3\,4\,5\,6 \\ 4\,3\,2\,5\,1\,6 \end{pmatrix} \\
ts &= \begin{pmatrix} 1\,2\,3\,4\,5\,6 \\ 1\,3\,2\,4\,5\,6 \end{pmatrix} \begin{pmatrix} 1\,2\,3\,4\,5\,6 \\ 4\,2\,3\,5\,1\,6 \end{pmatrix} = \begin{pmatrix} 1\,2\,3\,4\,5\,6 \\ 4\,3\,2\,5\,1\,6 \end{pmatrix}
\end{aligned}$$

This computation is of course very explicit, but the intuitive idea behind why s and t commute is the following: s only moves the elements 1, 4, and 5 among themselves, and in particular, it leaves the elements 2 and 3 untouched. On the other hand, t swaps the elements 2 and 3, and leaves the elements 1, 4 and 5 untouched. Since s and t operate on disjoint sets of elements, the action of s is not affected by t and the action of t is not effected by s. In particular, it makes no difference whether we perform s first and then t or the other way around.

Essentially these same ideas lead to the following:

Lemma 4.3.1. *Let s and t be any two disjoint cycles in S_n. Then s and t commute.*

> **Exercise 4.3.7.** Prove this assertion carefully by writing $s = (a_1, \ldots, a_d)$ and $t = (b_1, \ldots, b_e)$ for disjoint integers $a_1, \ldots, a_d, b_1, \ldots, b_d$, and writing out the effect of both st and ts on each integer $1, \ldots, n$. (See the notes on page 210 for some hints.)

Now let us consider another feature of permutations: it turns out that any permutation can be decomposed into a product of disjoint cycles! To take an example, consider the permutation $s = \binom{1\ 2\ 3\ 4\ 5\ 6}{3\ 2\ 1\ 6\ 4\ 5}$. Let us take the element 1 and follow it under repeated action of s: 1 goes to 3 which goes back to 1. Thus, the effect of s on the subset $\{1, 3\}$ is to act as a swap, or a transposition. Now pick another element not equal to either 1 or 3, say 2, and follow it under repeated action of s: 2 stays untouched. Thus, the effect of s on the subset $\{2\}$ is to act as the identity. So now, pick an element not equal to either 1 or 3 or 2, say 4: we find 4 goes to 6 which goes to 5 which then goes back to 4. Hence, the effect of s on the subset $\{4,\ 5,\ 6\}$ is to act as the 3-cycle $(4,\ 6,\ 5)$. It is now easy to see, either by explicitly computing, or by using the same intuition as we did above for why disjoint cycles commute, that $s = \binom{1\ 2\ 3\ 4\ 5\ 6}{3\ 2\ 1\ 6\ 4\ 5} = (4,\ 6,\ 5)(2)(1,\ 3)$. (Since (2) is just the identity permutation, it is typically omitted, and this product is written as $(4,\ 6,\ 5)(1,\ 3)$.

Notice that since disjoint cycles commute, $(4,\ 6,\ 5)(1,\ 3)$ is the same as $(1,\ 3)(4,\ 6,\ 5)$. Notice, too, that had we started with, for instance, 6 and followed it around, and then picked 3 and followed it around, we have found $s = (3,\ 1)(6,\ 5,\ 4)$. Any other decomposition of s into disjoint cycles must be related to the first decomposition $s = (4,\ 6,\ 5)(1,\ 3)$ in a similar manner as these two above: either the cycles could have been swapped, or internally, a cycle could have been rearranged cyclically (such as $(6,\ 5,\ 4)$ instead of $(4,\ 6,\ 5)$). This is because, the product of disjoint cycles simply follows, one by one, the various elements of $\{1\ 2\ 3\ 4\ 5\ 6\}$ under repeated action of s, and no matter in which manner the cycles are written, the repeated action of s must be the same.

These same ideas apply to arbitrary permutations, and we have the following (whose proof we omit because it is somewhat tedious to write in full generality):

Proposition 4.3.1. *Every permutation in S_n factors into a product of disjoint cycles. Two factorizations can only differ in the order in which the cycles appear, or, within any one cycle, by an internal cyclic rearrangement.*

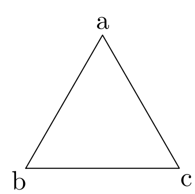
Corollary 4.3.1. *Every permutation in S_n can be written as a product of transpositions.*

Proof. This is just a combination of Proposition 4.3.1 and Exercise 4.3.6 above, which establishes that every cycle can be written as a product of transpositions. □

Remark 4.3.1. Unlike the factorization of a permutation into disjoint cycles, there is no uniqueness to the factorization into transpositions. (For instance, in addition to the factorization $(1,\ 3)(1,\ 2) = (1,\ 2,\ 3)$ we had before, we also find $(1,\ 2)(3,\ 2) = (1,\ 2,\ 3)$.) But something a little weaker than uniqueness holds even for factorizations into transpositions: if a permutation s has two factorizations $s = d_1 \cdot d_2 \cdot \cdots \cdot d_l$ and $s = e_1 \cdot e_2 \cdot \cdots \cdot e_m$ where the d_i and e_j are transpositions, then either l and m will both be even or both be odd! (The proof is slightly complicated, and we will omit it since this is an introduction to the subject.) This allows us to define unambiguously the *parity* of a permutation: we call a permutation *even* if the number of transpositions that appear in any factorization into transpositions is even, and likewise, we call it *odd* if this number is odd.

4.1.2 Dihedral groups

Example 4.4. Consider a piece of cardboard in the shape of an equilateral triangle. Now consider all operations we can perform on the piece of cardboard that do not shrink, stretch, or in anyway distort the triangle, but are such that after we perform the operation, nobody can tell that we did anything to the triangle! To help determine what such operations could be, pretend that the piece of cardboard has been placed at a fixed location on a table, and the location has been marked by lines drawn under the edges of the cardboard. Also, label the points on the table that lie directly under the vertices of the triangle as a, b, and c respectively. After we have done our (yet to be determined!) operation on the cardboard, the triangle should stay at the same location–otherwise it would be obvious that somebody has done something to the piece of cardboard. This means that after our operation, each vertex of the triangle must somehow end up once again on top of one of the three points a, b, and c marked on the table.



We will refer to our operations as *symmetries* of the equilateral triangle. We will also refer to each operation as a *rigid motion,* because, by not distorting the cardboard, it preserves its rigidity. Observe that since we are not allowed to distort the triangle, once we know where the vertices have gone to under our operation, we would immediately know where every other point on the triangle would have gone to. For, if a point P is at a distance x from a vertex A, a distance y from a vertex B and a distance z from the third vertex C, then the image of P must be at a distance x from the image of A, a distance y from the image of B and a distance z from the image of C, and this fixes the location of the image of P. (Actually, more is true: it is sufficient to know where any two vertices have gone to under our operation to know where every point has gone: see Remark 4.5.1 ahead if you are interested. But of course, if you know where two vertices have gone, then you automatically know where the third vertex has gone.) Hence, it is enough to study the possible rearrangements, or *permutations*, of the vertices of the triangle to determine our operations. A key sticking point is that while every symmetry of the triangle corresponds to a permutation of the vertices, it is conceivable that not every permutation of the vertices comes from a symmetry. As it turns out, this is not the case, as we will see below.

Let us, for example, write $\binom{a\ b\ c}{b\ c\ a}$ for the permutation of the vertices that takes whichever vertex that was on the point on the table marked a and moves it to the point marked b, whichever vertex that was on the point on the table marked b and moves it to the point marked c, and whichever vertex that was on the point on the table marked c and moves it to the point marked a. Notice that since there are three vertices, there are only six permutations to consider. With this notation, let us consider each of the six permutations in turn, and show that they can be realized as a symmetry of the triangle:

1. $id = \binom{a\ b\ c}{a\ b\ c}$. This of course corresponds to *doing nothing* to the triangle. This is a valid operation of the sort that we are seeking: it is clearly a rigid motion of the triangle (there is no distortion of the cardboard), and after we have performed this operation, we would not be able to tell whether anybody has disturbed the triangle or not!

2. $\sigma = \binom{a\ b\ c}{b\ c\ a}$. This can be realized by rotating the triangle counter-clockwise by 120°. This is a rigid motion (there is no stretching or other distortion involved), and of course, after the rotation is over, we would not be able to tell if the cardboard has been moved.

3. $\sigma^2 = \binom{a\ b\ c}{c\ a\ b}$. This can be realized as a counter-clockwise rotation by 240°, or what is the same thing, a clockwise rotation by 120°. Notice that if were to form the composition $\sigma \circ \sigma$, we would arrive at this permutation, and it is for this reason that we have denoted this permutation by σ^2.

4. $\tau_a = \binom{a\ b\ c}{a\ c\ b}$. This can be realized by flipping the triangle about the line joining the point a and the midpoint of the opposite side bc. This too is a rigid motion, and after the flip is over, we would not be able to tell if the cardboard has been moved.

5. $\tau_b = \binom{a\ b\ c}{c\ b\ a}$. This can be realized by flipping the triangle about the line joining the point b and the midpoint of the opposite side ac. Like τ_a, this too is a rigid motion, and after the flip is over, we would not be able to tell if the cardboard has been moved.

6. $\tau_c = \binom{a\ b\ c}{b\ a\ c}$. This is just like τ_a and τ_b, and can be realized by flipping the triangle about the line joining the point c and the midpoint of the opposite side ab.

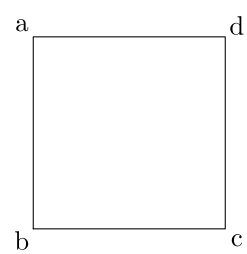
Thus, we have obtained all six permutations as symmetries of the triangle! Notice that these six symmetries compose as follows:

$\circ$	id	σ	σ^2	τ_a	τ_b	τ_c
id	id	σ	σ^2	τ_a	τ_b	τ_c
σ	σ	σ^2	id	τ_c	τ_a	τ_b
σ^2	σ^2	id	σ	τ_b	τ_c	τ_a
τ_a	τ_a	τ_b	τ_c	id	σ	σ^2
τ_b	τ_b	τ_c	τ_a	σ^2	id	σ
τ_c	τ_c	τ_a	τ_b	σ	σ^2	id

Notice that we get a group: the composition of any two symmetries is a symmetry, composition is associative since this is always true for composition of functions, the element id acts as the identity, and it is clear from the relations $\sigma\sigma^2 = \sigma^2\sigma = id$, $\tau_a\tau_a = \tau_b\tau_b = \tau_c\tau_c = id$ that every element has an inverse. This group is called the *dihedral group of index* 3 and is denoted D_3. (Notice the similarity between this group and the group S_3 of Example 4.2. We will take this up again when we consider isomorphisms later in this chapter.)

See Example 4.89 in the notes at the end of the chapter, where D_3 is interpreted as the group of symmetries of the equilateral triangle with the rigid structure.

Example 4.5. This example is similar in spirit to the previous example. We consider a piece of cardboard in the shape of a square. We wish to determine all operations we can perform on the piece of cardboard that do not shrink, stretch, or in anyway distort the square, but are such that after we perform the operation, nobody can tell that we did anything to the square! To help determine what such operations could be, pretend that the piece of cardboard has been placed at a fixed location on a table, and the location has been marked by lines drawn under the edges of the cardboard. Also, label the points on the table that lie directly under the vertices of the square as a, b, c, and d respectively. After we have done our (yet to be determined!) operation on the cardboard, the square should stay at the same location–otherwise it would be obvious that somebody has done something to the piece of cardboard.



We will refer to our operations as *symmetries* of the square and will refer to each operation as a *rigid motion.* Just as in the previous example, each vertex of the square must somehow end up once again on top of one of the four points a, b, c, and d marked on the table after the application of a symmetry. As before, the preservation of the rigidity of the square ensures that once we know where the vertices have gone to under the application of a symmetry, we would immediately know where every other point on the square would have gone to. (In fact, it is enough to know where two adjacent vertices have gone—see Remark 4.5.1 ahead.) Hence, it is enough to study the possible permutations of the vertices of the square to determine its symmetries. Unlike the previous example, however, it is not true that every permutation of the vertices comes from a symmetry.

As before, we write, for example, $\begin{pmatrix} a & b & c & d \\ b & c & d & a \end{pmatrix}$ for the permutation of the vertices that takes whichever vertex that was on the point on the table marked a and moves it to the point marked b, the vertex on b to c, the vertex on c to d, and the vertex on d to a. Notice that since there are four vertices, there are $4! = 24$ permutations to consider (see Exercise 4.3.3 above). With this notation, let us see which of these 24 permutations can be realized as a symmetry of the square:

1. $id = \begin{pmatrix} a & b & c & d \\ a & b & c & d \end{pmatrix}$. This of course corresponds to doing nothing to the square. As with the operation of the previous example that does nothing on the equilateral triangle, this operation on the square is a rigid motion of the square, and after we have performed this operation, we would not be able to tell whether anybody has disturbed the square or not.

2. $\sigma = \begin{pmatrix} a & b & c & d \\ b & c & d & a \end{pmatrix}$. This can be effected by rotating the square counter-clockwise by $90°$. This too is a rigid motion, and after the rotation is over, we would not be able to tell if the cardboard has been moved.

3. $\sigma^2 = \begin{pmatrix} a & b & c & d \\ c & d & a & b \end{pmatrix}$. This is effected by rotating the square counter-clockwise (or clockwise) by $180°$, and corresponds to the composition $\sigma\circ\sigma$. Hence the name σ^2 for this symmetry.

4. $\sigma^3 = \begin{pmatrix} a & b & c & d \\ d & a & b & c \end{pmatrix}$. This is effected by rotating the square counter-clockwise by $270°$ (or clockwise by $90°$), and corresponds to the composition $\sigma \circ \sigma \circ \sigma$.

5. $\tau_H = \begin{pmatrix} a & b & c & d \\ b & a & d & c \end{pmatrix}$. This corresponds to flipping the square about its horizontal axis (i.e., the line joining the midpoints of the sides ab and cd).

6. $\tau_V = \binom{a\ b\ c\ d}{d\ c\ b\ a}$. This corresponds to flipping the square about its vertical axis (i.e., the line joining the midpoints of the sides ad and bc).

7. $\tau_{ac} = \binom{a\ b\ c\ d}{a\ d\ c\ b}$. This corresponds to flipping the square about the top-left to bottom-right diagonal, i.e., the line joining the points a and c.

8. $\tau_{bd} = \binom{a\ b\ c\ d}{c\ b\ a\ d}$. This corresponds to flipping the square about the bottom-left to top-right diagonal, i.e., the line joining the points b and d.

One can check that the remaining permutations of the vertices cannot be realized by rigid motions. For instance, consider the permutation $\binom{a\ b\ c\ d}{b\ c\ a\ d}$. If, for example, vertex A of the square lies on the point a and vertex D of the square lies on d, then this permutation fixes D but moves the vertex A so that it lies on b. The segment AD therefore is converted from a diagonal of the square to a side of the square, and is hence shortened–this is clearly not a rigid motion!

Dispensing off all remaining permutations similarly, we find that our square with its structure of being a rigid object lying at the given location has just these eight symmetries above.

Remark 4.5.1. Here is another way of seeing that there are only eight symmetries: Observe that once we know where a pair of adjacent vertices have gone under the application of a symmetry, we immediately know where every other point on the square has gone, because of the rigidity. This is because if a point P of the square is at a distance x from a vertex A and a distance y from the adjacent vertex B, then the image of P must be at a distance x from the image of A and a distance y from the image of B. There is a unique point on or inside the four lines drawn on the table that satisfies this property, and this point will be the image of P.

Now consider a pair of adjacent vertices A and B. After the application of a symmetry, A can end up in one of four possible locations marked a,

b, c, or d that correspond to the four vertices of the square. Moreover, for the cardboard to not get distorted, B must end up at one of these locations that is adjacent to A. Hence, once a symmetry has placed A in one of four locations, there are only two possible locations where the symmetry could place B: either adjacent to A in the clockwise direction or adjacent to A in the counter-clockwise direction. Since the symmetries of the square are determined by where they send the adjacent vertices A and B, we find that there are $4 \cdot 2 = 8$ potential symmetries. On the other hand, we have explicitly exhibited eight distinct symmetries already. Hence the set of symmetries of the square consists precisely of these eight symmetries above.

