Foundations of Nonlinear Algebra

John Perry
University of Southern Mississippi
john.perry@usm.edu
http://www.math.usm.edu/perry/

Table of Contents

Reference sheet for notation

$[r]$	the element $r + n\mathbb{Z}$ of $\mathbb{Z}_n$
$\langle g \rangle$	the group (or ideal) generated by g
A_3	the alternating group on three elements
$A \triangleleft G$	for G a group, A is a normal subgroup of G
$A \triangleleft R$	for R a ring, A is an ideal of R
$[G, G]$	commutator subgroup of a group G
$[x, y]$	for x and y in a group G, the commutator of x and y
$\mathrm{Conj}_a(H)$	the group of conjugations of H by a
$\mathrm{conj}_g(x)$	the automorphism of conjugation by g
D_3	the symmetries of a triangle
$d \mid n$	d divides n
$\deg f$	the degree of the polynomial f
D_n	the dihedral group of symmetries of a regular polygon with n sides
$D_n(\mathbb{R})$	the set of all diagonal matrices whose values along the diagonal is constant
$d\mathbb{Z}$	the set of integer multiples of d
$f(G)$	for f a homomorphism and G a group (or ring), the image of G
$\mathbb{F}(\alpha)$	field extension of $\mathbb{F}$ by $alpha$
$\mathrm{Frac}(R)$	the set of fractions of a commutative ring R
F_S	the set of all functions mapping S to itself
G/A	the set of left cosets of A
$G\backslash A$	the set of right cosets of A
gA	the left coset of A with g
$G \cong H$	G is isomorphic to H
$\mathrm{GL}_m(\mathbb{R})$	the general linear group of invertible matrices
$\prod_{i=1}^{n} G_i$	the ordered n-tuples of G_1, G_2, …, G_n
$G \times H$	the ordered pairs of elements of G and H
g^z	for G a group and $g, z \in G$, the conjugation of g by z, or zgz^{-1}
$H < G$	for G a group, H is a subgroup of G
$\ker f$	the kernel of the homomorphism f
$\mathrm{lcm}(t, u)$	the least common multiple of the monomials t and u
$\mathrm{lm}(p)$	the leading monomial of the polynomial p
$\mathrm{lv}(p)$	the leading variable of a linear polynomial p
$\mathbb{M}$	the set of monomials in one variable
$\mathbb{M}_n$	the set of monomials in n variables
$N_G(H)$	the normalizer of a subgroup H of G
$\mathbb{N}$	the natural numbers $\{0, 1, 2, \ldots\}$
$\mathbb{N}^+$	positive integers
Ω_n	the nth roots of unity; that is, all roots of the polynomial $x^n - 1$
$\mathrm{ord}(x)$	the order of x
$P(S)$	the power set of S
Q_8	the group of quaternions
R/A	for R a ring and A an ideal subring of R, R/A is the quotient ring of R with respect to A

$\langle r_1, r_2, \ldots, r_m \rangle$	the ideal generated by $r_1, r_2, \ldots, r_m$
$R[x_1, x_2, \ldots, x_n]$	the ring of polynomials whose coefficients are in the ground ring R
S_n	the group of all permutations of a list of n elements
$S \times T$	the Cartesian product of the sets S and T
$\mathcal{T}_f$	the support of a polynomial f
T_f	the support of the polynomial f
$\mathrm{tts}(p)$	the trailing terms of p
$Z(G)$	centralizer of a group G
$\mathbb{Z}_n^*$	the set of elements of $\mathbb{Z}_n$ that are *not* zero divisors
$\mathbb{Z}/n\mathbb{Z}$	quotient group (resp. ring) of $\mathbb{Z}$ modulo the subgroup (resp. ideal) $n\mathbb{Z}$
$\mathbb{Z}$	integers
$\mathbb{Z}\left[\sqrt{-5}\right]$	the ring of integers, adjoin $\sqrt{-5}$
$\mathbb{Z}_n$	the quotient group $\mathbb{Z}/n\mathbb{Z}$

A few acknowledgements

These notes are inspired from some of my favorite algebra texts: [AF05, CLO97, HA88, KR00, vzGG99, Lau03, LP98, Rot06, Rot98]. (Believe it or not, that is *not* a comprehensive list.) They started out as notes to parallel [Lau03], but have since taken on a life of their own, and are now quite different. I have tried to cite a source when I followed a particular approach.

Thanks to the students who found typos, including (in no particular order) Jonathan Yarber, Kyle Fortenberry, Lisa Palchak, Ashley Sanders, Sedrick Jefferson, Shaina Barber, Blake Watkins, Kris Katterjohn, Taylor Kilman, Eric Gustaffson, Patrick Lambert, and others. Special thanks go to my graduate student Miao Yu, who endured the first drafts of Chapters 7, 8, and 11.

Rogério Brito of Universidade de São Paolo made several helpful comments, found some nasty errors, and suggested some of the exercises.

I have been lucky to have had great algebra professors; in chronological order:

- Vanessa Job at Marymount University;
- Adrian Riskin at Northern Arizona University;
- and at North Carolina State University:
    - Kwangil Koh,
    - Hoon Hong,
    - Erich Kaltofen,
    - Michael Singer, and
    - Agnes Szanto.

Boneheaded innovations of mine that looked good at the time but turned out bad in practice were entirely my idea. *This is not a peer-reviewed text*, which is why you have a supplementary text in the bookstore.

The following software helped prepare these notes:

- Sage 3.*x* and later[Ste08];
- Lyx [Lyx] (and therefore LaTeX [Lam86, Grä04] (and therefore TeX [Knu84])), along with the packages
    - `cc-beamer` [Pip07],
    - `hyperref` [RO08],
    - $\mathcal{AMS}$-LaTeX[Soc02],
    - `mathdesign` [Pic06],
    - `thmtools`, and
    - `algorithms` (modified slightly from the version released 2006/06/02) [Bri]; and
- Inkscape [Bah08].

I've likely forgotten some other non-trivial resources that I used. Let me know if another citation belongs here.

My wife forebore a number of late nights at the office (or at home) as I worked on these.

Ad maiorem Dei gloriam.

Preface

A two-semester sequence on modern algebra typically introduces students to the fundamental ideas in group and ring theory. Lots of textbooks do a good job of that, and I always recommend one or more to my classes.

However, most such books seem targeted at students with a strong mathematical background. Like many instructors these days, I encounter many students with a weaker background in mathematical thinking. These students arrive with enthusiasm, and find the material fascinating. Some may possess the great combination of talent, enthusiasm, and preparation, but most lack at least one of those three.

It wasn't until I taught algebra that I realized just *how many new ideas* a student meets, in contrast to other courses at the undergraduate level. Students seem to find algebra an "odd beast"; at most institutions, I suspect, only analysis is comparable. Unlike analysis, however, most every algebra text I've seen spends the first 50–100 pages *on material that is not algebra*. Authors have very, very good reasons for that; for example, the very concrete problems of number theory can motivate certain algebraic ideas. In my experience, however, students' unfamiliarity with number theory means we waste a lot of time and energy on information that isn't really *algebra*.

Desiring a mix of simplicity and utility, I decided to set out some notes that would throw the class into algebraic problems and ideas as soon as possible. As it happens, another interest of mine seems to have helped. Typically, an algebra text starts with groups, on account of their simplicity. Another option is to start with rings, on account of the familiarity of their operations. I've tried to marry the best of both worlds by starting with monoids, which are both simple and familiar. An added bonus is that one can introduce very deep notions, such as direct products, isomorphism, ideals (under another name), the Ascending Chain Condition, and even Hilbert Functions, in a fast, intuitive way that is not at all superficial.

As the notes diverged more and more from the textbooks I was using, I committed them to digital form, which allowed me to organize, edit, rearrange, modify, and extend them easily. By now, it is more or less an unofficial textbook.

Overview

These notes have two major parts: in one, we focus on an algebraic structure called a *group*; in the other, we focus on a special kind of group, a *ring*. They correspond roughly to a two-semester course in algebra.

In the first semester, I try to cover Chapters 1–5. Since a rigorous approach requires *some* sort of introduction, those chapters are preceded by a review of basic ideas you should have seen before – but only to set a foundation for what is to come.

We then move to monoids, relying on the natural numbers, matrices, and monomials as natural examples.[1] Monoids are not a popular way to start an algebra course, so much of that chapter is optional. However, a *brief* glance at monoids allows us to introduce prized ideas that we develop in much more depth with groups and rings, but in a context with which students are far more familiar.

Chapter 6, on number theory, serves as a bridge between the two main parts. Many books on algebra start with this material, I've pushed as much as I felt possible after group theory, so

[1] To some extent, I owe the idea of starting with monoids to a superb graduate-level text, [KR00].

that we can view number theory as an *application* of group theory. Ideally, the chapter on the RSA algorithm would provide a nice "bang" to end the first semester, but I haven't managed that in years, and even then it was too rushed. *Tempus fugit*, and all that.

In the second semester, we definitely cover Chapters 6 through 8, along with at least one of the later chapters. I include Chapter 12 for students who want to pursue a research project, and need an introduction that builds on what came before. As of this writing, some of those chapters still need major debugging, so don't take anything you read there too seriously.

It is not easy to jump around these notes. Not much of the material can be omitted. Within each chapter, many examples are used and reused; this applies to exercises, as well. I do try to concentrate on a few important examples, re-examining them in the light of each new topic. One consequence is that the presentation of groups depends on the introduction to monoids, and the presentation of rings depends on a thorough consideration of groups, which in turn depends on at least some of the material on monoids. On the other hand, most of the material on monoids can be postponed until after groups. In the first semester, I usually omit solvable groups (Section 3.6) and groups of automorphisms (Section 4.4).

To the student

Most people find advanced algebra quite difficult. There is no shame in that; I find it difficult, too. I'm a little unusual in that I find it difficult *but still love it.* No other branch of mathematics ever appealed to me the way algebra did. I sometimes joke that I earned a Ph.D. only because I was too dumb to quit.

I want you to learn algebra, and to see why its ideas have excited not just me, but thousands of others, most of whom are much, much smarter than me. My experiences teaching this class motivate the following remarks.

How to succeed at algebra

There are certain laws of success in algebra, which I'm pretty sure apply not only to me, but to everyone out there.

1. You won't "get it" right away.

One of the big shocks to students who study algebra is that they can't apply the same strategy that they have applied successfully in other mathematical courses. In many undergraduate textbooks, each section introduces some property or technique, *maybe* explains why it works, then illustrates an application of the property, asking you to repeat it on some problems. At most, they ask you to adapt the method used to apply the property.

Algebra isn't like that. The problems almost always require you to use some properties to derive or explain *other properties!* That requires a new style of solving problems, one where you develop the method of solution. Typically, this takes the form of a proof, a short explanation as to why some property is true. You're not really used to that, and you may even have thought that you were studying mathematics precisely to *escape* writing! Sorry!

2. Anything worth doing requires effort and time.

It will take more than 30 minutes per week to succeed with the homework problems in this class. It may well take more than 30 minutes *per problem*! Don't let that intimidate you.

To some extent, modern culture has left you ill-prepared for this class. Modern technology can execute in moments tasks that were once impossible, such as speaking across the ocean. Books and films tend to portray the process of discovery and invention as if it were also quick, but the

reality is far different. The people who developed these technologies did not do so with a snap of their fingers! They spent years, if not their entire lives, trying to solve difficult and important problems.

The same is true with mathematics. For example, the material covered in Chapter 9 is commonly called "Galois Theory". It's entirely possible that the reason it isn't called "Ruffini Theory" is that Paolo Ruffini, who discovered many of its principles, couldn't get anyone to take his notions seriously. None of the leading minds of his day would talk with him about it, which meant that he couldn't see easily the flaws in his work, let alone correct, develop, and deepen them. For that matter, the accomplishments of Evariste Galois were not recognized until decades after he stayed up all night before a duel to to write down ideas that had fermented in his mind. Eventually, they would inebriate the world with understanding.

Algebra is worth spending time on. Don't try to do it on the cheap, devoting only a few spare moments here and there.

3. You actually have to know the definitions.

I strongly suggest writing every definition down on a notecard, and creating flashcards to quiz yourself on basic definitions.

Most people no longer seem to think the meanings of words matter. This manifests itself even in mathematics, where students who walk around with A's in high school and college Calculus can't tell you the definition of a limit or a derivative! How do you earn a top score without learning what the fundamental ideas mean?

By its nature, you can't even understand the basic problems in algebra unless you know the meaning of the terms. I can talk myself blue in the face while helping students, but a student who can't state the definition of the technical words used in the problem will not understand the problem, let alone how to find the solution.

4. Don't be afraid to make a fool of yourself.

The only "dumb" questions in this class are the ones where someone asks me what a word means. That's a *definition;* if you can't be bothered to look it up, I can't be bothered to tell you.

All other questions pertinent to this material really are fair game. As I wrote above, I succeeded only because I was too dumb to quit. Every now and then some student works up the courage to ask a question she's sure will make her look stupid, but it's pretty much always a very good question. Often enough, I have to correct something stupid I said.

So, ask away. With any luck, you'll end up embarrassing *me*.

Ways these notes try to help you succeed

I have tried to present a large number of "concrete" examples. Some examples are more important than others, and you will notice that I return frequently to a few important objects. I am not unique in emphasizing these examples; most textbooks in algebra emphasize at least some of them.

Spend time familiarizing yourself with these examples. Students often make the mistake of thinking that the purpose of the examples is to show them how to solve the exercises. While that may be true in a textbook on, say, calculus, linear algebra, or differential equations, it can be a fatal assumption in non-linear algebra. Here, the purpose of the examples is to illustrate the objects and ideas that you have to understand in order to solve to the exercises. I suspect these notes are unusual in dedicating an entire section to the roots of unity, but if not, that only proves how important this example is.

I could say the same about the exercises. Even if an exercise isn't assigned, and you choose not to solve it, familiarize yourself with the statement of the exercise. A significant proportion of the exercises build on examples or even exercises that appear earlier.