It is a fun exercise for you to prove that these symmetries form a group (see below). Notice that this group is noncommutative: $\sigma\tau_H = \tau_H\sigma^3$ for example.

> **Exercise 4.5.1.** Create a table that shows how these symmetries compose and argue, as in Example 4.4 above, why this table shows that the set of symmetries forms a group.

This group is called the *dihedral group of index* 4, and is denoted D_4.

> **Exercise 4.5.2.** Use the group table of D_4 to show that every element of D_4 can be expressed as $\sigma^i\tau_H^j$ for *uniquely determined integers* $i \in \{0, 1, 2, 3\}$ and $j \in \{0, 1\}$.

> **Definition 4.5.1.** The *center* of a group is defined to be the set of all elements in the group that commute with all other elements in the group. (For instance, the identity element is always in the center of a group as it commutes with all other elements.)

> **Exercise 4.5.3.** Determine the elements in D_4 that lie in its center.

4.1.3 Cyclic groups

Example 4.6. Notice that the subset $\{1, -1\}$ of $\mathbb{Z}$ endowed with the usual multiplication operation of the integers is a group!

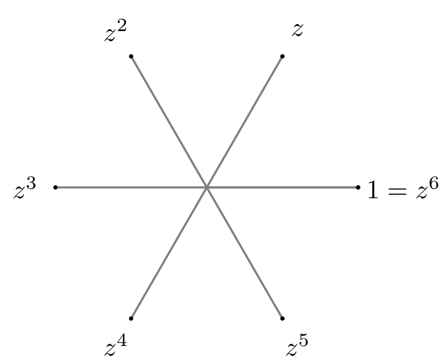
> **Question 4.6.1.** What similarities do you see between this group and the group $(\mathbb{Z}/2\mathbb{Z}, +)$?

Question 4.6.2. Let G be any group that has exactly two elements. Can you see that G must be similar to the group $(\mathbb{Z}/2\mathbb{Z}, +)$ in exactly the same way that this group $\{1, -1\}$ is similar to $(\mathbb{Z}/2\mathbb{Z}, +)$? Now that you have seen the notion of isomorphism in the context of rings and vector spaces, can you formulate precisely how any group with exactly two elements must be similar to $(\mathbb{Z}/2\mathbb{Z}, +)$?

We will now generalize Example 4.6.

Example 4.7. Let $n \geq 3$ be any integer. Recall that the complex number $z_n = \cos(2\pi/n) + \imath\sin(2\pi/n)$ has modulus 1, and is at an angle $\theta_n = 2\pi/n$ with respect to the positive real axis. DeMoivre's theorem ($(\cos(\theta) + \imath\sin(\theta))^k = \cos(k\theta) + \imath\sin(k\theta)$ for $k = 1, 2, \dots$) shows that the complex number $z_n^2 = \cos(4\pi/n) + \imath\sin(4\pi/n)$. This also has modulus 1, but is now at an angle $2\theta_n = 4\pi/n$. Proceeding, we find that the complex numbers $1, z_n, z_n^2, z_n^3, \dots, z_n^{n-1}$ are evenly spaced around the unit circle, and z_n^n gives you back the complex number 1.

The elements z_6^i are shown below:



Cyclic group of order 6

It is easy to see that the set $C_n = \{1, z_n, z_n^2, \dots, z_n^{n-1}\}$ is a group; it is known as *the cyclic group of order* n. (See Lemma 4.29.1 to note that $C_n = \langle z_n \rangle$, with notation as in that lemma. Thus, in the language of Definition

4.29.1 ahead, C_n is the group generated by z_n. See also Remark 4.32 ahead as well as Exercise 4.72.1.)

Question 4.7.1. If $z_n^i z_n^j = z_n^k$ for some k with $0 \le k < n$, what is k in terms of i and j?

Question 4.7.2. If $(z_n^i)^{-1} = z_n^j$ for some j with $0 \le j < n$, what is j in terms of i?

Question 4.7.3. Consider the group $(\mathbb{Z}/n\mathbb{Z}, +)$, for a fixed integer $n \ge 1$. Notice that every element in this group is obtained by adding $[1]_n$ to itself various number of times. For instance, $[2]_n = [1]_n + [1]_n$ (which we write as $2 \cdot [1]_n$), $[3]_n = [1]_n + [1]_n + [1]_n$ (which we write as $3 \cdot [1]_n$), etc. What similarities do you see between $(\mathbb{Z}/n\mathbb{Z}, +)$ and the group C_n above? Now that you have seen the notion of isomorphism in the context of rings and of vector spaces, can you formulate precisely how $(\mathbb{Z}/n\mathbb{Z}, +)$ and C_n are similar?

4.1.4 Direct product of groups

Example 4.8. Let G and H be groups. We endow the cartesian product $G \times H$ with the operation $(g_1, h_1)(g_2, h_2) = (g_1 g_2, h_1 h_2)$ (compare with Example 2.22 in Chapter 2). Here, the product $g_1 g_2$ refers to the operation in G, while the product $h_1 h_2$ refers to the operation in H.

Exercise 4.8.1. Verify that with this definition of operation, the set $G \times H$ forms a group.

This is known as the *direct product of G and H*.

Question 4.8.1. What is the identity element in $G \times H$? What is the inverse of an element (g, h)?

Question 4.8.2. If G and H are abelian, must $G \times H$ necessarily be abelian? If $G \times H$ is not abelian, can G or H be abelian? Can both G and H be abelian?

Exercise 4.8.2. Consider the direct product $(\mathbb{Z}/2\mathbb{Z}, +) \times (\mathbb{Z}/3\mathbb{Z}, +)$. Show by direct computation that every element of this group is a multiple of the element $([1]_2, [1]_3)$. What similarities do you see between this group and $(\mathbb{Z}/6\mathbb{Z}, +)$? With your experience with isomorphisms in the context of rings and vector spaces, can you formulate precisely how $(\mathbb{Z}/2\mathbb{Z}, +) \times (\mathbb{Z}/3\mathbb{Z}, +)$ and $(\mathbb{Z}/6\mathbb{Z}, +)$ are similar?

4.1.5 Matrix groups

Example 4.9. We know (see Exercise 2.115 in Chapter 2) that the set of invertible elements of a ring R, denoted R^*, forms a group under the multiplication operation in the ring. In particular, taking R to be $M_n(F)$ for a fixed field F, we find that the set of $n \times n$ *invertible* matrices with entries in F forms a group with respect to matrix multiplication. This is a very important group in mathematics, and has its own notation and its own name: it is denoted by $Gl_n(F)$ and is called the *general linear group of order n over F*. Recall (see the parenthetical remarks in Exercise 2.56.1 in Chapter 2) that a matrix with entries in a field is invertible if and only if its determinant is nonzero. Thus, $Gl_n(F)$ may be thought of as the group of all $n \times n$ matrices with entries in F whose determinant is nonzero.

> **Exercise 4.9.1.** Write down the group table for the group of units of the ring $Gl_2(\mathbb{Z}/2\mathbb{Z})$ (see Exercise 2.56.1 in Chapter 2). What familiar group is this isomorphic to?

Remark 4.9.1. Recall from Exercise 3.104 in Chapter 3 that if V is an n-dimensional vector space over F, then the ring of F-linear transformation from V to V, namely $End_F(V)$—see Exercise 3.101 of that same chapter as well—is isomorphic to the ring $M_n(F)$. As in Exercise 3.104, let $M : End_F(V) \to M_n(F)$ be the function that provides this isomorphism (M depends on a choice of basis for V, but that is not important at this point). It follows at once that the invertible elements of $End_F(V)$ will correspond to the invertible elements of $M_n(F)$ bijectively under M. Since the invertible elements of $M_n(F)$ are what we have denoted as $Gl_n(F)$ above, and since the invertible elements of $End_F(V)$ are just those linear transformations that are both injective and surjective (see Exercise 3.102 in Chapter 3), we find that elements of $Gl_n(F)$ correspond to the injective and surjective linear transformations of V. But more, if the matrix M_T corresponds to the linear transform T and the matrix M_S to the linear transform S under this isomorphism, then, since the map M "preserves multiplication," we find that the product matrix $M_T M_S$ corresponds to $T \circ S$. Although we have not studied the notion of isomorphisms of groups–we will in Section 4.4 ahead)–we have enough experience with isomorphisms in the context of rings and vector spaces by now to realize that $Gl_n(F)$ and the group of units of $End_F(V)$ are isomorphic as groups.

Many interesting groups arise as subgroups of $Gl_n(F)$ for suitable n and F. (Of course, we have not formally defined the term "subgroup" yet–we will in Section 4.2 ahead—but we have enough experience with subrings and subspaces already to know what that term should mean: a subset of a group G that is closed with respect to the operation in G and forms a group on its own under this operation.) In fact, every finite group that has n elements in it occurs as a subgroup of $Gl_n(F)$, for every field F. We consider a few examples of subgroups of $Gl_n(F)$ ahead. The identity element in all such groups will be the $n \times n$ identity matrix. Moreover, the multiplication in such groups will necessarily be associative, since matrix multiplication is known to be an associative operation (see Question 2.16.6 in Chapter 2).

Example 4.10. The set of matrices in $Gl_n(F)$ whose determinant is 1 forms a group under matrix multiplication, denoted $Sl_n(F)$, and called the *special linear group of order n over F*.

> **Question 4.10.1.** Why does this subset form a group? Check that the axioms hold.

Example 4.11. Let $B_2(\mathbb{R})$ be the set of matrices in $M_2(\mathbb{R})$ of the form $g = \begin{pmatrix} a & b \\ 0 & d \end{pmatrix}$ where $ad \neq 0$. Since the determinant of g is precisely ad, the condition $ad \neq 0$ shows that g is invertible, i.e., $g \in Gl_2(\mathbb{R})$. $B_2(\mathbb{R})$ is a group with respect to matrix multiplication.

> **Question 4.11.1.** What does the product of two such matrices g and h above look like?
>
> **Question 4.11.2.** What does the inverse of a matrix such as g above look like?

More generally, consider the subset $B_n(F)$ of upper triangular matrices in $M_n(F)$ whose product of diagonal entries is nonzero. Since the determinant of an upper triangular matrix is just the product of its diagonal entries, B_n is a subset of $Gl_n(F)$. $B_n(F)$ forms a group with respect to matrix multiplication It is referred to as the *upper triangular subgroup of* $Gl_n(F)$ or the *standard Borel subgroup of* $Gl_n(F)$.

> **Exercise 4.11.1.** Prove that the determinant of an upper triangular $n \times n$ matrix (with, say, entries in a field F), is just the product of its diagonal entries.

> **Exercise 4.11.2.** Show that the product of two upper triangular matrices is also upper triangular.

> **Exercise 4.11.3.** Show that the inverse of an invertible upper triangular matrix is also upper triangular.

Example 4.12. Let $U_2(\mathbb{R})$ be the subset of matrices in $M_2(\mathbb{R})$ of the form $g_a = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}$, where $a \in \mathbb{R}$. Note that the determinant of $g_a = 1$ for all a. $U_2(\mathbb{R})$ is a group with respect to matrix multiplication.

> **Question 4.12.1.** What is the product of g_a and g_b in terms of a and b?

> **Question 4.12.2.** What is the multiplicative inverse of g_a?

> **Question 4.12.3.** What similarity do you see between $U_2(\mathbb{R})$ and $(\mathbb{R}, +)$?

> **Question 4.12.4.** View the elements of $U_2(\mathbb{R})$ as the matrix of linear transformations of $\mathbb{R}^2$ with respect to the usual basis $\mathbf{i}$ and $\mathbf{j}$. Where do $\mathbf{i}$ and $\mathbf{j}$ go to under the action of g_a?

> **Question 4.12.5.** How would any of these calculations above in Questions (4.12.1) or (4.12.2) or (4.12.4) be changed if we had restricted a to be an integer? What similarity would you then have seen between this modified set (with a now restricted to be an integer) and $(\mathbb{Z}, +)$?

See Exercise 4.79 at the end of the chapter for a generalization to $n \times n$ matrices.

Example 4.13. As we will see in Exercise 4.78, the set of 2×2 matrices with entries in $\mathbb{R}$ satisfying $A^tA = I$, where I is the identity matrix and A^t stands for the transpose of A, forms a group: it is precisely the group of symmetries of the set described in Example 4.92 on Page 208. This set of matrices is indeed a subset of $Gl_2(\mathbb{R})$, since, as you are asked to prove as well in Exercise 4.78, the relation $A^tA = I$ yields that the determinant of A is ± 1, and in particular, nonzero. More generally, one can consider the set of $n \times n$ matrices A with entries in any field F satisfying $A^tA = I$. This set will form a group, known as the *orthogonal group of order n over F*. We will use the notation $O_n(F)$ for such groups.

See Remark 4.5 at the end of the chapter as well about more general orthogonal groups than the one we have introduced above.

> **Exercise 4.13.1.** Prove that the set of $n \times n$ matrices A with entries in G satisfying $A^tA = I$ forms a group under matrix multiplication. (Hint: you will need to show that $A^tA = I$ is equivalent to $AA^t = I$, and from this, that $(A^{-1})^t = (A^t)^{-1}$.)

4.1.6 Remarks and some general properties of groups

Remark 4.14. In an abstract group, several different symbols are used to denote the binary operation, the identity element, and the inverse of an element. Sometimes, one uses the symbol *id* for the identity element, as we have done above for the groups S_3, D_3, etc. Sometimes the symbol e is used for the identity element. Very often, one imagines the group operation to be some sort of "multiplication" between group elements (*Warning:* this is just an informal way to think about the operation—in general, the operation may not represent any sort of actual multiplication in the sense of multiplication in rings), and in such cases, one uses the familiar symbol 1 to represent the identity element. (In such a situation we say that the group is *written in multiplicative notation*, or *written multiplicatively*.) When writing the group multiplicatively, one simply writes the binary operation without any symbol, thus, the "product" of two elements a and b is simply written ab.

(We have followed this convention already with S_3, for example.) In the case where the group is abelian, one often imagines the group operation as some sort of "addition" in analogy with the addition operation in rings (*Same Warning:* this is just an informal way to think about the operation), and one writes + for the group operation and 0 for the identity. And continuing with the analogy, one writes $-a$ for the inverse of an element a. (In such a situation, we say that the group is *written in additive notation*, or *written additively*.)

Before proceeding further, here are some exercises that would be useful. You would have encountered many of these results already in the context of the additive group of a ring or a vector space. (For instance, see Remark 2.24, Exercise 2.114, and the notes on page 89 in Chapter 2):

Exercise 4.15. Show that the identity element in a group is unique.

Exercise 4.16. Show that the inverse of an element in a group is unique.

Exercise 4.17. Show that for any element a in a group G, $(a^{-1})^{-1} = a$.

Exercise 4.18. (Cancellation in Groups) If $ab = ac$ for elements a, b, and c in a group G, show that $b = c$ (left cancellation). Similarly, if $ba = ca$, show that $b = c$ (right cancellation).

Exercise 4.19. Let G be a group, and let a and b be elements in G. Show that $(ab)^{-1} = b^{-1}a^{-1}$.

Exercise 4.20. Let G be a group written multiplicatively. For any element $a \in G$ and for any positive integer j, it is customary to write a^j for $\underbrace{a \cdot a \cdots \cdot a}_{j\text{times}}$. Similarly, for any negative integer j, it is customary to write a^j for $\underbrace{a^{-1} \cdot a^{-1} \cdots \cdot a^{-1}}_{|j|\text{times}}$. Finally, it is customary to take a^0 to be 1. Prove the following:

1. If $y = a^j$ for some integer j, then $y^{-1} = a^{-j}$. (Hint: Compute $a^j a^{-j}$ by invoking the definition of a^j and a^{-j}—you would of course have to divide your proof into whether j is positive, negative, or zero.
2. For integers s and t, prove that $a^s a^t = a^{s+t}$. (Hint: First dispose of the case where either s or t is zero, and then divide your proof into four cases according to whether s is positive or negative and t is positive or negative.)