An approach I've used that seems uncommon, if not unique, is the presence of fill-in-the-blank exercises. I've designed these with two goals in mind. First, most algebra students are overwhelmed by the rush of ideas and objects — and have very little experience solving theoretical problems, where the "answer" is already given, and the "method of solution" is what they must produce! So, I've taken some of the problems that seem to present students with more difficulty, and sketched a proof where nearly every statement lacks either a phrase or a justification; students need merely fill the hole. Second, even when students have a basic understanding of the proof of a statement, they typically write a very poor proof. The fill-in-the-blank problems are meant to illustrate what a correct proof looks like — although, in my attempt to leave no stone unturned, they may seem pedantic.

Some interesting problems

We'd like to motivate this study of algebra with some problems that we hope you will find interesting. Although we eventually solve them in this text, it might surprise you that in this class, we're interested not in the solutions, but in *why the solutions work*. I could in fact tell you how to to solve them right here, and we'd be done soon enough; on to vacation! But then you wouldn't have learned what makes this course so beautiful *and important.* It would be like walking through a museum with me as your tour guide. I can summarize the purpose of each displayed article, but you can't learn enough in a few moments to appreciate it in the same way as someone familiar with fundamental notions in that field. The purpose of this course is to familiarize you with fundamental notions of non-linear algebra.

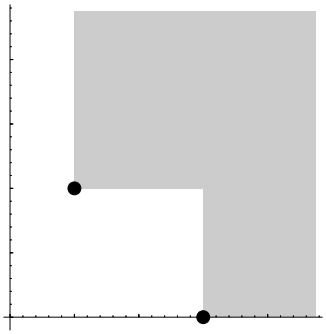
Still, let's take a preliminary stroll through the museum, and consider these exhibits.

Nimfinity

Consider the following game, which generalizes the ancient game of Nim. The playing board is the first quadrant of the x-y axis. Players take turns doing the following:

1. Choose some point (a,b) such that a and b are both integers, and that does not yet lie in a shaded region been shaded.
2. Shade the region of points (c,d) such that $c \geq a$ and $d \geq b$.

The winner is the player who forces the last move. In the example shown below, the players have chosen the points $(1,2)$ and $(3,0)$.



Questions:

- Must the game end? or is it possible to have a game that continues indefinitely? Is this true even if we use an n-dimensional playing board, where $n > 2$? And if so, why?
- Is there a way to count the number of moves remaining, even when there are infinitely many moves?
- Suppose that for each nonnegative integer d, you are forbidden from picking a certain number of points (a,b) such that $a+b=d$. It doesn't matter what the points are, only that you may choose a certain number, and no more. Is there a strategy to win?

We answer some of these questions at the end of Chapter 1.

A card trick

Take twelve cards. Ask a friend to choose one, to look at it without showing it to you, then to shuffle them thoroughly. Arrange the cards on a table face up, in rows of three. Ask your friend what column the card is in; call that number α.

Now collect the cards, making sure they remain in the same order as they were when you dealt them. Arrange them on a a table face up again, in rows of four. It is essential that you

maintain the same order; the first card you placed on the table in rows of three must be the first card you place on the table in rows of four; likewise the last card must remain last. The only difference is where it lies on the table. Ask your friend again what column the card is in; call that number β.

In your head, compute $x = 4\alpha - 3\beta$. If x does not lie between 1 and 12 inclusive, add or subtract 12 until it is. Starting with the first card, and following the order in which you laid the cards on the table, count to the xth card. This will be the card your friend chose.

Mastering this trick takes only a little practice. *Understanding* it requires quite a lot of background! We get to it in Chapter 6.

Internet commerce

Let's go shopping!!! No, wait. That's too inconvenient. *Let's go shopping… online!!!* Before the online company sends you your product, they'll want payment. This requires you to submit some sensitive information, namely, your credit card number. Once you submit that number, it will bounce happily around a few computers on its way to the company's server. Some of those computers might be in foreign countries. (It's quite possible. Don't ask.) Any one of those machines could have a snooper. How can you communicate the information *securely?*

The solution is *public-key cryptography*. The bank's computer tells your computer how to send it a message. It supplies a special number used to encrypt the message, called an *encryption key*. Since the bank broadcasts this in the clear over the internet, anyone in the world can see it. What's more, anyone in the world can look up the method used to decrypt the message.

You might wonder, *How on earth is this secure?!?* Public-key cryptography works because there's the *decryption key* remains with the company, hopefully secret. Secret? Whew! …or so you think. A snooper could reverse-engineer this key using a "simple" mathematical procedure that you learned in grade school: factoring an integer into primes, like, say, $21 = 3 \cdot 7$.

*How on earth is **this** secure?!?* Although the procedure is "simple", the size of the integers in use now is about 40 digits. Believe it or not, even a 40 digit integer takes even a computer far too long to factor! So your internet commerce is completely safe. For now.

Factorization

How can we factor polynomials like $p(x) = x^6 + 7x^5 + 19x^4 + 27x^3 + 26x^2 + 20x + 8$? There are a number of ways to do it, but the most efficient ways involve *modular arithmetic*. We discuss the theory of modular arithmetic later in the course, but for now the general principle will do: pretend that the only numbers we can use are those on a clock that runs from 1 to 51. As with the twelve-hour clock, when we hit the integer 52, we reset to 1; when we hit the integer 53, we reset to 2; and in general for any number that does not lie between 1 and 51, we divide by 51 and take the remainder. For example,

$$20 \cdot 3 + 8 = 68 \rightsquigarrow 17.$$

How does this help us factor? When looking for factors of the polynomial p, we can simplify multiplication by working in this modular arithmetic. This makes it easy for us to reject many possible factorizations before we start. In addition, the set $\{1, 2, \ldots, 51\}$ has many interesting properties under modular arithmetic that we can exploit further.

Conclusion

Non-linear algebra deals with interesting and important problems, while retaining a deep, theoretical character: we wonder more about *why things are true* than about *how we can do things*. Algebraists can at times be concerned more with elegance and beauty than applicability and efficiency. You may be tempted on many occasions to ask yourself the point of all this abstraction and theory. *Who needs this stuff?*

Keep the examples above in mind; they show that algebra is not only useful, but necessary. Its applications have been profound and broad. Eventually you will see how algebra addresses the problems above; for now, you can only start to imagine.

The class "begins" here. Wipe your mind clean: unless it says otherwise here or in the following pages, everything you've learned until now is suspect, and cannot be used to explain anything. You should adopt the Cartesian philosophy of doubt.[2]

[2]Named after the mathematician and philosopher René Descartes, who inaugurated modern philosophy and claimed to have spent a moment wondering whether he even existed. *Cogito, ergo sum* and all that.

Chapter 0:
Foundations

This chapter re-presents ideas you have seen before, but may not have acquired comfort with them. We will emphasize precise definitions and rely heavily on deductive precision, rather than intuitive vagueness — sometimes called "hand waving". Too often, people speak vaguely to each other, and words contain different meanings for different people.

Do not mistake this dismissal for disdain; intuition is *very* important in the problem solving process, and you *will* have to develop some intuition to succeed with this material. We *will* emphasize intuitive notions as we introduce new terms. However, you should already have an intuitive familiarity with most of the ideas presented in this section, so any weaknesses you have will be with your ability to *deduce* a solution in *precise* words.

Gauss, no slouch in either mathematics or science, felt that mathematics is not merely a science, but the queen of the sciences. Good science depends on clarity and reproducibility. This can be hard going for a while, but if you accept it and engage it, you will find it very rewarding.

0.1: Sets and relations

Let's start with some general tools of mathematics that you should have seen before now.

Sets

The most fundamental object in mathematics is the **set**. Sets can possess a property called **inclusion** when all the elements of one set are also members of the other. More commonly, people say that the set A is a **subset** of the set B if every element of A is also an element of B. If A is a subset of B but not equal to B, we say that A is a **proper subset** of B. All sets have the **empty set** as a subset; some people write the empty set as $\{\}$, but we will use $\emptyset$, which is also common.

Notation 0.1. If A is a subset of B, we write $A \subseteq B$. If A is a proper subset, we can still write $A \subsetneq B$, but if we want to emphasize that they are not equal, we write $A \subsetneq B$.

You should recognize these sets:

- the **positive integers**, $\mathbb{N}^+ = \{1, 2, 3, \ldots\}$, also called the **counting numbers**,
- the **natural numbers**, $\mathbb{N} = \{0, 1, 2, \ldots\}$, and
- the **integers**, $\mathbb{Z} = \{\ldots, -2, 1, 0, 1, 2, \ldots\}$, which extend $\mathbb{N}^+$ to "complete" subtraction.

You are already familiar with the intuitive motivation for these numbers and also how they are applied, so we won't waste time rehashing that. Instead, let's spend time re-presenting some basic ideas of sets, especially the integers.

Notation 0.2. Notice that both $\mathbb{N}^+ \subseteq \mathbb{N} \subseteq \mathbb{N}$ and $\mathbb{N}^+ \subsetneq \mathbb{N} \subseteq \mathbb{Z}$ are true.

We can put sets together in several ways.

Definition 0.3. Let S and T be two sets. The **Cartesian product of S and T** is the set of ordered pairs

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

The **union** of S and T is the set

$$S \cup T = \{x : x \in S \text{ or } x \in T\},$$

the **intersection** of S and T is the set

$$S \cap T = \{x : x \in S \text{ and } x \in T\},$$

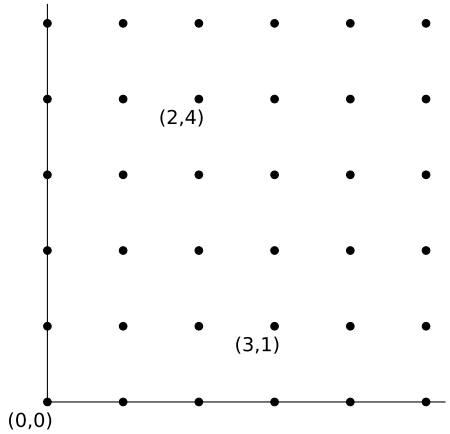
and the **difference** of S and T is the set

$$S \backslash T = \{x : x \in S \text{ and } x \notin T\}.$$

Example 0.4. Suppose $S = \{a, b\}$ and $T = \{x+1, y-1\}$. By definition,

$$S \times T = \{(a, x+1), (a, y-1), (b, x+1), (b, y-1)\}.$$

Example 0.5. If we let $S = T = \mathbb{N}$, then $S \times T = \mathbb{N} \times \mathbb{N}$, the set of all ordered pairs whose entries are natural numbers. We can visualize this as a **lattice**, where points must have integer co-ordinates:



Let $\mathcal{B} = \{S, T, Z\}$ where

- S is the set of positive integers,
- T is the set of negative integers, and
- $Z = \{0\}$.

The elements of $\mathcal{B}$ are **disjoint** sets, by which we mean that they have nothing in common. In addition, the elements of $\mathcal{B}$ **cover** $\mathbb{Z}$, by which we mean that their union produces the entire set of integers. This phenomenon, where a set can be described the union of smaller, disjoint sets, is important enough to highlight with a definition.

Definition 0.6. Suppose that A is a set and $\mathcal{B}$ is a family of subsets of A, called **classes**. We say that $\mathcal{B}$ is a **partition** of A if

- the classes **cover** A: that is, $A = \bigcup_{B \in \mathcal{B}} B$; and
- distinct classes are disjoint: that is, if $B_1, B_2 \in \mathcal{B}$ are distinct ($B_1 \neq B_2$), then $B_1 \cap B_2 = \emptyset$.

The next section introduces a very important kind of partition.

Relations

We often want to describe a relationship between two elements of two or more sets. It turns out that this relationship is also a set. Defining it this way can seem unnatural at first, but in the long run, the benefits far outweigh the costs.

Definition 0.7. Any subset of $S \times T$ is **relation on the sets S and T**. A **function** is any relation f such that $(a,b) \in f$ implies $(a,c) \notin f$ for any $c \neq b$. An **equivalence relation on** S is a subset R of $S \times S$ that satisfies the properties

reflexive:	for all $a \in S$, $(a,a) \in R$;
symmetric:	for all $a, b \in S$, if $(a,b) \in R$ then $(b,a) \in R$; and
transitive:	for all $a, b, c \in S$, if $(a,b) \in R$ and $(b,c) \in R$ then $(a,c) \in R$.

Notation 0.8. Even though relations and functions are sets, we usually write them in the manner to which you are accustomed.

- We typically denote relations that are not functions by symbols such as $<$ or $\subseteq$. If we want a generic symbol for a relation, we usually write $\sim$.
- If $\sim$ is a relation, and we want to say that a and b are members of the relation, we write not $(a,b) \in \sim$, but $a \sim b$, instead. For example, in a moment we will discuss the subset relation $\subseteq$, and we always write $a \subseteq b$ instead of "$(a,b) \in \subseteq$".
- We typically denote functions by letters, typically f, g, or h, or sometimes the Greek letters, η, φ, ψ, or μ. Instead of writing $f \subseteq S \times T$, we write $f : S \to T$. If f is a function and $(a,b) \in f$, we write $f(a) = b$.
- The definition and notation of relations and sets imply that we can write $a \sim b$ and $a \sim c$ for a relation $\sim$, but we cannot write $f(a) = b$ and $f(a) = c$ for a function f.

Example 0.9. Define a relation $\sim$ on $\mathbb{Z}$ in the following way. We say that $a \sim b$ if $ab \in \mathbb{N}$. Is this an equivalence relation?

Reflexive? Let $a \in \mathbb{Z}$. By properties of arithmetic, $a^2 \in \mathbb{N}$. By definition, $a \sim a$, and the relation is reflexive.

Symmetric? Let $a, b \in \mathbb{Z}$. Assume that $a \sim b$; by definition, $ab \in \mathbb{N}$. By the commutative property of arithmetic, $ba \in \mathbb{N}$ also, so $b \sim a$, and the relation is reflexive.

Transitive? Let $a, b, c \in \mathbb{Z}$. Assume that $a \sim b$ and $b \sim c$. By definition, $ab \in \mathbb{N}$ and $bc \in \mathbb{N}$. I could argue that $ac \in \mathbb{N}$ using the trick

$$ac = \frac{(ab)(bc)}{b^2},$$

and pointing out that ab, bc, and b^2 are all natural, which suggests that ac is also natural. However, this argument contains a fatal flaw. Do you see it?

It lies in the fact that we don't know whether $b = 0$. If $b \neq 0$, then the argument above works just fine, but if $b = 0$, then we encounter division by 0, which you surely know is not allowed! (If you're not sure *why* it is not allowed, fret not. We explain this in a moment.)