4.2 Subgroups, Cosets, Lagrange's Theorem

After our practice with subrings and subspaces, the following concept must now be quite intuitive:

Definition 4.21. Let G be a group. A *subgroup* of G is a subset H that is closed with respect to the binary operation such that with respect to this operation, H is itself a group.

Exercise 4.22. Let G be a group and let H be a subgroup. Prove that the identity element of H must be the same as the identity element of G. (Hint: Write id_G and id_H for the respective identities. Then $id_H id_H = id_H$. Also, $id_G id_H = id_H$. So?)

The following lemma allows us to check if a nonempty subset of a group is a subgroup.

Lemma 4.23. *(Subgroup Test) Let G be a group, and let H be a nonempty subset. If for all a and b in H the product ab^{-1} is also in H, then H is a subgroup of G.*

Proof. Since H is nonempty (note that we are invoking the nonemptiness hypothesis!), H has at least one element in it, call it a. Then, taking $b = a$ in the the statement of the lemma, we find $aa^{-1} = e \in H$. Thus, H contains the identity. Next, given any $x \in H$, we take $a = e$ and $b = x$ in the statement of the lemma to find that the product $ex^{-1} = x^{-1}$ must be in H, so H contains inverses of all its elements. Finally, given any x and y in H, note that y^{-1} must also be in H by what we just saw, so, taking $a = x$ and $b = y^{-1}$, we find $x(y^{-1})^{-1} = xy$ must be in H. Putting all this together, we find that H is closed with respect to the group operation, contains the identity, and contains inverses of all its elements. Since the associativity of the group operation is simply inherited from the fact that the operation is associative on all of G, we find that H satisfies all group axioms, and hence, H is a subgroup of G.

□

Example 4.24. Let G be a group. The subset $\{1_G\}$ is a subgroup, called the *trivial subgroup*.

> **Exercise 4.24.1.** Prove this by applying the subgroup test (Lemma 4.23).

Example 4.25. In the group S_3, the subset $\{id, r_1, r_2\}$ is a subgroup, as are the subsets $\{id, f_1\}$, $\{id, f_2\}$, and $\{id, f_3\}$.

> **Exercise 4.25.1.** Prove these assertions by studying the group table of S_3 on page 160.

Example 4.26. In the group S_n of permutations of $\{1, 2, \ldots, n\}$, let H be the subset consisting of all permutations that act as the identity on n.

> **Exercise 4.26.1.** Prove that H is a subgroup of S_n using the subgroup test (Lemma 4.23).

> **Exercise 4.26.2.** Compare H with S_{n-1}. What similarities do you see?

Example 4.27. The various matrix groups we considered above such as $Sl_n(F)$, $B_n(F)$, $O_n(F)$, etc., are all subgroups of $Gl_n(F)$.

Example 4.28. Let G be a group. Recall that we have defined the *center* of G (see Definition 4.5.1) to be the subset consisting of all elements of G that commute with every other element of G.

> **Exercise 4.28.1.** Prove that the center of G is a subgroup of G.
>
> **Question 4.28.1.** What can you say about the center of G when G is abelian?

4.2.1 Subgroup generated by an element

Example 4.29. Let G be a group, and let a be an element in G. What would be the *smallest* subgroup of G that contains a? (By smallest, we mean smallest with respect to set theoretic inclusion, that is, we seek a subgroup H of G that contains a such that if K is any other subgroup of G that contains a, then $H \subseteq K$.) Let us write G multiplicatively. Then, *any* subgroup H that contains a must contain, along with a, the elements $a \cdot a = a^2$, $a \cdot a^2 = a^3$, $\dots$, because the subgroup must be closed with respect to the group operation. It must contain the identity 1 $(= a^0)$ since it is a subgroup. Similarly, it must contain the inverse a^{-1}, and then, it must contain all products $a^{-1} \cdot a^{-1} = a^{-2}$, $a^{-1} \cdot a^{-2} = a^{-3}$, $\dots$. We have the following:

Lemma 4.29.1. *The set $\langle a \rangle = \{a^n \mid n \in \mathbb{Z}\}$ is a subgroup of G. It is the smallest subgroup of G that contains a, in the sense that if H is any subgroup of G that contains a, then $\langle a \rangle \subseteq H$.*

Proof. The discussions just before the statement of this lemma show that if H is any subgroup of G that contains a, then H must contain all a^i, for $i \in \mathbb{Z}$, that is, H must contain $\langle a \rangle$. Thus, we only need to show that $\langle a \rangle$ is a subgroup of G. But this is easy by the subgroup test (Lemma 4.23): $\langle a \rangle$ is nonempty since a is in there. Given any two elements x and y in $\langle a \rangle$, $x = a^i$

for some $i \in \mathbb{Z}$, and $y = a^j$ for some $j \in \mathbb{Z}$. Note that $y^{-1} = a^{-j}$ (Exercise 4.20). Then $xy^{-1} = a^i a^{-j}$. Hence (Exercise 4.20 again), $xy^{-1} = a^{i-j} \in \langle a \rangle$, proving that $\langle a \rangle$ is indeed a subgroup. □

Before proceeding further, we pause to give a name to the object considered in the lemma above:

> **Definition 4.29.1.** Let G be a group, and let a be an element in G. The subgroup $\langle a \rangle$ is called the *subgroup generated by* a. A subgroup H of G is called *cyclic* if $H = \langle g \rangle$ for some $g \in G$. In particular, G itself is called cyclic if $G = \langle g \rangle$ for some $g \in G$.

> **Exercise 4.29.1.** Let G be a group written multiplicatively, and let $a \in G$. For integers s and t, prove that $a^s a^t = a^{s+t}$ by mimicking the proof that $(a^j)^{-1} = a^{-j}$ in the lemma above. (Hint: First dispose of the case where either s or t is zero, and then divide your proof into four cases according to whether s is positive or negative and t is positive or negative.)

In S_3, for instance, we see that $\langle r_1 \rangle$ is the (finite) set $\{id, r_1, r_1^2 = r_2\}$. This is because we need no further powers: $r_1^3 = id$, so $r_1^4 = r_1^7 = \cdots = r_1$, and $r_1^5 = r_1^8 = \cdots = r_1^2 = r_2$. Similarly, $r_1^{-1} = r_2$, so $r_1^{-2} = r_1^{-1} r_1^{-1} = r_2^2 = r_1$, and from this, we see that all powers r_1^{-n} $(n = 1, 2, \ldots)$ is one of id, r_1, or r_2.

By contrast, the subgroup $\langle 1 \rangle$ of the additive group $(\mathbb{Z}, +)$ is all of $\mathbb{Z}$. This is easy to see: $\langle 1 \rangle = \{1, 1+1 = 2, 1+2 = 3, \ldots, 0, -1, (-1)+(-1) = -2, (-1)+(-2) = -3, \ldots\}$.

These examples suggest an interesting and important concept:

> **Definition 4.29.2.** Let G be a group (written multiplicatively), and let $a \in G$. The *order* of a (written $o(a)$)is the least positive integer n (if it exists) such that $a^n = 1$. If no such integer exists, we say that a has infinite order.

We now have the following:

Lemma 4.29.2. *Let G be a group and let $a \in G$. Then $o(a)$ is finite if and only if $\langle a \rangle$ is a finite set. When these (equivalent) conditions hold,*

$o(a)$ equals the number of elements in the subgroup $\langle a\rangle$, and if this common integer is m, then the elements $1, a, \ldots, a^{m-1}$, are all distinct, and $\langle a\rangle = \{1, a, \ldots, a^{m-1}\}$.

Proof. Assume that $o(a)$ is finite, say m. Then, any integer l can be written as $bm + q$ for $0 \le q < m$, so $a^l = a^{bm+q} = (a^m)^b a^q = 1 \cdot a^q$. Hence, every power of a can be written as a^q for some q between 0 and $m-1$, that is, $\langle a\rangle = \{1, a, \ldots, a^{m-1}\}$. This shows that $\langle a\rangle$ is a finite set.

Now assume that $\langle a\rangle$ is a finite set. Then, the powers $1, a, a^2, \ldots$ cannot all be distinct (otherwise $\langle a\rangle$ would be infinite), so there exist nonnegative integers k and l, with $k < l$, such that $a^k = a^l$. Multiplying by $a^{-k} = (a^k)^{-1}$, we find $1 = a^{l-k}$. Note that $l-k$ is positive. Thus, the set of positive integers t such that $a^t = 1$ is nonempty, since $l-k$ is in this set. By the well-ordering principle, there is a *least positive integer* s such that $a^s = 1$. This shows that a has finite order, namely s.

Now assume that these equivalent conditions hold. We have already seen above that if $o(a)$ is finite and equal to some m, then $\langle a\rangle = \{1, a, \ldots, a^{m-1}\}$. Note that these elements are all distinct, since if $a^j = a^k$ for $0 \le j < k \le m-1$, then, multiplying both sides by $a^{-j} = (a^j)^{-1}$, we would find $1 = a^{k-j}$, and since $0 < k - j < m$, this would contradict the fact that m is the *least* positive integer l such that $a^l = 1$. It follows that the number of elements in $\langle a\rangle$ is precisely m, the order of a. $\square$

The following result is useful, its proof uses an idea that we have already encountered in the proof of Lemma 4.29.2 above:

Lemma 4.29.3. *Let a be an element of a group G and suppose that $a^l = 1$ for some integer l. Then the order of a divides l. (In particular, the order of a is finite.)*

Proof. Note that if l is negative, then $a^{-l} = (a^l)^{-1} = 1$. Hence, the set of positive integers n such that $g^n = 1$ is nonempty, since either l or $-l$ is in that set. By the Well-Ordering principle, this set has a least element, so indeed the order of g is finite.

Now suppose the order of a is m. Write $l = bm + r$ for integers b and r with $0 \leq r < m$. Then $a^r = a^l a^{-bm} = a^l (a^m)^{-b} = 1$, because both a^l and a^m equal 1. Since m is the least positive integer n such that $a^n = 1$, it follows that $r = 0$, i.e., that m divides l. □

We now prove a result that determines the order of a^d in terms of the order of a, in the case where the order of a is finite:

Lemma 4.29.4. *Let G be a group and let $a \in G$ have finite order m. Then, for any integer d, the order of a^d equals $m/gcd(m, d)$.*

Proof. Let us assume first that d is positive. Denote the order of a^d by t. Thus, the integer t will correspond to the first time that 1 occurs in the list a^d, a^{2d}, a^{3d}, ... By Lemma 4.29.3 above, the integer t will correspond to the first time m divides a member of the list d, $2d$, $3d$, ... This member will then be a common multiple of d and m, and since this is the first common multiple in the list, it will the least common multiple of d and m. In other words, t will be such that dt will be the least common multiple of m and d. Since $dm = \gcd(m, d)\mathrm{lcm}(m, d) = \gcd(m, d)dt$, we find $t = m/\gcd(m, d)$, as desired.

If d is zero or negative, choose a positive integer p so that $pm + d \geq 0$. Now observe that $a^{pm+d} = (a^m)^p a^d = 1^p a^d = a^d$. Hence, the order of a^d is the same as the order of a^{pm+d}, and since $pm+d > 0$, we may apply the result of the last paragraph to find that the order of $a^d = m/\gcd(m, pm + d)$. To finish the proof, we will show that $\gcd(m, d) = \gcd(m, pm + d)$. It is enough to show that the set of common divisors of the two pairs of integers are the same. But this is easy: if l divides both m and d, then it must also divide the linear combination $pm + d$ by Lemma 1.2, so l is a common divisor of m and $pm + d$. Thus the set of common divisors of m and d is a subset of the common divisors of m and $pm + d$. On the other hand, if l divides both m

and $pm+d$, then l must divide the linear combination $(-p)m+pm+d=d$, so l is a common divisor of m and d. Thus, we have the reverse inclusion: the set of common divisors of m and $pm+d$ is a subset of the common divisors of m and d. The two sets are hence equal. □

We have the following immediately:

Corollary 4.29.1. *With a and m as above,*

1. *If d divides m, then a^d has order m/d.*
2. *If d is relatively prime to m, then a^d has order m, and $\langle a^d \rangle = \langle a \rangle$.*

Proof. It follows directly from the lemma that if d divides m then a^d has order m/d, and that if d and m are relatively prime, then a^d has order m. To see that $\langle a^d \rangle = \langle a \rangle$ in the case where d and m are relatively prime, note that by Lemma 4.29.2, the subgroup $\langle a \rangle$ has m elements since a has order m, and likewise, the subgroup $\langle a^d \rangle$ has m elements since a^d has order m. But $\langle a^d \rangle$ is a subset of $\langle a \rangle$, since any power of a^d is also a power of a. Since a subset T of a finite set S that has the same number of elements as S must equal S, we find that $\langle a^d \rangle = \langle a \rangle$. □

Here is a quick exercise to show you that cyclic groups can come in hidden forms!

> **Exercise 4.29.2.** Show that $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$ is cyclic. (Hint: What is the order of the element $([1]_2, [1]_3)$?) After you work this out, see Exercise 4.72.1 ahead and Exercise 4.86 at the end of the chapter.

Now let H be a subgroup of a group G, and assume that the number of elements in G is finite. Then there is a very tight restriction on the possible number of elements in H (Theorem 4.39 ahead), and we will work towards understanding this restriction in the next two subsections. First, a definition:

Definition 4.30. Let G be a group. The *order* of G (written $o(G)$) is the number of elements in G, if this number is finite. If the number of elements in G is infinite, we say G is of infinite order.

Remark 4.31. Do not confuse the order of an element a in a group G with the order of the group G, these refer to two separate concepts. (All the same, even though these are separate concepts, we will see (Corollary 4.40 ahead) that the two integers are related.) Note that Lemma 4.29.2 above says that the order of a equals the order of the subgroup generated by a. Thus in the special case when G is cyclic, i.e., when $G = \langle a \rangle$ for some $a \in G$ (See Definition 4.29.1 above), the order of a and the order of the group $G = \langle a \rangle$ are indeed the same integers, even though they arise out of different concepts.

Remark 4.32. Continuing with the special situation at the end of Remark 4.31 above, let G be any cyclic group of order n. Thus $G = \langle a \rangle$ for some $a \in G$, and since G has order n, Lemma 4.29.2 shows that the element a must have order n, and that $G = \{1, a, \dots, a^{n-1}\}$. Notice the similarity with Example 4.7 above. See also Exercise 4.72.1.

4.2.2 Cosets

We have already seen the notion of the coset of a subgroup with respect to an element before. We saw this in the context of subgroups I of abelian groups of the form $(R, +)$, where R is a ring and I is an ideal (see page 57). We also saw this in the context of subgroups W of abelian groups of the form $(V, +)$, where V is a vector space and W is a subspace (see page 130). The following should therefore come as no surprise, the only novel feature is that we need to distinguish between right and left cosets since the group operation in an arbitrary group need not be commutative:

Definition 4.33. Let G be a group and let H be a subgroup. Given any $a \in G$, the *left coset of H with respect to a* is the set of all elements of the form ah as h varies in H, and is denote aH. Similarly, the *right coset of H*

with respect to a is the set of all elements of the form ha as h varies in H, and is denoted Ha.

Example 4.34. Let us consider an example that will show that indeed left and right cosets can be different. Take G to be S_3 (see Example 4.2), and let $H = \langle f_1 \rangle$. Since the order of f_1 is 2 (see the group table for S_3 on page 160), $\langle f_1 \rangle = \{1, f_1\}$ by Lemma 4.29.2. Take a to be the element r_1. Then the left coset $r_1\langle f_1 \rangle = \{r_1, r_1 f_1\} = \{r_1, f_3\}$ (see the group table), while the right coset $\langle f_1 \rangle r_1 = \{r_1, f_1 r_1\} = \{r_1, f_2\}$. Clearly, the left and right cosets of r_1 with respect to the subgroup $\langle f_1 \rangle$ are not equal!