This apparent failure should not discourage you; in fact, it gives us the answer to our original question. We asked if $\sim$ was an equivalence relation. In fact, *it is not,* and what's more, it illustrates an important principle of mathematical study. Failures like this should prompt you to explore whether you've found an unexpected avenue to answer a question. In this case, the fact that $a \cdot 0 = 0 \in \mathbb{N}$ for any $a \in \mathbb{Z}$ implies that $1 \sim 0$ and $-1 \sim 0$. However, $1 \not\sim -1$! The relation is *not* transitive, so it *cannot* be an equivalence relation!

Binary operations

Another important relation is defined by an operation.

Definition 0.10. Let S and T be sets. A **binary operation from S to T** is a function $f : S \times S \to T$. If $S = T$, we say that f is a binary operation **on** S. A binary operation f on S is **closed** if $f(a, b)$ is defined for all $a, b \in S$.

Example 0.11. Addition of the natural numbers is a function, $+ : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$; the sentence, $2 + 3 = 5$ can be thought of as $+(2, 3) = 5$. Hence, addition is a binary operation on $\mathbb{N}$. Addition is defined for all natural numbers, so it is closed.

Subtraction of natural numbers can be viewed as a function, as well: $- : \mathbb{N} \times \mathbb{N} \to \mathbb{Z}$. However, while subtraction is a binary operation, it is not closed, since it is not "on $\mathbb{N}$": the range ($\mathbb{Z}$) is not the same as the domain ($\mathbb{N}$). This is the reason we need the integers: they "close" subtraction of natural numbers.

In each set described above, you can perform arithmetic: add, subtract, multiply, and (in most cases) divide. We need to make the meaning of these operations precise.[3]

Addition of positive integers is defined in the usual way: it counts the number of objects in the union of two sets with no common element. To obtain the integers $\mathbb{Z}$, we extend $\mathbb{N}^+$ with two kinds of new objects.

- 0 is an object such that $a + 0 = a$ for all $a \in \mathbb{N}^+$ (the *additive identity*). This models the union of a set of a objects and an empty set.
- For any $a \in \mathbb{N}^+$, we define its *additive inverse*, $-a$, as an object with the property that $a + (-a) = 0$. This models *removing* a objects from a set of a objects, so that an empty set remains.

Since $0 + 0 = 0$, we are comfortable deciding that $-0 = 0$. To add with negative integers, let $a, b \in \mathbb{N}^+$ and consider $a + (-b)$:

- If $a = b$, then substitution implies that $a + (-b) = b + (-b) = 0$.
- Otherwise, let A be any set with a objects.

[3] We will not make the meanings as precise as possible; at this level, some things are better left to intuition. For example, I will write later, "If I can remove a set with b objects from [a set with a objects]…" What does this mean? We will not define this, but leave it to your intuition.

- If I can remove a set with b objects from A, and have at least one object left over, let $c \in \mathbb{N}^+$ be the number of objects left over; then we define $a + (-b) = c$.
- If I *cannot* remove a set with b objects from A, then let $c \in \mathbb{N}^+$ be the smallest number of objects I would need to add to A so that I could remove b objects. This satisfies the equation $a + c = b$; we then define $a + (-b) = -c$.

For multiplication, let $a \in \mathbb{N}^+$ and $b \in \mathbb{Z}$.

- $0 \cdot b = 0$ and $b \cdot 0 = 0$;
- $a \cdot b$ is the result of adding a copies of b, or

$$\underbrace{(((b+b)+b)+\cdots b)}_{a};$$

  and
- $(-a) \cdot b = -(a \cdot b)$.

We won't bother with a proof, but we assert that such an addition and multiplication are defined for all integers, and satisfy the following properties:

- $a + b = b + a$ and $ab = ba$ for all $a, b \in \mathbb{N}^+$ (the *commutative property*).
- $a + (b + c) = (a + b) + c$ and $(ab)c = a(bc)$ for all $a, b, c \in \mathbb{N}^+$ (the *associative property*).
- $a(b + c) = ab + ac$ for all $a, b, c \in \mathbb{Z}$ (the *distributive property*).

Notation 0.12. For convenience, we usually write $a - b$ instead of $a + (-b)$.

We have not yet talked about the additive inverses of additive inverses. Suppose $b \in \mathbb{Z}\backslash\mathbb{N}$; by definition, b is an additive inverse of some $a \in \mathbb{N}^+$, $a + b = 0$, and $b = -a$. Since we want addition to satisfy the commutative property, we *must* have $b + a = 0$, which suggests that we can think of a as the additive inverse of b, as well! That is, $-b = a$. Written another way, $-(-a) = a$. This also allows us to define the **absolute value** of an integer,

$$|a| = \begin{cases} a, & a \in \mathbb{N}, \\ -a, & a \notin \mathbb{N}. \end{cases}$$

Orderings

We have said nothing about the "ordering" of the natural numbers; that is, we do not "know" yet whether 1 comes before 2, or vice versa. However, our definition of adding negatives has imposed a natural ordering.

Definition 0.13. For any two elements $a, b \in \mathbb{Z}$, we say that:

- $a \leq b$ if $b - a \in \mathbb{N}$;
- $a > b$ if $b - a \notin \mathbb{N}$;
- $a < b$ if $b - a \in \mathbb{N}^+$;
- $a \geq b$ if $b - a \notin \mathbb{N}^+$.

So $3 < 5$ because $5 - 3 \in \mathbb{N}^+$. Notice how the negations work: the negation of $<$ is *not* $>$.

Remark 0.14. Do not yet assume certain "natural" properties of these orderings. For example, it is true that if $a \leq b$, then either $a < b$ or $a = b$. But why? You can reason to it from the definitions given here, so you should do so.

More importantly, you cannot yet assume that if $a \leq b$, then $a + c \leq b + c$. You can reason to this property from the definitions, and you will do so in the exercises.

Some orderings enjoy special properties.

Definition 0.15. Let S be any set. A **linear ordering** on S is a relation $\sim$ where for any $a, b \in S$ one of the following holds:

$$a \sim b, a = b, \text{ or } b \sim a.$$

Suppose we define a relation on the subsets of a set S by inclusion; that is, $A \sim B$ if and only if $A \subseteq B$. This relation is *not* a linear ordering, since

$$\{a, b\} \not\subseteq \{c, d\}, \quad \{a, b\} \neq \{c, d\}, \quad \text{and} \quad \{c, d\} \not\subseteq \{a, b\}.$$

By contrast, the orderings of $\mathbb{Z}$ *are* linear.

Theorem 0.16. The relations $<$, $>$, $\leq$, and $\geq$ are linear orderings of $\mathbb{Z}$.

Our "proof" relies on some unspoken assumptions: in particular, the arithmetic on $\mathbb{Z}$ that we described before. Try to identify where these assumptions are used, because when you write your own proofs, you have to ask yourself constantly: Where am I using unspoken assumptions? In such places, either the assertion must be something accepted by the audience,[4] or you have to cite a reference your audience accepts, or you have to prove it explicitly. It's beyond the scope of this course to discuss these assumptions in detail, but you should at least try to find them.

Proof. We show that $<$ is linear; the rest are proved similarly.

Let $a, b \in \mathbb{Z}$. Subtraction is closed for $\mathbb{Z}$, so $b - a \in \mathbb{Z}$. By definition, $\mathbb{Z} = \mathbb{N}^+ \cup \{0\} \cup \{-1, -2, \ldots\}$. Since $b - a$ must be in one of those three subsets, let's consider each possibility.

- If $b - a \in \mathbb{N}^+$, then $a < b$.
- If $b - a = 0$, then recall that our definition of subtraction was that $b - a = b + (-a)$. Since $b + (-b) = 0$, reasoning on the meaning of natural numbers tells us that $-a = -b$, and thus $a = b$.
- Otherwise, $b - a \in \{-1, -2, \ldots\}$. By definition, $-(b - a) \in \mathbb{N}^+$. We know that $(b - a) + [-(b - a)] = 0$. It is not hard to show that $(b - a) + (a - b) = 0$, and reasoning on the meaning of natural numbers tells us again that $a - b = -(b - a)$. In other words, and thus $b < a$.

We have shown that $a < b$, $a = b$, or $b < a$. Since a and b were arbitrary in $\mathbb{Z}$, $<$ is a linear ordering. □

It should be easy to see that the orderings and their linear property apply to all subsets of $\mathbb{Z}$, in particular $\mathbb{N}^+$ and $\mathbb{N}$. That said, this relation behaves differently in $\mathbb{N}$ than it does in $\mathbb{Z}$.

Linear orderings are already special, but some are *extra* special.

Definition 0.17. Let S be a set and $\prec$ a linear ordering on S. We say that $\prec$ is a **well-ordering** if

Every nonempty subset T of S has a **smallest element** a;
that is, there exists $a \in T$ such that for all $b \in T$, $a \prec b$ or $a = b$.

[4] In your case, the *instructor* is the audience.

Example 0.18. The relation $<$ is *not* a well-ordering of $\mathbb{Z}$, because $\mathbb{Z}$ itself has no smallest element under the ordering.

Why not? Proceed by way of contradiction. Assume that $\mathbb{Z}$ has a smallest element, and call it a. Certainly $a-1 \in \mathbb{Z}$ also, but

$$(a-1)-a = -1 \notin \mathbb{N}^+,$$

so $a \not< a-1$. Likewise $a \neq a-1$. This contradicts the definition of a smallest element, so $\mathbb{Z}$ is not well-ordered by $<$.

We now assume, *without proof*, the following principle.

The relations $<$ and $\leq$ are well-orderings of $\mathbb{N}$.

That is, any subset of $\mathbb{N}$, ordered by these orderings, has a smallest element. This may sound obvious, but it is very important, and what is remarkable is that *no one can prove it.*[5] It is an assumption about the natural numbers. This is why we state it as a principle (or axiom, if you prefer). In the future, if we talk about the well-ordering of $\mathbb{N}$, we mean the well-ordering $<$.

One consequence of the well-ordering property is the following fact.

Theorem 0.19. Let $a_1 \geq a_2 \geq \cdots$ be a nonincreasing sequence of natural numbers. The sequence eventually stabilizes; that is, at some index i, $a_i = a_{i+1} = \cdots$.

Proof. Let $T = \{a_1, a_2, \ldots\}$. By definition, $T \subseteq \mathbb{N}$. By the well-ordering principle, T has a least element; call it b. Let $i \in \mathbb{N}^+$ such that $a_i = b$. The definition of the sequence tells us that $b = a_i \geq a_{i+1} \geq \cdots$. Thus, $b \geq a_{i+k}$ for all $k \in \mathbb{N}$. Since b is the *smallest* element of T, we know that $a_{i+k} \geq b$ for all $k \in \mathbb{N}$. We have $b \geq a_{i+k} \geq b$, which is possible only if $b = a_{i+k}$. Thus, $a_i = a_{i+1} = \cdots$, as claimed. □

Another consequence of the well-ordering property is the principle of:

Theorem 0.20 (Mathematical Induction)**.** Let P be a subset of $\mathbb{N}^+$. If P satisfies (IB) and (IS) where
(IB) $1 \in P$;
(IS) for every $i \in P$, we know that $i+1$ is also in P;
then $P = \mathbb{N}^+$.

There are several versions of mathematical induction that appear: generalized induction, strong induction, weak induction, etc. We present only this one as a theorem, but we use the others without comment.

Proof. Let $S = \mathbb{N}^+ \backslash P$. We will prove the contrapositive, so assume that $P \neq \mathbb{N}^+$. Thus $S \neq \emptyset$. Note that $S \subseteq \mathbb{N}^+$. By the well-ordering principle, S has a smallest element; call it n.

- If $n = 1$, then $1 \in S$, so $1 \notin P$. Thus P does not satisfy (IB).

[5] You might try to prove the well-ordering of $\mathbb{N}$ using induction. You would in fact succeed, but that requires you to assume induction. Why is induction true? In fact, you cannot explain that induction is true without the well-ordering of $\mathbb{N}$. In other words, *well-ordering is equivalent to induction:* each implies the other.

Claim: Explain precisely why $0 < a$ for any $a \in \mathbb{N}^+$, and $0 \leq a$ for any $a \in \mathbb{N}$.
Proof:

1. Let $a \in \mathbb{N}^+$ be arbitrary.
2. By _____, $a + 0 = a$.
3. By _____, $0 = -0$.
4. By _____, $a + (-0) = a$.
5. By definition of _____, $a - 0 = a$.
6. By _____, $a - 0 \in \mathbb{N}^+$.
7. By definition of _____, $0 < a$.
8. A similar argument tells us that if $a \in \mathbb{N}$, then $0 \leq a$.

Figure 0.1. Material for Exercise 0.21

- If $n \neq 1$, then $n > 1$ by the properties of arithmetic. Since n is the smallest element of S and $n - 1 < n$, we deduce that $n - 1 \notin S$. Thus $n - 1 \in P$. Let $i = n - 1$; then $i \in P$ and $i + 1 = n \notin P$. Thus P does not satisfy (IS).

We have shown that if $P \neq \mathbb{N}^+$, then P fails to satisfy at least one of (IB) or (IS). This is the contrapositive of the theorem. □

Induction is an enormously useful tool, and we will make use of it from time to time. You may have seen induction stated differently, and that's okay. There are several kinds of induction which are all equivalent. We use the form given here for convenience.

Exercises.

In this first set of exercises, we assume that you are not terribly familiar with creating and writing proofs, so we provide a few outlines, leaving blanks for you to fill in. As we proceed through the material, we expect you to grow more familiar and comfortable with thinking, so we provide fewer outlines, and in the outlines that we do provide, we require you to fill in the blanks with more than one or two words.

Exercise 0.21.
(a) Fill in each blank of Figure 0.1 with the justification.
(b) Why would someone writing a proof of the claim think to look at $a - 0$?
(c) Why would that person start with $a + 0$ instead?

Exercise 0.22.
(a) Fill in each blank of Figure 0.2 with the justification.
(b) Why would someone writing a proof of this claim think to look at the values of $a - b$ and $b - a$?
(c) Why is the introduction of S essential, rather than a distraction?

Exercise 0.23. Let $a \in \mathbb{Z}$. Show that:
(a) $a < a + 1$;
(b) if $a \in \mathbb{N}$, then $0 \leq a$; and
(c) if $a \in \mathbb{N}^+$, then $1 \leq a$.