Continuing with this example, let us make a table of all left and right cosets of $\langle f_1 \rangle$.

a	Left Coset $a\langle f_1 \rangle$	Right Coset $\langle f_1 \rangle a$
id	$\{1, f_1\}$	$\{1, f_1\}$
r_1	$\{r_1, f_3\}$	$\{r_1, f_2\}$
r_2	$\{r_2, f_2\}$	$\{r_2, f_3\}$
f_1	$\{1, f_1\}$	$\{1, f_1\}$
f_2	$\{f_2, r_2\}$	$\{f_2, r_1\}$
f_3	$\{f_3, r_1\}$	$\{f_3, r_2\}$

Notice that every coset (left or right) has exactly two elements, which is the same number as the number of elements in the subgroup $\langle f_1 \rangle$ that we are considering. This will be useful in understanding the proof of Lagrange's theorem (Theorem 4.39) below.

> **Exercise 4.35.** Take $G = S_3$ and take $H = \langle r_1 \rangle$. Write down all left cosets of H and all right cosets of H with respect to all the elements of G. What observation do you make?

The following equivalence relation in Lemma 4.36 below is analogous to the corresponding equivalence relations for rings (see page 57) and vector spaces (see page 130), except that once again, we need to distinguish two cases because the group operation need not be commutative. Note that in the case of rings, for example, we define $a \sim b$ if and only if $a - b \in I$ (where I is some given ideal). Now note that $a - b$ is really $a + (-b)$. Thus, in the group situation, the expression analogous to $a + (-b)$ would be ab^{-1}, and this is indeed the expression we consider in the lemma below. (And while $a + (-b) = (-b) + a$ in the situation of rings, the operation in a group need not be commutative, so we need to consider the expression analogous to $(-b) + a$ as well, which is $b^{-1}a$.)

Lemma 4.36. *Let G be a group and H a subgroup. Define two relations on G, denoted "$\sim_L$" and "$\sim_R$," by the following rules: $a \sim_L b$ if and only if $b^{-1}a \in H$, and $a \sim_R b$ if and only if $ab^{-1} \in H$. Then $\sim_L$ and $\sim_R$ are both equivalence relations on G. The equivalence class $[a]_L$ of an element a with respect to the relation $\sim_L$ is the left coset aH, while its equivalence class $[a]_R$ with respect the relation $\sim_R$ is the right coset Ha*

Proof. The proof that $\sim_L$ is an equivalence relation is similar to the proof of Lemma 2.78 in Chapter 2, except that we have to account for the fact that the group operation need not be commutative.

To show that $a \sim_L a$, simply note that $a^{-1}a = 1 \in H$. To show that $a \sim_L b$ implies that $b \sim_L a$, note that $a \sim_L b$ gives (by definition) $b^{-1}a = h$ for some $h \in H$, and taking inverses of both sides (see Exercise 4.19 above), we find $(b^{-1}a)^{-1} = a^{-1}b = h^{-1}$. Since h^{-1} is also in H as H is a subgroup, we find $a^{-1}b$ is in H, which shows that $b \sim_L a$. Finally, given $a \sim_L b$ and $b \sim_L c$, note that (by definition) $b^{-1}a = h_1$ and $c^{-1}b = h_2$ for some h_1 and h_2 in H. Then $h_2h_1 = c^{-1}b \cdot b^{-1}a = c^{-1}a$, and since h_2h_1 is also in H (as H is a subgroup), we find $a \sim_L c$ as well.

The proof that $\sim_R$ is an equivalence relation is similar.

To prove that $[a]_L = aH$, note that any element b in aH is of the form ah for some $h \in H$. Multiplying by a^{-1}, we find $a^{-1}b = h$ and hence $a^{-1}b \in H$. This shows that $b \sim_L a$. Thus, all elements in aH are in the equivalence class of a, i.e., $aH \subseteq [a]_L$. For the other direction, take any $b \in [a]_L$. Then $b \sim_L a$, so (by definition) $a^{-1}b = h$ for some $h \in H$. Thus, multiplying both sides by a, we find $b = ah$, so $b \in aH$. Hence, $[a]_L \subseteq aH$ as well.

The proof that $[a]_R = Ha$ is similar.

□

Note that we immediately have:

Corollary 4.37. *Any two left cosets aH and bH are either equal or disjoint. Similarly any two right cosets Ha and Hb are either equal or disjoint.*

Proof. This follows from the fact that $aH = [a]_L$ and $bH = [b]_L$, and the fact that any two equivalence classes arising from an equivalence relation are either equal or disjoint. (The proof is identical for right cosets.) □

Here is a quick exercise:

> **Exercise 4.38.** Show that H is itself a left coset, as well as a right coset. Now show that the left coset aH equals H if and only if $a \in H$. Similarly, show that the right coset Hb equals H if and only if $b \in H$. More generally, show that left coset aH equals the left coset bH if and only if $a \in bH$ if and only if $b \in aH$ (and similarly for right cosets).

4.2.3 Lagrange's theorem

Theorem 4.39. *(Lagrange's Theorem) Let G be a group of finite order, and let H be a subgroup. Then the order of H divides the order of G.*

Proof. The crux of the proof is to show that any two left cosets of H have the same number of elements (recall that we have already seen this phenomenon in Example 4.34 above: see the table of left and right cosets in that example). Once we have shown this, it will follow that every coset has $o(H)$ elements in it, since H itself is one of these left cosets (it is the left coset hH for any $h \in H$, for instance, see Exercise 4.38). From this it is trivial to conclude that $o(H)$ must divide $o(G)$: since the left cosets are disjoint and their union is G, and since each left coset has $o(H)$ elements, we would find that $o(H)$

times the number of distinct left cosets of H must equal $o(G)$, i.e., $o(H)$ must divide $o(G)$.

To prove that any two left cosets of H have the same number of elements, take two left cosets aH and bH. Every element of aH can be written as ah for some *unique* $h \in H$. For, by definition, every element of aH is already of the form ah for some $h \in H$. We only have to show that h is unique. But this is clear: if $ah = ah'$, then by cancellation (see Exercise 4.18), h must equal h'. Hence, the following function $f : aH \to bH$ is well defined: take an element in aH, it is expressible as ah for some uniquely determined h, so send this element to bh which lives in the left coset bH. This is a one-to-one function, since if $ah_1 = ah_2$, then cancelling a, we would find $h_1 = h_2$. It is onto as well because every element in bH is of the form bh for some (uniquely determined) element h in H, and hence, $f(ah) = bh$. It follows that the number of elements in aH equals the number of elements in bH. The rest of the proof of the theorem is as described in the first paragraph.

(Note that essentially the same proof shows that any two right cosets of H have the same number of elements.)

□

Here is an immediate corollary containing a result promised in Remark 4.31 above:

Corollary 4.40. *Let G be a group of finite order. Then the order of any element of G divides the order of G.*

Proof. By Lemma 4.29.2, the order of any element a equal the order of the subgroup $\langle a \rangle$ generated by a. But, by the theorem above, $o(\langle a \rangle)$ divides $o(G)$. It follows that $o(a)$ divides $o(G)$.

□

Here is is a corollary to Corollary 4.40:

Corollary 4.41. *Let G be a group of finite order d. Then $a^d = 1$ for all $a \in G$.*

Proof. Let the order of a be q. We saw in Corollary 4.40 that $q|d$, so $d = mq$ for some integer m. Then, $a^d = (a^q)^m = 1^m = 1$. $\square$

One of the prettiest theorems in elementary number theory is Fermat's Little Theorem, which is purely an application of the corollary above.

Theorem 4.42. *Let p be a prime, and let a be any integer. Then $a^p \equiv a \pmod{p}$.*

Proof. If $p|a$, then clearly $p|a^p$, so both a^p and a are congruent to 0 mod p. In particular, this means that $a^p \equiv a \pmod{p}$ for such a. Thus, we only need to consider the case where $p \nmid a$. In that case, note that $[a]_p = [r]_p$, where r is one of $1, 2, \ldots, p-1$. Now we have already observed in our earlier examples (see Example 2.59 in Chapter 2) that $\mathbb{Z}/p\mathbb{Z}$ is a field. By Exercise 2.48 of the same chapter, the nonzero elements of a field form a group under multiplication. In particular, this means that the nonzero elements of $\mathbb{Z}/p\mathbb{Z}$ form a group under multiplication, and this group clearly has order $p-1$. So, by the corollary above, $[r]_p^{p-1} = [1]_p$. Multiplying both sides by $[r]_p$, we find $[r]_p^p = [r]_p$. Since $[r]_p$ is just $[a]_p$, we find $[a]_p^p = [a]_p$ in the ring $\mathbb{Z}/p\mathbb{Z}$, so, back in $\mathbb{Z}$, we find $a^p \equiv a \pmod{p}$.

$\square$

4.3 Normal Subgroups, Quotient Groups

Recall how we formed a quotient ring R/I(see page 57) from a ring R and an ideal I; the elements of R/I were the cosets $a+I$ as a ranged through R, and the addition and multiplication were defined respectively by $(a+I)+(b+I) = (a+b)+I$, and $(a+I)\cdot(b+I) = (ab)+I$. We showed that these rules for addition and multiplication were well-defined (Lemma 2.80) and then went on to show (Theorem 2.82) that R/I with these operations was indeed a ring. Similarly, recall how we formed a quotient space V/W (see page 130) from a vector space V over a field F and a subspace W: the elements of V/W were the cosets $u+W$ as u ranged through V, and the vector addition and scalar multiplication were defined respectively by $(u+W)+(v+W) = (u+v)+W$, and $f(u+W) = fu+W$. Once again, we observed that these rules for vector addition and scalar multiplication were well-defined (Exercise 3.69) and then went on to show that V/W with these operations was indeed a vector space over F (Theorem 3.70).

We would of course like to mimic these constructions and form a quotient group G/H from a group G and a subgroup H: we would take the elements of G/H to be the various (say, left) cosets gH as g ranges through G, and we would define the group operation on G/H by $aH \cdot bH = (ab)H$. But when we carry out this program, we run into a slight problem: in general, the operation $aH \cdot bH = (ab)H$ is not well defined! For, suppose that $aH = a'H$ and $bH = b'H$. Viewing aH as $a'H$ and bH as $b'H$, our desired operation should yield that $aH \cdot bH = a'H \cdot b'H = (a'b')H$. Thus, $(ab)H$ ought to equal $(a'b')H$ whenever $aH = a'H$ and $bH = b'H$ (or put differently, whenever $a \sim_L a'$ and $b \sim_L b'$).

In general, this need not happen. For instance, take $G = S_3$, and take $H = \langle f_1 \rangle$. Consider the cosets $r_1\langle f_1 \rangle = \{r_1, f_3\}$ and $r_2\langle f_1 \rangle = \{r_2, f_2\}$ (see the table in Example 4.34). Now, it is clear from the table that $r_1\langle f_1 \rangle = f_3\langle f_1 \rangle$ and that $r_2\langle f_1 \rangle = f_2\langle f_1 \rangle$. So, the question is: is $(r_1r_2)\langle f_1 \rangle = (f_3f_2)\langle f_1 \rangle$? The answer is no! We find that $(r_1r_2)\langle f_1 \rangle = 1\langle f_1 \rangle = \{1, f_1\}$, while $(f_3f_2)\langle f_1 \rangle = r_2\langle f_1 \rangle = \{r_2, f_2\}$.

So how should one fix this problem? Let us first analyze the situation some more. Since $a' = a' \cdot 1 \in a'H$ and since $a'H = aH$, we find $a' \in aH$, so $a' = ah$ for some $h \in H$. Similarly, $b' = bk$ for some $k \in H$. Then $a'b' = ahbk$. If $(ab)H$ ought to equal $(a'b')H$, then $a'b'$ ought to equal abl for some $l \in H$ (see Exercise 4.38). We have gotten $a'b'$ to look like $ahbk$, let us massage this a bit and write it as $abb^{-1}hbk$. Now, *suppose* that $b^{-1}hb$ is also in H by some miracle, say that $b^{-1}hb = j$ for some $j \in H$. Then, $a'b' = ahbk = abb^{-1}hbk = abjk$, and of course, $jk \in H$ as both j and k are in H. It would follow that if this miracle were to happen, then $a'b'$ would look like ab times an element of H, and therefore, abH would equal $a'b'H$.

As the example of $G = S_3$ and $H = \langle f_1 \rangle$ above shows, this miracle will not always happen, but there are some special situations where this will happen, and we give this a name:

Definition 4.43. Let G be a group. A subgroup H of G is called a *normal* subgroup if for any $g \in G$, $g^{-1}hg \in H$ for all $h \in H$.

Remark 4.44. Alternatively, write $g^{-1}Hg$ for the set $\{g^{-1}hg \mid h \in H\}$. Then we may rewrite the definition above as follows: H is said to be normal if $g^{-1}Hg \subseteq H$ for all $g \in G$. Note that this is equivalent to requiring that $gHG^{-1} \subseteq H$ for all $g \in G$. For, setting y to be g^{-1}, note that as g ranges through all the elements of G, $y = g^{-1}$ ranges through all the elements of G as well.

Example 4.45. Take $G = S_3$ again, but this time around, take $H = \langle r_1 \rangle = \{1, r_1, r_2\}$. Let us consider the sets $g^{-1}Hg$ as g ranges through S_3. We can obtain the various products by using the group table for S_3 on page 160, for instance, $f_1\{1, r_1, r_2\}f_1^{-1} = f_1\{1, r_1, r_2\}f_1 = \{f_1f_1, f_1r_1f_1, f_1r_2f_1\} = \{1, r2, r_1\}$, etc. Doing so, we obtain the following:

g	$g\{1, r_1, r_2\}g^{-1}$
id	$\{1, r_1, r_2\}$
r_1	$\{1, r_1, r_2\}$
r_2	$\{1, r_1, r_2\}$
f_1	$\{1, r_2, r_1\}$
f_2	$\{1, r_2, r_1\}$
f_3	$\{1, r_2, r_1\}$

Thus, for each $y \in G$, we find $yHy^{-1} = H$ (so most definitely, $yHy^{-1} \subseteq H$ as needed in Definition 4.44), so indeed H is a normal subgroup of G.

It was not a coincidence in the example above that yHy^{-1} actually turned out be *equal* to H instead of merely being a subset. We have the following easy result:

Lemma 4.46. *Prove that if H is a normal subgroup of G, then indeed $yHy^{-1} = H$ for all $y \in G$.*

Proof. Fix a $y \in G$. Since H is normal, we know that $yHy^{-1} \subseteq H$. We wish to show that $H \subseteq yHy^{-1}$ as well. But since H is normal, $y^{-1}H(y^{-1})^{-1} \subseteq H$, so $y^{-1}Hy \subseteq H$. Thus, for any $h \in H$, $y^{-1}hy = k$ for some $k \in H$. We may rewrite this as $h = yky^{-1}$ by pre-multiplying by y and post-multiplying by y^{-1}. But yky^{-1} is an element of yHy^{-1} as $k \in H$, so we find that for each $h \in H$, $h \in yHy^{-1}$. Thus $H \subseteq yHy^{-1}$ as desired. □

There is an immediate corollary to this:

Corollary 4.47. *Let G be a group, and let N be a normal subgroup. Then for any $g \in G$, the left coset gN and the right coset Ng are equal.*

Proof. Since N is normal, we may apply the lemma above with $y = g$ to find $gNg^{-1} = N$. Hence, for any $n \in N$, we have $gng^{-1} = m$ for some $m \in N$. Post-multiplying this by g, we find $gn = mg$. Thus, $gn \in Ng$. Since this is true for arbitrary $n \in N$, we find $gN \subseteq Ng$. For the reverse inclusion, take $y = g^{-1}$ in Lemma 4.46 to find $g^{-1}Ng = N$ as well. Hence, given any $n \in N$, $g^{-1}ng = m$ for some $m \in N$. Pre-multiplying this by g, we find $ng = gm$, so $ng \in gNH$. Since this is true for arbitrary $n \in N$, we find $Ng \subseteq gN$. □

Remark 4.48. As a result of this corollary, if N is normal in G, we may simply talk of *the cosets of* N without specifying whether these are left or right coset.

> **Exercise 4.49.** Prove the converse of Corollary 4.47: If N is a subgroup of G such that for every $g \in G$, the left coset gN equals the right coset Ng, then N is normal.

Exercise 4.50. Prove that the center of a group (see Definition 4.5.1) is a normal subgroup.

Exercise 4.51. Prove that every subgroup of an abelian group is normal.

The following is now a consequence of all our discussions:

Lemma 4.52. *Let G be a group, and let N be a normal subgroup. Denote by G/N the set of cosets (see Remark 4.48) of N. Then, the binary operation defined on G/N by $(aN)(bN) = (ab)N$ is well-defined.*

Proof. The proof of this lemma is contained in the discussions just before Definition 4.43. In fact, it was precisely the analysis of what would make the operation $(aH)(bH) = (ab)H$ on the (left) cosets of an arbitrary group H well-defined that led us to the definition of normal subgroups. It would be a good idea to read that discussion and furnish the proof of this lemma yourselves.

□

Theorem 4.53. *Let G be a group, and let N be a normal subgroup. Then, the set G/N, with the operations as defined in the statement of Lemma 4.52, is a group.*

Proof. We have observed in Lemma 4.52 that these operations are well-defined. We have to check if all group axioms are satisfied.

1. *Associativity:* Given aN, bN, and cN in G/N, we have $(aN)[(bN)(cN)] = (aN)[(bc)N] = [a(bc)]N = [(ab)c]N$ (the last equality because of associativity in G). On the other hand, $[(aN)(bN)](cN) = [(ab)N](cN) = [(ab)c]N$. Hence, $(aN)[(bN)(cN)] = [(aN)(bN)](cN)$.

2. *Identity element:* The coset $N = 1{\cdot}N$ acts as the identity element. For, for any aN, we have $(aN)(1 \cdot N) = (a \cdot 1)N = aN$, and $(1 \cdot N)(aN) = (1 \cdot a)N = aN$.

3. *Existence of inverses:* For any aN, consider the coset $a^{-1}N$. We find $(aN)(a^{-1}N) = (aa^{-1})N = 1 \cdot N = N$, and similarly, $(a^{-1}N)(aN) = (a^{-1}a)N = 1 \cdot N = N$. Since the coset N is the identity element in G/N, we find that the coset $a^{-1}N$ is the inverse of the coset aN.

This proves that G/N with the operation as above is indeed a group.

□

Definition 4.54. The set G/N with the binary operation defined in the statement of Lemma 4.52 is called the *quotient group* of G by the normal subgroup N.

Exactly as with quotient rings and quotient vector spaces, the intuition behind quotient groups is that it is a group obtained from a group G by "killing" or "dividing out" all elements in a given normal subgroup N. Thus, G/N is to be thought of as the set of all remainders left after dividing out by N, endowed with the natural "quotient" operation of Lemma 4.52.

Exercise 4.55. If the order of G is finite, show that $o(G/N) = o(G)/o(N)$.

Exercise 4.56. Take G to be D_4, and take N to be the group generated by σ^2 (see Example 4.5 and also Exercise 4.5.2).

1. Prove that N is normal in G.
2. Prove that G/N has order 4.
3. Prove that G/N is abelian.
4. Prove that G/N is not cyclic.

4.4 Group Homomorphisms and Isomorphisms

Having had enough experience with quantifying the fact that sometimes the ring operations in two given rings may be "the same except perhaps for dividing out by an ideal," or that, sometimes, the vector space operations in two given vector spaces over a field may be "the same except perhaps for dividing out by a subspace," the following concept should now be very intuitive:

Definition 4.57. Let G and H be groups. A function $f : G \to H$ is called a group homomorphism if $f(g)f(h) = f(gh)$ for all $g, h \in G$.

Remark 4.58. Just as with the definitions of ring homomorphisms and vector space homomorphisms (linear transformations) there are some features of this definition that are worth noting:

1. In the equation $f(g)f(h) = f(gh)$, note that the operation on the left side represents the group operation in the group H, while the operation on the right side represents the group operation in the group G.

2. By the very definition of a function, f is defined on all of G. The image of G under f, however, need not be all of H (i.e, f need not be surjective). We will see examples of this ahead (see Example 4.63 and Example 4.64 for instance). However, the image of G under f is not an arbitrary subset of H: the definition of a group homomorphism ensures that the image of G under f is actually a *subgroup* of H (see Lemma 4.68 later in this section).

3. Note that it is not necessary to stipulate that $f(1_G) = 1_H$ since the property holds automatically, see Lemma 4.59 below.

Lemma 4.59. *Let $f : G \to H$ be a group homomorphism. Then $f(1_G) = 1_H$.*

Proof. We have already seen the proof of this in the context of ring homomorphisms (Lemma 2.90 in Chapter 2) and of vector space homomorphisms (Lemma 3.77 in Chapter 3). For completeness, we will prove it again: you should read this proof and go back and re-read the proofs of the corresponding lemmas on ring homomorphisms and vector space homomorphisms. We have $f(1_G) = f(1_G \cdot 1_G) = f(1_G) \cdot f(1_G)$, so putting this together, we have

$f(1_G) = f(1_G) \cdot 1_H = f(1_G) \cdot f(1_G)$. Invoking left cancellation (see Exercise 4.18), we find $1_H = f(1_G)$. □

We get an immediate corollary to this (see Corollary 2.91 in Chapter 2, as also Remark 3.81 in Chapter 3):

Corollary 4.60. *Let $f\colon G \to H$ be a group homomorphism. Then for any $g \in G$, $f(g^{-1}) = (f(g))^{-1}$.*

Proof. Since $gg^{-1} = 1_G$, we have $f(gg^{-1}) = f(g)f(g^{-1}) = f(1_G) = 1_H$, and similarly, from $g^{-1}g = 1_G$ we find $f(g^{-1})f(g) = 1_H$. This shows that $f(g)$ and $f(g^{-1})$ are inverses in H. □

The following definition should be natural at this point, after your experiences with ring homomorphisms and vector space homomorphisms:

Definition 4.61. Given a group homomorphism $f : G \to H$, the *kernel* of f is the set $\{g \in G \mid f(g) = 1_H\}$. It is denoted $ker(f)$.

No surprise here:

Proposition 4.62. *The kernel of a group homomorphism $f : G \to H$ is a normal subgroup of G.*

Proof. Let us prove first that $ker(f)$ is a subgroup. Since $1_G \in ker(f)$ (see Lemma 4.59), $ker(f)$ is certainly nonempty. Now that we know it is nonempty, by Lemma 4.23, it is sufficient to show that whenever g and k are in $ker(f)$, then gk^{-1} is also in $ker(f)$. First note that by Corollary 4.60, $f(k)$ and $f(k)^{-1}$ are inverses of each other in the group H. With this at hand, we have $f(gk^{-1}) = f(g)f(k^{-1}) = f(g)(f(k))^{-1} = 1_H 1_H = 1_H$ (we have invoked the fact here that both g and k are in the kernel of f so they get mapped to 1_H under f). We thus find $gk^{-1} \in ker(f)$ as desired.

To show $ker(f)$ is normal, we need to show that $gkg^{-1} \in ker(f)$ for all $g \in G$ and all $k \in ker(f)$. But this is easy: for any $g \in G$ and $k \in ker(f)$,

$f(gkg^{-1}) = f(g)f(k)f(g^{-1}) = f(g)1_H(f(g))^{-1} = f(g)(f(g))^{-1} = 1_H$, so indeed, $gkg^{-1} \in ker(f)$.

$\square$

Example 4.63. Given groups G and H, the map $f : G \to H$ that sends every $g \in G$ to 1_H is a group homomorphism.

> **Question 4.63.1.** Why is this f a group homomorphism? What is the kernel of f?

Notice that if H has more than just the identity element, then f is not surjective.

Example 4.64. Let R and S be rings, and let $f : R \to S$ be a ring homomorphism. Then, focusing just on the addition operations in R and S (with respect to which we know that R and S are abelian groups), the function $f : (R, +) \to (S, +)$ is a group homomorphism. In particular, if f is not surjective as a ring homomorphism (for example, the natural inclusion map $\mathbb{Z} \to \mathbb{Q}$, see Example 2.97 in Chapter 2), then f is not surjective as a group homomorphism either.

Example 4.65. Let G and H be groups (see Example 4.8). Define a function $f : G \times H \to H$ by $f(g, h) = h$.

> **Question 4.65.1.** Why is this f a group homomorphism? What is the kernel of f?

Example 4.66. Define a function $f : S_3 \to \{1, -1\}$ (see Example 4.6) by $f(r_1^i f_1^j) = (-1)^j$ (see Exercise 4.2.1).

> **Question 4.66.1.** Why is this f a group homomorphism? What is the kernel of f?

We now come to group isomorphisms. Just as ring isomorphisms capture the notion that the addition and multiplication in two rings are "essentially the same" *without even having to divide out by any ideal,* and just as vector space isomorphisms capture the notion that the vector space operations in two vector spaces are "essentially the same" *without even having to divide out by any subspace,* group isomorphisms capture the notion that the group operations in two groups are "essentially the same" *without even having to divide out by any normal subgroup.*

As with rings and vector spaces, we need a couple of lemmas first:

Lemma 4.67. *Let G and H be two groups and let $f : G \to H$ be a group homomorphism. Then f is an injective function if and only if $ker(f)$ is the trivial subgroup $\{1_G\}$.*

Proof. The proof of this is very similar to the proof of the corresponding Lemma 2.102 in Chapter 2: let us redo that proof in the context of groups. Suppose f is injective. Suppose that $g \in ker(f)$, so $f(g) = 1_H$. By Lemma 4.59, $f(1_G) = 1_H$. Since both g and 1_G map to the same element in H and since f is injective, we find $g = 1_G$. Thus, the kernel of f consists of just the element 1_G, which is precisely the trivial subgroup. Conversely, suppose that $ker(f) = \{1_G\}$. Suppose that $f(g_1) = f(g_2)$ for g_1, g_2 in G. Since f is a group homomorphism, we find $f(g_1g_2^{-1}) = f(g_1)f(g_2^{-1}) = f(g_1)(f(g_2))^{-1}$ (the last equality is because of Remark 4.60), and of course $f(g_1)(f(g_2))^{-1} = f(g_1)(f(g_1))^{-1} = 1_H$. Thus, $g_1g_2^{-1} \in ker(f)$. But $ker(f) = \{1_G\}$, so $g_1g_2^{-1} = 1_G$, i.e., $g_1 = g_2$. Hence, f is injective.

□

Our next lemma is analogous to Lemma 2.103 of Chapter 2 and Lemma 3.88 of Chapter 3:

Lemma 4.68. *Let G and H be two groups and let $f : G \to H$ be a group homomorphism. Write $f(G)$ for the image of G under f. Then $f(G)$ is a subgroup of H.*

Proof. Note that $1_H \in f(G)$ by Lemma 4.59, so $f(G)$ is nonempty. We can hence apply Lemma 4.23, so given h and k in $f(G)$, we need to show that hk^{-1} is also in $f(G)$. By definition of being in $f(G)$, there exist g and j in G

such that $f(g) = h$ and $f(j) = k$. Note that $f(j^{-1}) = k^{-1}$, by Remark 4.60. Hence $f(gj^{-1}) = f(g)f(j^{-1}) = hk^{-1}$, showing that $hk^{-1} \in f(G)$. Hence $f(G)$ is a subgroup of H. □

We are now ready for

Definition 4.69. Let $f : G \to H$ be a group homomorphism. If f is both injective and surjective, then f is said to be an *isomorphism* between G and H. Two groups G and H are said to be *isomorphic* (written $G \cong H$) if there is some function $f : G \to H$ that is an isomorphism between G and H.

Here are some examples:

Example 4.70. The function $f : S_2 \to (\mathbb{Z}/2\mathbb{Z}, +)$ that sends 1_{S_2} to $[0]_2$ and the element $(1, 2)$ (written in cycle notation) to $[1]_2$ is an isomorphism (see Exercise 4.3.2). Verify this!

Example 4.71. The groups S_3 and D_3 are isomorphic.

> **Question 4.71.1.** Compare their group tables on pages 160 and 168. Can you determine a function $f : S_3 \to D_3$ that effects an isomorphism between S_3 and D_3.

Example 4.72. Recall that Remark 4.32 showed that if G is a cyclic group of order n, generated by an element g, then g also has order n and that $G = \{1, g, \dots, g^{n-1}\}$.

> **Exercise 4.72.1.** Extend this statement to prove: If G and H are any two cyclic groups of order n, then $G \cong H$.

Example 4.73. Let G be a cyclic group of order n and H a cyclic group of order m. If m and n are relatively prime, then the direct product $G \times H$ is isomorphic to C_{nm}. (See Exercise 4.29.2, as also, Exercise 4.86 at the end of the chapter.)

> **Exercise 4.73.1.** Prove this by showing first that if $G = \langle g \rangle$ and $H = \langle h \rangle$, then (g, h) must have order mn. Since mn is also the order of $G \times H$, $G \times H$ must equal the cyclic subgroup generated by (g, h). Now use Exercise 4.72.1 above.

Example 4.74. Recall the group G/N where $G = D_4$ and N is the subgroup generated by σ^2 (see Exercise 4.56).

> **Exercise 4.74.1.** Prove that G/N is isomorphic to $(\mathbb{Z}/2\mathbb{Z}, +) \times (\mathbb{Z}/2\mathbb{Z}, +)$.

Finally, we have the following:

Theorem 4.75. *(Fundamental Theorem of Homomorphisms of Groups.) Let $f : G \to H$ be a homomorphism of groups, and write $f(G)$ for the image of G under f. Then the function $\tilde{f} : G/ker(f) \to f(G)$ defined by $\tilde{f}(g \cdot ker(f)) = f(g)$ is well-defined, and provides an isomorphism between $G/ker(f)$ and $f(G)$.*

Proof. The proof is similar to the proofs of the corresponding theorems for rings (Theorem 2.110 of Chapter 2) and vector spaces (Theorem 3.94 of Chapter 3). We first check that $\tilde{f}$ is well-defined. Suppose that $g \cdot ker(f) = h \cdot ker(f)$. Then $gh^{-1} \in ker(f)$, so $f(gh^{-1}) = 1_H$. But also, $f(gh^{-1}) = f(g)f(h^{-1}) = f(g)(f(h))^{-1}$. Thus, $1_H = f(g)(f(h))^{-1}$, so $f(g) = f(h)$.

Now we check that $\tilde{f}$ is a homomorphism. We have $\tilde{f}(g \cdot ker(f)) \cdot \tilde{f}(h \cdot ker(f)) = f(g)f(h) = f(gh)$ (as f is a group homomorphism). On the other hand, $\tilde{f}((gh) \cdot ker(f)) = f(gh)$. Hence $\tilde{f}$ is a group homomorphism.

We check that $\tilde{f}$ is surjective: Note that any $h \in f(G)$ is by definition of the form $f(g)$ for some $g \in G$. It is clear that $\tilde{f}(g \cdot ker(f)) = f(g) = h$, so $\tilde{f}$ is surjective.

Now we check that $\tilde{f}$ is injective. Suppose that $\tilde{f}(g \cdot ker(f)) = 1_H$. Then, $f(g) = 1_H$, so $g \in ker(f)$. It follows that $g \cdot ker(f) = ker(f)$, i.e. $g \cdot ker(f) = 1_{G/ker(f)}$. Hence f is injective.

□

Here is a quick exercise that uses this theorem:

> **Exercise 4.76.** Let G be a group of finite order, and let $f : G \to H$ be a *surjective* group homomorphism. Prove that H also has finite order, and that the order of H divides the order of G. (Hint: Combine Exercise 4.55 and the theorem above.)

4.5 Further Exercises

Exercise 4.77. You have seen the dihedral groups of index 3 and 4 in the text (Examples 4.4 and 4.5). The groups D_n are defined analogously for $n \geq 5$. Determine the group table for D_5 and determine its center.