Exercise 0.24. Let $a, b, c \in \mathbb{Z}$.

Claim: We can order any subset of $\mathbb{Z}$ linearly by $<$.
Proof:

1. Let $S \subseteq \mathbb{Z}$.
2. Let $a, b \in$_____. We consider three cases.
3. If $a - b \in \mathbb{N}^+$, then by $a < b$ by _____.
4. If $a - b = 0$, then simple arithmetic shows that _____.
5. Otherwise, $a - b \in \mathbb{Z} \backslash \mathbb{N}$. By definition of opposites, $b - a \in$_____.
   (a) Then $a < b$ by _____.
6. We have shown that we can order a and b linearly. Since a and b were arbitrary in _____, we can order *any* two elements of that set linearly.

Figure 0.2. Material for Exercise 0.22

(a) Prove that if $a \leq b$, then $a = b$ or $a < b$.
(b) Prove that if both $a \leq b$ and $b \leq a$, then $a = b$.
(c) Prove that if $a \leq b$ and $b \leq c$, then $a \leq c$.

Exercise 0.25. Let $a, b \in \mathbb{N}$ and assume that $0 < a < b$. Let $d = b - a$. Show that $d < b$.

Exercise 0.26. Let $a, b, c \in \mathbb{Z}$ and assume that $a \leq b$. Prove that
(a) $a + c \leq b + c$;
(b) if $c \in \mathbb{N}^+$, then $a \leq ac$; and
(c) if $c \in \mathbb{N}^+$, then $ac \leq bc$.

Note: You may henceforth assume this for *all* the inequalities given in Definition 0.13.

Exercise 0.27. Let $S \subseteq \mathbb{N}$. We know from the well-ordering property that S has a smallest element. Prove that this smallest element is unique.

Exercise 0.28. Show that $>$ is not a well-ordering of $\mathbb{N}$.

Exercise 0.29. Show that the ordering $<$ of $\mathbb{Z}$ generalizes "naturally" to an ordering $<$ of $\mathbb{Q}$ that is also a linear ordering.

Exercise 0.30. By definition, a function is a relation. Can a function be an equivalence relation?

Exercise 0.31.
(a) Fill in each blank of Figure 0.3 with the justification.
(b) Why would someone writing a proof of the claim think to write that $a_i < a_{i+1}$?
(c) Why would someone want to look at the smallest element of A?

Definition 0.32. Let $f : S \to U$ be a mapping of sets.

- We say that f is **one-to-one** if for every $a, b \in S$ where $f(a) = f(b)$, we have $a = b$.
- We say that f is **onto** if for every $x \in U$, there exists $a \in S$ such that $f(a) = x$.

Let S be a well-ordered set.
Claim: Every strictly decreasing sequence of elements of S is finite.
Proof:

1. Let $a_1, a_2, \ldots \in$_____.
   (a) Assume that the sequence is _____.
   (b) In other words, $a_{i+1} < a_i$ for all $i \in$_____.
2. By way of contradiction, suppose the sequence is _____.
   (a) Let $A = \{a_1, a_2, \ldots\}$.
   (b) By definition of _____, A has a smallest element. Let's call that smallest element b.
   (c) By definition of _____, $b = a_i$ for some $i \in \mathbb{N}^+$.
   (d) By _____, $a_{i+1} < a_i$.
   (e) By definition of _____, $a_{i+1} \in A$.
   (f) This contradicts the choice of b as the _____.
3. The assumption that the sequence is _____ is therefore not consistent with the assumption that the sequence is _____.
4. As claimed, then, _____.

Figure 0.3. Material for Exercise 0.31

Exercise 0.33. Suppose that $f : S \to U$ is a one-to-one, onto function. Let $g : U \to S$ by

$$g(u) = s \iff f(s) = u.$$

(a) Show that g is also a one-to-one, onto function.
(b) Show that g *undoes* f, in the sense that for any $s \in S$, we have $g(f(s)) = s$.

This justifies the notation of an **inverse function**; if two functions f and g satisfy the relationship of Exercise 0.33, then each is the inverse function of the other, and we write $g = f^{-1}$ and $f = g^{-1}$. Notice how this implies that $f = (f^{-1})^{-1}$.

0.2: Division

Before looking at algebraic objectrs, we need one more property of the integers.

The Division Theorem

The last "arithmetic operation" that you know about is division, but this operation is… "interesting".

Theorem 0.34 (The Division Theorem for Integers)**.** Let $n, d \in \mathbb{Z}$ with $d \neq 0$. There exist unique $q \in \mathbb{Z}$ and $r \in \mathbb{Z}$ satisfying (D1) and (D2) where
(D1) $n = qd + r$;
(D2) $0 \leq r < |d|$.

One implication of this theorem is that division *is not an operation on* $\mathbb{Z}$**!** An operation on $\mathbb{Z}$ is a relation $f : \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$, but the quotient and remainder imply that division is a relation of the

form $\div : (\mathbb{Z} \times (\mathbb{Z} \setminus \{0\})) \to \mathbb{Z} \times \mathbb{Z}$. That is not a binary operation *on* $\mathbb{Z}$. We explore this further in a moment, but for now let's turn to a proof of the theorem.

Proof. We consider two cases: $d \in \mathbb{N}^+$, and $d \in \mathbb{Z} \setminus \mathbb{N}$. First we consider $d \in \mathbb{N}^+$; by definitino of absolute value, $|d| = d$. We must show two things: first, that q and r exist; second, that r is unique.

Existence of q and r: First we show the existence of q and r that satisfy (D1). Let $S = \{n - qd : q \in \mathbb{Z}\}$ and $M = S \cap \mathbb{N}$. You will show in Exercise 0.51 that M is non-empty. By the well-ordering of $\mathbb{N}$, M has a smallest element; call it r. By definition of S, there exists $q \in \mathbb{Z}$ such that $n - qd = r$. Properties of arithmetic imply that $n = qd + r$.

Does r satisfy (D2)? By way of contradiction, assume that it does not; then $|d| \leq r$. We had assumed that $d \in \mathbb{N}^+$, so Exercises 0.21 and 0.25 implies that $0 \leq r - d < r$. Rewrite property (D1) using properties of arithmetic:

$$\begin{aligned} n &= qd + r \\ &= qd + d + (r - d) \\ &= (q+1)\,d + (r - d)\,. \end{aligned}$$

Rewrite this as $r - d = n - (q+1)\,d$, which shows that $r - d \in S$. Recall $0 \leq r - d$; by definition, $r - d \in \mathbb{N}$. We have $r - d \in S$ and $r - d \in \mathbb{N}$, so $r - d \in S \cap \mathbb{N} = M$. But recall that $r - d < r$, which contradicts the choice of r as the *smallest* element of M. This contradiction implies that r satisfies (D2).

Hence $n = qd + r$ and $0 \leq r < d$; q and r satisfy (D1) and (D2).