Exercise 4.78. We will determine the group of symmetries of the set in Example 4.92 (see Page 208). Recall from Example 3.82 in Chapter 3 that after fixing a basis of $\mathbb{R}^2$, we can identify the set of all linear transformations of $\mathbb{R}^2$ with $M_2(\mathbb{R})$.

Let T be a linear transformation that preserves the structure of our set, and let M_T be its matrix representation with respect to, say, the standard basis $\{\mathbf{i}, \mathbf{j}\}$ of $\mathbb{R}^2$. Then,

$$M_T = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

for suitable real numbers a, b, c, and d. We will describe the conditions that a, b, c, and d must satisfy:

1. By considering the lengths of an arbitrary vector (x, y) before and after applying T, prove that $(ax + by)^2 + (cx + dy)^2 = x^2 + y^2$ for all x and y in $\mathbb{R}$.

2. Show that this relation leads to the following necessary and sufficient conditions for M_T to represent a symmetry of our set:

   (a) $a^2 + c^2 = 1$

   (b) $b^2 + d^2 = 1$, and

   (c) $ab + cd = 0$

3. Show that the conditions in (2) above are equivalent to the condition $(M_T)^t M_T = I$, where I is the identity matrix, and $(M_T)^t$ stands for the transpose of M_T. Conclude from this that any matrix that satisfies the condition in (2) above must have determinant equal to ± 1.

4. Now assume that M_T satisfies these conditions. Observe that this means that the columns of M_T are of length 1, and that the two columns are perpendicular to each other. (Such a matrix is called *orthonormal.*) We have thus determined the symmetries of the set in Example 4.92 to be the set of 2×2 orthonormal matrices with entries in $\mathbb{R}$. Now prove that this set actually forms a group under matrix multiplication. This group is known as the *orthogonal group of order* 2 *over* $\mathbb{R}$. You should go back

and revisit Example 4.13 as well. In alignment with that example, the group in this exercise should be denoted $O_2(\mathbb{R})$.

Exercise 4.79. Here is a group that generalizes Example 4.12. Let $U_n(\mathbb{R})$ denote the set of $n \times n$ upper triangular matrices with entries in $\mathbb{R}$, all of whose diagonal entries equal 1. Thus, every matrix in $U_n(\mathbb{R})$ can be expressed as the sum of the identity matrix I and a *strictly* upper triangular matrix N. We will show that $U_n(\mathbb{R})$ forms a group with respect to multiplication.

1. For any matrix M in $M_n(\mathbb{R})$, define the *level* l of M as follows: the level of the zero matrix is ∞, and for nonzero matrices M, $l(M) = \min\{j - i \mid M_{i,j} \neq 0\}$, where $M_{i,j}$ stands for the (i,j) entry of M. Thus, a matrix is of level 0 if and only if it is upper triangular with at least one nonzero entry along the diagonal, and it is of level 1 if and only if it is strictly upper triangular, with at least one nonzero entry along the "super diagonal" (the super diagonal is the set of entries that run diagonally from the $(1,2)$ slot down to the $(n-1,n)$ slot), etc. Show that $l(MN) \geq l(M)+l(N)$. Give an example of matrices M and N such that $MN \neq 0$, and $l(MN) > l(M) + l(N)$.

2. Conclude that any strictly upper triangular matrix N is nilpotent (see Exercise 2.122 in Chapter 2).

3. Now show using Parts (1) and (2) above that $U_n(\mathbb{R})$ forms a group with respect to matrix multiplication. (You may also want to look at Exercise 2.122 in Chapter 2 for some ideas.)

Exercise 4.80. Let G be a group with an even number of elements. Prove that G has at least one *nonidentity* element a such $a^2 = 1$. (Hint: To say that $a^2 = 1$ is to say that $a = a^{-1}$. Now pair the elements in the group suitably, and invoke the fact that the group has an even number of elements.)

Exercise 4.81. Prove that a group G is abelian if and only if the function $f : G \to G$ that sends any g to g^{-1} is a homomorphism.

Exercise 4.82. Prove that a group G is abelian if and only if $(gh)^2 = g^2h^2$ for all g and h in G.

Exercise 4.83. The discussions preceding Definition 4.43 established the following: if N is a normal subgroup of G, then the operation on the left cosets of N determined by $(aN)(bN) = abN$ is well-defined. Prove the converse of this: if N is a subgroup of G such that the operation on the left cosets of N determined by $(aN)(bN) = abN$ is well-defined, then N must be normal in G. (Hint:

For any $g \in G$, consider the product of left cosets $g^{-1}N \cdot gN = 1_G N = N$. For any $n \in N$, note that the coset $(g^{-1}n)N$ is the same as $g^{-1}N$ (why?). Since the product is well-defined, $(g^{-1}n)N \cdot gN$ should also equal N. So?)

Exercise 4.84. What is the last digit in 43^{99999}? (Hint: Work mod 2 and mod 5 separately, applying Corollary 4.41 above, and then combine the result.)

Exercise 4.85. By Exercise 4.51, the center $Z(G)$ of a group G is a normal subgroup of G. Hence, it makes sense to talk of the quotient group $G/Z(G)$. Prove that if $G/Z(G)$ is cyclic, then G must be abelian, and thus, $Z(G)$ must equal G.

Exercise 4.86. Let G be a cyclic group of order m and H a cyclic group of order n. Show that $G \times H$ is cyclic if and only if $\gcd(m, n) = 1$.

Notes

Remarks on sets with structure and their symmetry groups Recall from the text (Page 158) that a set with structure is simply a set with a certain feature that we wish to focus on, and a symmetry of such a set is a one-to-one onto map from the set to itself that preserves this feature. If f and g are two such maps, then the composition $f \circ g$ as well as $g \circ f$ will also be one-to-one onto maps that preserve the feature. (Recall that if $f : S \to S$ and $g : S \to S$ are two functions from a set S to itself, then the *composition* of f and g, written $f \circ g$ takes $s \in S$ to $f(g(s))$, and similarly, $g \circ f$ takes s to $g(f(s))$.) Often, if f is such a feature-preserving map, then f^{-1} (which exists because f is a bijection) will also preserve the feature, although, this is not always guaranteed. (See the remarks on Page 211 later in these notes for some examples where the inverse of a structure-preserving map is not structure-preserving.) So, if we restrict our attention to those structure-preserving maps whose inverse is also structure-preserving, then these maps constitute a group, called the *symmetry group* of the set with the given structure.

We consider some examples below of sets with structure and their symmetry groups:

Example 4.87. The set in question could be any set, such as $\Sigma_3 = \{1, 2, 3\}$, with the feature of interest being merely the fact that it is a set. (This structure is called the *trivial* structure.) Of course this particular set has lots of other features (for example, each element in Σ_3 corresponds to a length on a number line–see Example 4.88 below), but we do not focus on any other feature for the moment. Any one-to-one onto map from a set such as Σ_3 to itself will certainly preserve the feature that Σ_3 is a set, so, the symmetries of a set with trivial structure are precisely the various one-to-one maps from the set to itself. We have already considered this group in Example 4.2, it is S_3.

Example 4.88. If we consider instead Σ_3 with the feature that each element corresponds to a length on a number line—1 to the unit length, 2 to twice the unit-length, and 3 to three times the unit length—then our symmetry group would be different. Any symmetry f would now have to satisfy the property that if $n \in \Sigma_3$ corresponds to a certain length, then $f(n)$ should also correspond to the same length. It follows immediately that $f(1)$ has to equal 1: $f(1)$ cannot be 2 or 3 since 1 has unit length while 2 has twice the unit length and 3 has three times the unit length. Similary $f(2) = 2$ and $f(3) = 3$. Hence, f can only be the identity map. Thus, the symmetry group of Σ_3 with the feature that each element corresponds to a length on the number line is the trivial group consisting of just the identity.

Example 4.89. The set could be a piece of cardboard cut in the shape of an equilateral triangle with the feature that it is a rigid object. Again, this set could have other features (for example, the cardboard could be colored in alternating horizontal strips of black and white–see Question 4.89.1 below), but we will ignore those. The symmetries of this set would be those one-to-one and onto maps f from the triangle to itself that preserve the rigidity of the triangle, i.e, that do not distort the cardboard. (Put differently, if p and q are any two points on the triangle, then the distance between p and q should be the same as the distance between $f(p)$ and $f(q)$). We have seen this group before: it is the group D_3 (Example 4.4).

> **Question 4.89.1.** Pick one edge of the triangle, and refer to its direction as the "horizontal" direction. Suppose the piece of cardboard of this example had, additionally, been colored in alternating horizontal strips of black and white. Suppose that the total number of strips is odd (and at least three), so that the strips along the two horizontal edges are both of the same color and there is at least one strip of the other color. What would be the symmetries of this new set? What would be the symmetries if the total number of strips were even, so that the strips along the two horizontal edges are of different color?

Example 4.90. Just as in the last example, the set could be a piece of cardboard cut in the shape of a square, with the structure that it is a rigid object. We have seen the symmetries of this set, it is D_4 (Example 4.5).

> **Question 4.90.1.** Pick one edge of the square, and refer to its direction as the "horizontal" direction. Suppose the piece of cardboard of this example had, additionally, been colored in alternating horizontal strips of black and white. What would be the symmetries of this new set?

Example 4.91. The set could be $\mathbb{R}^2$, and the feature of interest could be the fact that it is a vector space over the reals. What would be the symmetries of a vector space that preserve its vector space structure? In fact, what should it mean to preserve the vector space structure? A vector space is characterized by two operations, an addition operation on vectors, and a scalar multiplication operation between a scalar and a vector. We say that a map $f : \mathbb{R}^2 \to \mathbb{R}^2$ preserves its vector space structure if for any two vectors v and w in $\mathbb{R}^2$ and for any scalar $a \in \mathbb{R}$, "f respects these operations," i.e., if f sends v to some $f(v)$ and w to some $f(w)$, then f must send $v + w$ to $f(v) + f(w)$ and av to $af(v)$. We have seen such maps f before in Chapter 3: f must be a *linear transformation* of $\mathbb{R}^2$. Hence, a symmetry of $\mathbb{R}^2$ with its vector space structure is a linear transformation of $\mathbb{R}^2$ that is both injective and surjective.

> **Exercise 4.91.1.** Show that if f is a one-to-one onto linear transformation of $\mathbb{R}^2$, then the inverse map f^{-1} is also a linear transformation of $\mathbb{R}^2$. (Hence, the inverse also preserves the vector space structure of $\mathbb{R}^2$. It follows that the symmetry group of $\mathbb{R}^2$ with its vector space structure is precisely the set of injective and surjective linear transformations of $\mathbb{R}^2$.)

Example 4.92. This examples puts further conditions on Example 4.91: We could take the set to be $\mathbb{R}^2$ as before, with the structure being that it is a vector space over the reals, and that *each vector has a length.* (Recall that the length of the vector (a, b) in $\mathbb{R}^2$ is taken to be $\sqrt{a^2 + b^2}$.) The symmetries of this set would be one-to-one onto linear transformations of $\mathbb{R}^2$ that in addition preserve the length of a vector. (See Exercise 4.78 at the end of the chapter.)

Example 4.93. This example and the next are central in Galois theory, and you may wish to postpone them for a future reading. The set could be the field $\mathbb{Q}[\sqrt{2}]$, and the structure could be that (i) $\mathbb{Q}[\sqrt{2}]$ is a field, and (ii) every element in $\mathbb{Q}[\sqrt{2}]$ satisfies a family of polynomials over the rationals.

The symmetries of $\mathbb{Q}[\sqrt{2}]$ that preserve the fact that it is a field are those one-to-one onto maps $f : \mathbb{Q}[\sqrt{2}] \to \mathbb{Q}[\sqrt{2}]$ that satisfy the property that for all a and b in $\mathbb{Q}[\sqrt{2}]$, $f(a+b) = f(a) + f(b)$, $f(ab) = f(a)f(b)$ and $f(1) = 1$: in other words, f must also be a ring homomorphism. (Note that once f satisfies the property that it is a ring homomorphism, the relations $ab = 1$ will mean that $f(a)f(b) = 1$, so pairs of multiplicative inverses will go to pairs of multiplicative inverses under f. Thus, the essential character of $\mathbb{Q}[\sqrt{2}]$ that gives it the structure of not just a ring but a field will automatically be preserved.) Put differently, we find that f must be a ring *isomorphism* from $\mathbb{Q}[\sqrt{2}]$ to $\mathbb{Q}[\sqrt{2}]$ if it is to preserve the field structure of $\mathbb{Q}[\sqrt{2}]$.

As for the second feature, we say that a one-to-one onto map $f : \mathbb{Q}[\sqrt{2}] \to \mathbb{Q}[\sqrt{2}]$ preserves the minimal polynomial over the rationals if any $a \in \mathbb{Q}[\sqrt{2}]$, the element $f(a)$ satisfies the same minimal polynomial over the rationals as a. Now take an arbitrary $a \in \mathbb{Q}$ (note). It will have minimal polynomial $x - a$ (why?). Hence $f(a)$ must also have minimal polynomial $x - a$ if the minimal polynomial is to be preserved. In particular, $f(a)$ must satisfy $f(a) - a = 0$, i.e, $f(a)$ must equal a. Since this is true for arbitrary $a \in \mathbb{Q}$, we find that any symmetry of $\mathbb{Q}[\sqrt{2}]$ that preserves the field structure and the minimal polynomial over the rationals must be a ring isomorphism that act as the identity map on the rationals. Moreover, it is easy to see that *any* ring isomorphism from $\mathbb{Q}[\sqrt{2}]$ to $\mathbb{Q}[\sqrt{2}]$ that is the identity on the rationals necessarily preserves the minimal polynomial over the rationals of an arbitrary $a \in \mathbb{Q}[\sqrt{2}]$: if a satisfies the polynomial $p(x) = x^t + q_{t-1}x^{t-1} + \cdots + q_1 x + q_0$ with coefficients in $\mathbb{Q}$, then applying f to the equation $a^t + q_{t-1}a^{t-1} + \cdots + q_1 a + q_0 = 0$ and using the fact that f is a ring homomorphism that is the identity on the rationals, we find that $f(a)^t + q_{t-1}f(a)^{t-1} + \cdots + q_1 f(a) + q_0 = 0$, i.e., $f(a)$ also satisfies $p(x)$. Conversely, if $f(a)$ satisfies some monic polynomial $q(x)$ with coefficients in the rationals, then applying f^{-1}, we find that a also satisfies $q(x)$. In particular, since a and $f(a)$ satisfy the same monic polynomials with rational coefficients, they must both have the same minimal polynomial over the rationals.

Thus, the symmetries of $\mathbb{Q}[\sqrt{2}]$ that preserve both the field structure and the minimal polynomial of elements must be ring isomorphisms from $\mathbb{Q}[\sqrt{2}]$ to $\mathbb{Q}[\sqrt{2}]$ which act as the identity on $\mathbb{Q}$. But by Exercise 2.137 in Chapter 2, any ring homomorphism from $\mathbb{Q}$ to $\mathbb{Q}[\sqrt{2}]$ must automatically be the identity map on the rationals, so this extra condition is not necessary. It follows that the symmetries of $\mathbb{Q}[\sqrt{2}]$ that preserve both the field structure and the minimal polynomial of elements over the rationals are precisely the set of ring isomorphisms from $\mathbb{Q}[\sqrt{2}]$ to itself.

> **Exercise 4.93.1.** Using the ideas developed in these remarks and using Exercise 2.109.1 in Chapter 2, prove that the only non-trivial symmetry of $\mathbb{Q}[\sqrt{2}]$ with the structure above is the familiar conjugation map that sends each $a+b\sqrt{2}$ (a, b, in the rationals) to $a-b\sqrt{2}$. Hence, there are precisely two symmetries of this set: the *do nothing* symmetry, that is, the identity map on $\mathbb{Q}[\sqrt{2}]$, and this conjugation map.

Example 4.94. (This is also from Galois theory, and as with the previous example, you may wish to postpone this for a future reading.) The set could be the field $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$, and the structure could be that (i) $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$ is a field (see Exercise 2.119 in Chapter 2), and (ii) every element in $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$ satisfies a family of polynomials over the field $\mathbb{Q}[\sqrt{2}]$.

The same considerations as in Example 4.93 above apply: The symmetries of $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$ that preserve the field structure are precisely the set of ring isomorphisms from $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$ to itself. The symmetries which also preserve the minimal polynomial of elements over $\mathbb{Q}[\sqrt{2}]$ can be determined exactly as above: these must be the ring isomorphisms from $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$ to itself that act as the identity on $\mathbb{Q}[\sqrt{2}]$. (Note that unlike the previous example, it is not true that every ring isomorphism from $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$ to itself acts as the identity on $\mathbb{Q}[\sqrt{2}]$.)

> **Exercise 4.94.1.** Using the ideas developed in these remarks and using Exercise 2.138 in Chapter 2, prove that the only non-trivial symmetry of $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$ with the structure above is the map that sends $a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6}$ to $a + b\sqrt{2} - c\sqrt{3} - d\sqrt{6}$ (here a, b, c, and d are rational numbers). Thus, there are precisely two symmetries of this set. Note, however, that Exercise 2.138 of Chapter 2 shows that there are other ring isomorphisms from $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$ to itself: these however do not act as the identity on $\mathbb{Q}[\sqrt{2}]$.

Remarks on Exercise 4.3.7 Here is how you may start this exercise: Let j be an integer in the set Σ_n. Let us consider the case where j is one of the b's, say $j = b_k$, for some k. Then $t(j) = b_{k+1}$, where the subscript is taken modulo e so as to lie in the set $\{1, 2, \ldots, n\}$. Hence $st(j)$ would be $s(b_{k+1})$. Now, because s and t are disjoint cycles, b_{k+1} will not appear among the a's, and hence, $s(b_{k+1})$ would equal b_{k+1}. Now work out $ts(j)$ for this particular case. Next consider the case where j is not one of the b's and work out the details.

Remarks on structure-preserving maps forming a group It is not always true that the inverse of a structure preserving map also preserves the structure. Typically this is so, but occasionally this is not the case. It is for this reason that we only consider symmetries of a set whose inverse also preserves the given structure when viewing the symmetries as a group. Here is an example:

We consider the real numbers with its "differentiable structure." What this means is that there exists a notion of differentiability of functions $\mathbb{R} \to \mathbb{R}$. A symmetry of $\mathbb{R}$ with its differential structure would be a one-to-one onto function $f\colon \mathbb{R} \to \mathbb{R}$ that "preserves this differentiable structure." This means that f should satisfy the condition that for any differentiable map $g\colon \mathbb{R} \to \mathbb{R}$, the composite $g \circ f$ must also be differentiable. A necessary and sufficient condition for this to happen is that f itself must be differentiable. It is now easy to find bijections $f\colon \mathbb{R} \to \mathbb{R}$ that are differentiable, but whose inverse is not differentiable. One example is the function $f(x) = x^3$. It is differentiable at all values of x, but its inverse function $f^{-1}(x) = x^{1/3}$ fails to be differentiable at $x = 0$.

Remarks on orthogonal groups Orthogonal groups come in more guises than the one we have described in Example 4.13. Recall the origins of the $n = 2$ case over $\mathbb{R}$ that we exhibited in Exercise 4.78: the group $O_2(\mathbb{R})$ is the set of symmetries of $\mathbb{R}^2$ with the structure that it is a vector space over the reals, and that every vector has a length (Example 4.92). Now let us examine "length" more closely. The length of a vector $p\mathbf{i} + q\mathbf{j}$ is defined to be $\sqrt{p^2 + q^2}$. Temporarily ignoring the square root (we will put it back later), the squared-length of a general vector $x\mathbf{i}+y\mathbf{j}$ is thus given by $x^2 + y^2$. This is an example of a "quadratic form"—a polynomial all of whose monomials are of degree 2 in—in two variables. Now note that the polynomial $x^2 + y^2$ can be written as

$$(x, y) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} (x, y)^t$$

(Here, recall from elementary linear algebra that the product of a row vector (s, t) and a column vector $(p, q)^t$ is given by $sp + tq$. Thus, since $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} (x, y)^t$ is just $(x, y)^t$, the product above becomes $(x, y)(x, y)^t = x^2 + y^2$, as claimed.)

Mathematicians have found it useful to define length differently as well (we will see a famous example of this ahead). More generally, let $q = ax^2 + 2bxy + cy^2$ be any quadratic form with coefficients a, b, c from the reals. (It is convenient to write the coefficient of xy as $2b$.) Then, q may be written as

$$(x, y)\begin{pmatrix} a & b \\ b & c \end{pmatrix}(x, y)^t$$

(Check this! Notice how the fact that we wrote the coefficient of xy as $2b$ allows us to write the $(1, 2)$ and $(2, 1)$ entries of the matrix above as b. Had we taken the coefficient of xy as b, then these entries would have had to be $b/2$.) Using this quadratic form, we define the q-length of a vector $p\mathbf{i}+q\mathbf{j}$ as $\sqrt{ap^2 + 2bpq + cq^2}$. (The length may well turn out to be imaginary—the quantity under the square root sign may be negative—but that only makes matters more interesting!) Moreover, we define the q-dot product of two vectors $s\mathbf{i}+t\mathbf{j}$ and $p\mathbf{i}+q\mathbf{j}$ to be $asp+b(sq+tp)+ctq$. Writing M_q for the matrix $\begin{pmatrix} a & b \\ b & c \end{pmatrix}$ above, we find that the q-length of $p\mathbf{i} + q\mathbf{j}$ is given by

$$(p, q)\begin{pmatrix} a & b \\ b & c \end{pmatrix}(p, q)^t,$$

and the q-dot-product of $s\mathbf{i} + t\mathbf{j}$ and $p\mathbf{i} + q\mathbf{j}$ is given by

$$(s, t)\begin{pmatrix} a & b \\ b & c \end{pmatrix}(p, q)^t,$$

The matrix M_q allows us to compute q-lengths and q-dot products; notice that M_q is a *symmetric* matrix (the $(1, 2)$ entry and $(2, 1)$ entry are equal).

Given an arbitrary quadratic form q in two variables with coefficients in the reals, we may now consider the symmetries of $\mathbb{R}^2$ with q-structure: this is the structure that $\mathbb{R}^2$ is a vector space over the reals, and that every vector has a q-length. The symmetries then would be one-to-one onto linear transformations of $\mathbb{R}^2$ that in addition preserve q-length. These symmetries form a group that we will denote $O_2(\mathbb{R}, q)$. It is called the *orthogonal group of q over $\mathbb{R}$.*

Exercise 4.95. Prove that $O_2(\mathbb{R}, q)$ consists of those 2×2 matrices A with entries in $\mathbb{R}$ satisfying $A^t M_q A = M_q$.

Exercise 4.96. Given a one-to-one onto linear transform T, let us say that it satisfies Property (1) if the q-length of $T(v)$ is the same as the q-length of v for all v in $\mathbb{R}^2$. Let us say that it satisfies Property (2) if the q-dot product of $T(v)$ and $T(w)$ is the same as the q-dot product of v and w, for all v and w in $\mathbb{R}^2$. Show that T satisfies Property (1) if and only if it satisfies Property (2).

More generally, an arbitrary quadratic form q in n variables $x_1, x_2, \ldots, x_n$ over a field F is a polynomial in these variables with coefficients in F, all of whose monomials are of degree 2. As long as $2 \neq 0$ in this field (so we rule out fields like $\mathbb{Z}/2\mathbb{Z}$), we may form a symmetric $n \times n$ matrix M_q as above, where the entries in the slots (i, j) and (j, i) both equal *half* the coefficient of $x_i x_j$ in the quadratic form q. (We have to impose the $2 \neq 0$ condition, because otherwise, we would not be able to divide by 2!) The set of $n \times n$ matrices A satisfying $A^t M_q A = M_q$ forms a group $O_n(F, q)$, referred to as the *orthogonal group of q over F*.

Perhaps the most famous example of the length of vectors in $\mathbb{R}^n$ being measured by quadratic forms other than $x_1^2 + x_2^2 + \cdots + x_n^2$ is given by Einstein's theory of relativity. There, space-time is considered as a four dimensional space, and the length of the vector $(t, x, y, z)^t$, where t is the time coordinate and x, y, and z are the usual spatial coordinates, is given by $\sqrt{t^2 - x^2 - y^2 - z^2}$. (Actually, this is a drastic simplification: space-time is not really a vector space but a four-dimensional manifold, and the length formula above applies on the tangent spaces–which are actual vector spaces–but that is too mathematically advanced for now.) This quadratic form $t^2 - x^2 - y^2 - z^2$ has associated symmetric matrix

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

The associated orthogonal group is called the *Lorentz* group.

Appendix A

Sets, Functions, and Relations

We review here some basic notions that you would have seen in an earlier course "on proofs" or on discrete mathematics.

A *set* is simply a collection of *objects*. We are of course being informal here: there are more formal definitions of sets that are based on various axioms designed to avoid paradoxes, but we will not go into such depths in this appendix. If A is a set, the objects whose collection make up the set A are also referred to as the *elements* of A. You will be familiar with both notations for sets: the explicit notation, such as $A = \{2, 3, 5, 7\}$ as well as the implicit or *set-builder* notation, such as $A = \{n \,|n \text{ is a prime integer between 2 and 10}\}$. You will also be familiar with the notation "$\in$" for "element of."

If A and B are two sets, we say A is a subset of B (written $A \subseteq B$) if $x \in A$ implies $x \in B$. If $A \subseteq B$ and $B \subseteq A$, we say $A = B$. If $A \subseteq B$ but $A \neq B$, we say that A is a *proper subset* of B, and we write $A \subset B$.

The *union* of two sets A and B, denoted $A \cup B$, is simply the set $\{x \mid x \in A \text{ or } x \in B\}$. The *intersection* of two sets A and B, denoted $A \cap B$ is the set $\{x \mid x \in A \text{ and } x \in B\}$. The *difference* of two sets A and B, denoted $A - B$, is the set $\{x \mid x \in A \text{ and } x \notin B\}$. (Note that in general, $A - B \neq B - A$.)

A *function* f from A to B (written $f\colon A \to B$) is a rule that assigns to *each* element of A a *unique* element of B. A function $f\colon A \to B$ is called *injective* or *one-to-one* if $f(a_1) = f(a_2)$ for $a_1, a_2 \in A$ implies that $a_1 = a_2$ (or alternatively, if $a_1 \neq a_2$, then $f(a_1) \neq f(a_2)$). A function $f\colon A \to B$ is called *surjective* or *onto* if for each $b \in B$, there exists $a \in A$ such that $f(a) = b$. A function $f\colon A \to B$ that is both injective and surjective is said to be *bijective*; also, f is said to provide a *bijection* or a *one-to-one correspondence* between A and B.

Example A.1. Consider the following functions from the integers to itself:

1. $f(n) = 2n$.
2. $g(n) = \begin{cases} n, & \text{if } n \text{ is odd} \\ n/2, & \text{if } n \text{ is even} \end{cases}$
3. $h(n) = n^2 + 1$.
4. $b(n) = n + 1$.

Then f is injective but not surjective, g is surjective but not injective, h is neither injective nor surjective, and b is bijective.

The *Cartesian Product* of two sets A and B, denoted $A \times B$, is simply the set of all ordered pairs $\{(a, b) \mid a \in A,\ b \in B\}$. A *relation* on a set A is simply a subset of $A \times A$. Let R be a relation on a set A. If $(a, b) \in R$, we say *a is related to b* and we often write $a\ R\ b$ to indicate that a is related to b under the relation R. The relation R is said to be *reflexive* if for each $a \in A$, $a\ R\ a$. R is said to be *symmetric* if whenever $a\ R\ b$, then $b\ R\ a$ as well. Finally, R is said to *transitive* if whenever $a\ R\ b$ and $b\ R\ c$, then $a\ R\ c$ as well.

A relation R on a set A that is reflexive, symmetric, and transitive is called an *equivalence relation* on A. For any $a \in A$, let us write $[a]$ for the set of all elements of B that are related to a, that is, $[a] = \{b \mid a\ R\ b\}$. The set $[a]$ is called the *equivalence class* of a. We have the following: if R is an equivalence relation on A, then for any two elements a and b in A, either $[a] = [b]$ or else, $[a]$ and $[b]$ are disjoint. In particular, this means that the equivalence classes divide A into disjoint sets of the form $[a]$, whose union is all of A.

The symbol $\sim$ is often used instead of "R" to denote a relation on a set.

Example A.2. The easiest and most central example perhaps of an equivalence relation on a set is the relation $\sim$ on $\mathbb{Z}$ defined by saying that m is related to n (or $m \sim n$) iff $m - n$ is even. Convince yourself that this relation is indeed an equivalence relation, and that there are precisely two equivalence classes: the class [0] and the class [1].

A *binary operation* on a set A is simply a function $f : A \times A \to A$. As we have seen, the usual operations of addition and multiplication in, for example, the integers, are just binary operations on $\mathbb{Z}$, that is, functions $\mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$.

> **Question A.3.** Is division a binary operation on the rationals? How about on the set $\mathbb{Q} - \{0\}$?

A set A is said to be *countable* if there exists a one-to-one correspondence between A and some subset of $\mathbb{N}$. If no such correspondence exists, then A is said to be *uncountable*. If there exists a one-to-one correspondence between A and the subset $\{1, 2, \ldots, n\}$ of $\mathbb{Z}$ (for some n), then A is said to be *finite*. If no such $n \in \mathbb{Z}$ exists for which there is a one-to-one correspondence between A and $\{1, 2, \ldots, n\}$, then A is said to be *infinite*. Note that an infinite set can be either countable or uncountable.

Example A.4. Any set with a finite number of elements is countable, by definition of finiteness and countability.

Example A.5. Any subset of a countable set is also countable.

Example A.6. The integers are countable. One one-to-one correspondence between $\mathbb{Z}$ and $\mathbb{N}$ is the one that sends a to $2a$ if $a \geq 0$, and a to $2(-a) - 1$ if $a < 0$.

Example A.7. The Cartesian product of two countable sets is countable. Here is a sketch of a proof when both A and B are infinite. There exists a one-to-one correspondence between A and $\mathbb{N}$ (why?), and turn, there exists a one-to-one correspondence between $\mathbb{N}$ and the set $\{2^n \; n \in \mathbb{N}\}$. Composing, we get a one-to-one correspondence f between A and the set $\{2^n \; n \in \mathbb{N}\}$. Similarly, we have a one-to-one correspondence g between B and the set the set $\{3^n \; n \in \mathbb{N}\}$. Now define the map $h : A \times B \to \mathbb{N}$ by $h(a, b) = f(a)g(b)$, and show that this is a bijection.

Example A.8. The rationals $\mathbb{Q}$ are countable. This is because we may view $\mathbb{Q} \subseteq \mathbb{Z} \times \mathbb{Z}$ by identifying the rational number a/b, written in lowest terms, with the ordered pair (a, b). By Example A.7 above, $\mathbb{Z} \times \mathbb{Z}$ is countable, and hence by Example A.5, $\mathbb{Q}$ is also countable.

Example A.9. The reals $\mathbb{R}$ are uncountable. The proof of this is the famous Cantor diagaonalization argument.

Appendix B

Partially Ordered Sets, and Zorn's Lemma

A *Partial Order* on a set S is a relation "$\leq$" on S that is reflexive, *antisymmetric* (i.e., $a \leq b$ and $b \leq a$ imply that $a = b$), and transitive. Here are two examples:

Example B.1. Define a relation "$\leq$" on the positive integers by the rule $m \leq n$ if and only if m divides n. Since $m|m$ for all positive integers, $\leq$ is reflexive. Since $m|n$ and $n|m$ imply $m = n$ (recall that we are only allowing positive integers in our set), our relation $\leq$ is indeed antisymmetric. Finally, if $m|n$ and $n|q$, then indeed $m|q$, so $\leq$ is transitive.

Example B.2. Let S be a nonempty set, and write T for the set of all *proper* subsets of S. Define a relation "$\leq$" on T by defining $X \leq Y$ if and only if $X \subseteq Y$. You should be able to verify easily that $\leq$ is a partial order on T.

This partial order could also have been defined on the set of *all* subsets of S, we chose to define it only on the set of proper subsets to make the situation more interesting (see Example B.4 ahead, for instance)!

Given a partial order "$\leq$" on a set, two elements x and y are said to be *comparable* if either $x \leq y$ or $y \leq x$. If neither $x \leq y$ or $y \leq x$, then x and y are said to be *incomparable*. For instance, in Example B.1, 2 and 3 are incomparable, since neither $2|3$ nor $3|2$. Similarly, in the set of all proper subsets of, say, the set $\{1, 2, 3\}$, the subsets $\{1, 2\}$ and $\{1, 3\}$ are incomparable, since neither of these sets is a subset of the other.

Given a partial order $\leq$ on a set S, and given a subset A of S, an *upper bound* of A is an element $z \in S$ such that $x \leq z$ for all $x \in A$.

Example B.3. In Example B.1, if we take A to be the set $\{1, 2, 3, 4, 5, 6\}$, then $\mathrm{lcm}(1, 2, 3, 4, 5, 6) = 60$ is an upper bound for A.

Note that not all subsets of S need have an upper bound. For instance, if we take B in this same example to be the set of all powers of 2, then there is no integer divisible by 2^m for all values of m, so B will not have an upper bound.

Given a partial order $\leq$ on a set S, a *maximal* element in S is an element x such that for any other element y, either $y \leq x$ or else x and y are incomparable.

Example B.4. In Example B.2, suppose we took $S = \{1, 2, 3\}$, so

$$T = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}\}.$$

Then $\{1, 2\}$ is maximal: each of $\{1\}$, $\{2\}$,, and $\{2\}$ are $\leq \{1, 2\}$, while $\{1, 3\}$and $\{2, 3\}$ cannot be compared with $\{1, 2\}$.

Of course, these same arguments show that $\{1, 3\}$and $\{2, 3\}$ are also maximal elements.

Note that instead, we had taken T to be the set of *all* subsets of $\{1, 2, 3\}$, then there would only have been one maximal element, namely $\{1, 2, 3, \}$, and all other subset X would have satisfied $X \leq \{1, 2, 3\}$. Having several maximal elements incomparable to one another is certainly a more intriguing situation!

A partial order on a set that has the further property that any two elements are comparable is called a *linear order.* For example, the usual order relation on $\mathbb{R}$ is a linear order.

Given a partial order $\leq$ on a set S, a *chain* in S is a nonempty subset A of S that is linearly ordered with respect to $\leq$, i.e., for all x and y from A, either $x \leq y$ or $y \leq x$.

Example B.5. In Example B.3, note that B is a chain, since every element of B is a power of 2, and given elements 2^m and 2^n in B, if $m \leq n$ then $2^m | 2^n$, else $2^n | 2^m$. On the other hand, A is not a chain: we have already seen that 2 and 3 are incomparable.

Zorn's Lemma, in spite of its name, is really not a lemma, but a universally accepted *axiom* of logic. It states the following:

> *Zorn's Lemma*: Let S be a nonempty set with a partial order $\leq$. If every chain in S has an upper bound in S, then S has a maximal element.

Zorn's Lemma is equivalent to certain other axioms of logic, most famously, to the *Axiom of Choice.* What this means that if one were to accept the statement of Zorn's Lemma as a fundamental axiom of logic, then in conjunction with other accepted axioms of logic, one can derive the statement of the Axiom of Choice. Conversely, if one were to accept the Axiom of Choice as a fundamental axiom of logic, then in conjunction with other accepted axioms of logic, one can derive the statement of Zorn's Lemma.

Here is a typical application of Zorn's Lemma. Recall from Exercise 2.135 of Chapter 2 the definition of maximal ideals.

Theorem B.6. *Let R be a ring. Then R contains maximal ideals.*

Proof. Let S be the set of all *proper* ideals of R. Note that S is nonempty, since the zero ideal $\langle 0\rangle$ is in S. We define a partial order $\leq$ on S by $I \leq J$ if and only if $I \subseteq J$ (see Example B.2 above). Let T be a chain in S. Recall what this means: T is a collection of proper ideals of R such that if I and J are in the collection, then either $I \subseteq J$ or else $J \subseteq I$. We claim that T has an upper bound in S, i.e., there exists a proper ideal K in R such that $I \subseteq K$ for all I in our chain T. The proof of the claim is simple. By the definition of being a chain, T is nonempty, so T contains at least one ideal of R. We define K, *as a set,* to be the union of all the ideals I in T. We need to show that K is a proper ideal of R. This is easy. Note that since there is at least one ideal in T, and since this ideal contains 0, K must be nonempty as it must contain at least the element 0. Now given a and b in K, note that a must live in some ideal I in T and b must live in some ideal J in T, since K is, after all, the union of all the ideals in T. Since T is linearly ordered (this is where the property that chains are linearly ordered comes in), either $I \subseteq J$ or else $J \subseteq I$. Say $I \subseteq J$. Then both a and b are in J. Hence, $a+b$ is also in J as J is an ideal. Since J in turn is contained in K, we find $a+b \in K$. This shows that K is closed under addition. Now given any $a \in K$, as before, $a \in I$ for some ideal I in T. Since I is an ideal, both ar and ra are in I for all $r \in R$. Since $I \subseteq K$, we find ar and ra are in K. By Lemma 2.67 of Chapter 2, we find K is an ideal. Of course, K is clearly an upper bound for T, since $I \subseteq K$ for all I in T by the very manner in which we have defined K.

Note that indeed K is a proper ideal of R, i.e., K is in S. For, if not, then $K = R$, so in particular, this means that $1 \in K$. Since K is the union of the ideals in T, we find $1 \in I$ for some ideal I in T. But this is a contradiction, since I is a proper ideal of R (remember that the set S was defined as the set of all proper ideals of R, and I is a member of S).

Since T was arbitrary, we have found that every chain in S has an upper bound in S. By Zorn's lemma, S has a maximal element. But a maximal element of S is precisely a maximal ideal of R!

□

Now we will present the proof that bases exist in all vector spaces, not just in those with a finite spanning set; this proof invokes Zorn's Lemma. Recall that we can assume that our vector space is nontrivial, thanks to Example 3.35 of Chapter 3.

Theorem B.7. *Every vector space has a basis.*

Proof. Let S be the set of all linearly independent subsets of V. Since V is not trivial by assumption, it has at least one nonzero vector, say v, and the set $\{v\}$ is then linearly independent (Exercise 3.22.1). It follows that S is a nonempty set.

Define a partial order on S by declaring, for any two linearly independent subsets X and Y, that $X \leq Y$ if and only if $X \subseteq Y$. It is easy to check that this is indeed a partial order: First, given any linearly independent subset X of V, clearly $X \subseteq X$, so indeed $X \leq X$. Next, if X and Y are two linearly independent subsets of V and if $X \leq Y$ and $Y \leq X$, this means that $X \subseteq Y$ and $Y \subseteq X$, so indeed $X = Y$. Finally, if $X \leq Y$ and $Y \leq Z$ for three linearly independent subsets X, Y, and Z of V, then this means that $X \subseteq Y \subseteq Z$, i.e., $X \subseteq Z$, so indeed $X \leq Z$.

Our strategy will be to first establish that S has a maximal element with respect to this partial order, and then to show that this maximal element must be a basis for V.

Given any chain T in S (recall that this means that T consists of linearly independent subsets of S with the property that if X and Y are in T, then either $X \subseteq Y$ or $Y \subseteq X$), we will show that T has an upper bound in S. Write K for the union of all linearly independent subsets X that are contained in T. We claim that K is an upper bound for T. Let us first show that K is a linearly independent subset of V. By Definition 3.22 of Chapter 3, we need to show that every finite subset of K is linearly independent. Given any finite set of vectors $v_1, \ldots, v_n$ from K, note that each v_i must live in some linearly independent subset X_i in the chain T. Since T is a chain, the subsets in T are linearly ordered (this is where we use the defining property that the elements of a chain are linearly ordered!), we must have $X_{i_1} \leq X_{i_2} \leq \cdots \leq X_{i_n}$ for some permutation $(i_1, i_2, \ldots, i_n)$ of the integers $(1, 2, \ldots n)$. Thus, all the vectors $v_1, \ldots, v_n$ belong to X_{i_n}. But since $X_{i,n}$ is a linearly independent set, Definition 3.22 of Chapter 3 implies that the vectors v_1, $\ldots$, v_n must be linearly independent! Since this is true for any finite set of vectors in K, we find that K is a linearly independent set. In particular, K is in S.

Now note that given any linearly independent subset X contained in the chain T, we have $X \subseteq K$ by the very definition of K, so by definition of the order relation, $X \leq K$. This shows that indeed T has an upper bound in S.

By Zorn's Lemma, S has a maximal element, call it B. We will show that B must be a basis of V. Since B is already linearly independent, we only need to show that B spans V. So let v be any nonzero vector in V: we need to show that v can be written as a linear combination of elements of B. If v is already in B, there is nothing to prove (why?). If v is not in B, $B \cup \{v\}$ must be linearly dependent, otherwise, $B \cup \{v\}$ would be a linearly independent subset of V strictly containing B, violating the maximality of B. Thus, there exists a relation $f_0 v + f_1 b_1 + f_2 b_2 + \cdots f_k b_k = 0$ for some scalars f_0, f_1, $\cdots$, f_k (not all zero), and some vectors b_1, b_2, $\cdots$, b_k of B. Notice that $f_0 \neq 0$, since otherwise, our relation would read $f_1 b_1 + f_2 b_2 + \cdots f_k b_k = 0$ (with not all f_i equal to zero), which is impossible since the b_i are in B and B is a linearly independent set. Therefore, we can divide by f_0 to find $v = (-f_1/f_0)b_1 + (-f_2/f_0)b_2 + \cdots + (-f_k/f_0)b_k$. Hence v can be written as a linear combination of elements of B, so B spans V.

Thus, B is a basis of V. □

Remarks on Proposition 3.37, Chapter 3: Shrinking infinite spanning sets down to a basis The proof that any spanning set of V can be shrunk to basis, even when V is infinite-dimensional, involves a modification of the proof of Theorem B.7.

Let us use Σ to denote the given spanning set of V, and as in the proof of Theorem B.7, let S denote the set of all linearly independent sets of V *that are contained in S*. (The italicized condition is where we depart from the proof of

Theorem B.7.) Note that S is not empty, since Σ is nonempty (recall V is not the trivial space), and therefore, for any nonzero $v \in \Sigma$, $\{v\} \subseteq \Sigma$ will be a linearly independent set, so $\{v\}$ will be an element of S.

Now impose the same partial order on S as in the proof of Theorem B.7: $X \leq Y$ if and only if $X \subseteq Y$ for two sets X and Y in S. Argue exactly as in that proof that S must have a maximal element. (Note that if T is a chain in S, then K, the union of all the sets contained in T, will also be contained in Σ, since every set in T is contained in Σ.) Let B be a maximal element of S. (Note that by construction $B \subseteq \Sigma$.) The claim is that B is a basis for V.

To prove this it is of course sufficient to prove that B spans V since B is already linearly independent. For this, we claim that it is sufficient to show that *every vector in* Σ is expressible as a linear combination of elements of B. For, assume that we have shown this. Then, given any vector $v \in V$, first write it as $v = f_1u_1 + \cdots + f_nu_n$ for suitable vectors $u_i \in \Sigma$ and scalars f_i, invoking the fact that Σ spans V. Next, since we would have shown that every vector in Σ is expressible as a linear combination of elements in B, we find that each u_i is expressible as $u_i = f_{i,1}b_{i,1} + \cdots + f_{i,n_i}b_{i,n_i}$ for some vectors $b_{i,j} \in B$ and scalars $f_{i,j}$. Substituting these expressions for each u_i into the expression above for v, we find that v is expressible as a linear combination of elements of B, i.e., that B spans V.

To show that every vector in Σ is expressible as a linear combination of elements of B, assume that some $u \in \Sigma$ is not expressible as a linear combination of elements of B. Then, exactly as in the proof of Proposition 3.49 (see how we showed $C_1 = C \cup \{v_{t+1}\}$ must be linearly independent), we would find that $B \cup \{u\}$ is linearly independent. But this contradicts the maximality of B! Hence every vector in Σ must be expressible as a linear combination of elements of B, which means that B must be a basis. Since $B \subseteq \Sigma$, we have succeeded in shrinking Σ down to a basis.

Remarks on Proposition 3.49, Chapter 3: the general case The proof of this proposition when V is not assumed to be finite-dimensional involves just a minor modification of the proof of Theorem B.7. What we need to show is that there is a maximal linearly independent subset B of V *that contains C.* Then, exactly as in the proof of Theorem B.7, this maximal linearly independent set would be a basis of V, and of course, it would have been chosen so as to contain C. To show the existence of B, we need to consider the set S of all linearly independent subsets of V *that contain* C. One would impose a partial order on this set exactly as in the proof of Theorem B.7. Once again, S, with this partial order, will turn out to satisfy the extra hypothesis of Zorn's Lemma, and will hence have a maximal element. That maximal element would be our desired maximal linearly independent subset of V that contains C.

Appendix C

GNU Free Documentation License

Preamble

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "**Document**", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "**you**". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "**Modified Version**" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "**Secondary Section**" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "**Invariant Sections**" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "**Cover Texts**" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "**Transparent**" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "**Opaque**".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "**Title Page**" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

The "**publisher**" means any person or entity that distributes copies of the Document to the public.

A section "**Entitled XYZ**" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "**Acknowledgements**", "**Dedications**", "**Endorsements**", or "**History**".) To "**Preserve the Title**" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.

C. State on the Title page the name of the publisher of the Modified Version, as the publisher.

D. Preserve all the copyright notices of the Document.

E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.

F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.

G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.

H. Include an unaltered copy of this License.

I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.

J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.

L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.

N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.

O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties—for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number.

Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled “History” in the various original documents, forming one section Entitled “History”; likewise combine any sections Entitled “Acknowledgements”, and any sections Entitled “Dedications”. You must delete all sections Entitled “Endorsements”.

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an “aggregate” if the copyright resulting from the compilation is not used to limit the legal rights of the compilation’s users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document’s Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled “Acknowledgements”, “Dedications”, or “History”, the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, or distribute it is void, and will automatically terminate your rights under this License.

However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means, this is the first time you have received notice of violation of this License (for any work) from that copyright holder, and you cure the violation prior to 30 days after your receipt of the notice.

Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, receipt of a copy of some or all of the same material does not give you any rights to use it.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See `http://www.gnu.org/copyleft/`.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation. If the Document specifies that a proxy can decide which future versions of this License can be used, that proxy's public statement of acceptance of a version permanently authorizes you to choose that version for the Document.

11. RELICENSING

"Massive Multiauthor Collaboration Site" (or "MMC Site") means any World Wide Web server that publishes copyrightable works and also provides prominent facilities for anybody to edit those works. A public wiki that anybody can edit is an example of such a server. A "Massive Multiauthor Collaboration" (or "MMC") contained in the site means any set of copyrightable works thus published on the MMC site.

"CC-BY-SA" means the Creative Commons Attribution-Share Alike 3.0 license published by Creative Commons Corporation, a not-for-profit corporation with a principal place of business in San Francisco, California, as well as future copyleft versions of that license published by that same organization.

"Incorporate" means to publish or republish a Document, in whole or in part, as part of another Document.

An MMC is "eligible for relicensing" if it is licensed under this License, and if all works that were first published under this License somewhere other than this MMC, and subsequently incorporated in whole or in part into the MMC, (1) had no cover texts or invariant sections, and (2) were thus incorporated prior to November 1, 2008.

The operator of an MMC Site may republish an MMC contained in the site under CC-BY-SA on the same site at any time before August 1, 2009, provided the MMC is eligible for relicensing.

ADDENDUM: How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

> Copyright © YEAR YOUR NAME. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with … Texts." line with this:

> with the Invariant Sections being LIST THEIR TITLES, with the Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.

Index