Uniqueness of q and r: Suppose that there exist $q', r' \in \mathbb{Z}$ such that $n = q'd + r'$ and $0 \leq r' < d$. By definition of S, $r' = n - q'd \in S$; by assumption, $r' \in \mathbb{N}$, so $r' \in S \cap \mathbb{N} = M$. We chose r to be minimal in M, so $0 \leq r \leq r' < d$. By substitution,

$$\begin{aligned} r' - r &= (n - q'd) - (n - qd) \\ &= (q - q')\,d. \end{aligned}$$

Moreover, $r \leq r'$ implies that $r' - r \in \mathbb{N}$, so by substitution, $(q - q')\,d \in \mathbb{N}$. Similarly, $0 \leq r \leq r'$ implies that $0 \leq r' - r \leq r'$. By substitution, $0 \leq (q - q')\,d \leq r'$. Since $d \in \mathbb{N}^+$, it must be that $q - q' \in \mathbb{N}$ also (repeated addition of a negative giving a negative), so $0 \leq q - q'$. If $0 \neq q - q'$, then $1 \leq q - q'$. By Exercise 0.26, $d \leq (q - q')\,d$. By Exercise 0.24, we see that $d \leq (q - q')\,d \leq r' < d$. This states that $d < d$, a contradiction. Hence $q - q' = 0$, and by substitution, $r - r' = 0$.

We have shown that if $0 < d$, then there exist unique $q, r \in \mathbb{Z}$ satisfying (D1) and (D2). We still have to show that this is true for $d < 0$. In this case, $0 < |d|$, so we can find unique $q, r \in \mathbb{Z}$ such that $n = q\,|d| + r$ and $0 \leq r < |d|$. By properties of arithmetic, $q\,|d| = q\,(-d) = (-q)\,d$, so $n = (-q)\,d + r$. □

Definition 0.35 (terms associated with division)**.** Let $n, d \in \mathbb{Z}$ and suppose that $q, r \in \mathbb{Z}$ satisfy the Division Theorem. We call n the **dividend**, d the **divisor**, q the **quotient**, and r the **remainder**.

Moreover, if $r = 0$, then $n = qd$. In this case, we say that d **divides** n, and write $d \mid n$. We also say that n is **divisible by** d. If we cannot find such an integer q, then d **does not divide** n, and we write $d \nmid n$.

In the past, you have probably heard of this as "divides evenly." In advanced mathematics, we typically leave off the word "evenly".

As noted, division is not a binary operation on $\mathbb{Z}$, or even on $\mathbb{Z}\setminus\{0\}$. That doesn't seem especially tidy, so we define a set that allows us to make an operation of division:

- the **rational numbers**, sometimes called the **fractions**, $\mathbb{Q} = \{a/b : a, b \in \mathbb{Z} \text{ and } b \neq 0\}$.

We observe the conventions that $a/1 = a$ and $a/b = c/d$ if $ad = bc$. This makes division into a binary operation on $\mathbb{Q}\setminus\{0\}$, though not on $\mathbb{Q}$ since division by zero remains undefined.

Remark 0.36. Why do we insist that $b \neq 0$? Basically, it doesn't make sense. The very idea of division means that if $a/b = c$, then $a = bc$. So, let $a/0 = c$. In that case, $a = 0c$. This is true only if $a = 0$, so we can't have $b = 0$. On the other hand, this reasoning doesn't apply to $0/0$, so what about allowing that to be in $\mathbb{Q}$? Actually, that offends our notion of an operation! Why? because if we put $0/0 \in \mathbb{Q}$, it is not hard to show that both $0/0 = 1$ and $0/0 = 2$, which would imply that $1 = 2$!

We have built a chain of sets $\mathbb{N}^+ \subsetneq \mathbb{N} \subsetneq \mathbb{Z} \subsetneq \mathbb{Q}$, extending each set with some useful elements. Even this last extension of this still doesn't complete arithmetic, since the fundamental *Pythagorean Theorem* isn't closed in $\mathbb{Q}$! Take a right triangle with two legs, each of length 1; the hypotenuse must have length $\sqrt{2}$. As we show later in the course, *this number is not rational!* That means we cannot compute all measurements along a line using $\mathbb{Q}$ alone. This motivates a definition to remedy the situation:

- the **real numbers** contain a number for every possible measurement of distance along a line.[6]

We now have

$$\mathbb{N}^+ \subsetneq \mathbb{N} \subsetneq \mathbb{Z} \subsetneq \mathbb{Q} \subsetneq \mathbb{R}.$$

In the exercises, you will generalize the ordering $<$ to the set $\mathbb{Q}$. As for an ordering on $\mathbb{R}$, we leave that to a class in analysis, but you can treat it as you have in the past.

Do we need anything else? Indeed, we do: before long, we will see that even these sets are insufficient for algebra.

Equivalence classes

Recall that an equivalence relation satisfies the reflexive, symmetric, and transitive properties. Under an equivalence relation, different elements of a set are considered "equivalent".

Example 0.37. Let $\sim$ be a relation on $\mathbb{Z}$ such that $a \sim b$ if and only if a and b have the same remainder after division by 4. Then $7 \sim 3$ and $7 \sim 19$ but $7 \not\sim 6$.

[6] Speaking precisely, $\mathbb{R}$ is the set of limits of "nice sequences" of rational numbers. By "nice", we mean that the elements of the sequence eventually grow closer together than any rational number. The technical term for this is a **Cauchy sequence**. For more on this, see any textbook on real analysis.

We will find it *very useful* to group elements that are equivalent under a certain relation.

Definition 0.38. Let $\sim$ be an equivalence relation on a set A, and let $a \in A$. The **equivalence class** of a in A with respect to $\sim$ is

$$[a] = \{b \in S : a \sim b\}.$$

Example 0.39. Continuing our example above, $3, 19 \in [7]$ but $6 \notin [7]$.

It turns out that equivalence relations partition a set! We will prove this in a moment, but we should look at a concrete example first.

Normally, we think of the division of n by d as dividing a set of n objects into q groups, where each group contains d elements, and r elements are left over. For example, $n = 23$ apples divided among $d = 6$ bags gives $q = 3$ apples per bag and $r = 5$ apples left over.

Another way to look at division by d is that it divides $\mathbb{Z}$ into d sets of integers. Each integer falls into a set according to its remainder after division. An illustration using $n = 4$:

$\mathbb{Z}$:	…	-2	-1	0	1	2	3	4	5	6	…
		↓	↓	↓	↓	↓	↓	↓	↓	↓	
division by 4:	…	2	3	0	1	2	3	0	1	2	…

Here $\mathbb{Z}$ is divided into four sets

$$\begin{aligned} A &= \{\ldots, -4, 0, 4, 8, \ldots\} = [0] \\ B &= \{\ldots, -3, 1, 5, 9, \ldots\} = [1] \\ C &= \{\ldots, -2, 2, 6, 10, \ldots\} = [2] \\ D &= \{\ldots, -1, 3, 7, 11, \ldots\} = [3]. \end{aligned} \tag{1}$$

Observe two important facts:

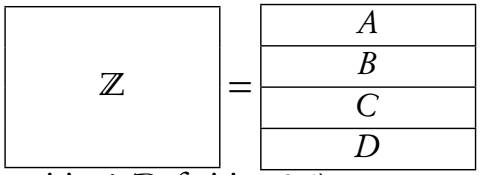
- the sets A, B, C, and D *cover* $\mathbb{Z}$; that is,

$$\mathbb{Z} = A \cup B \cup C \cup D;$$

  and
- the sets A, B, C, and D are *disjoint;* that is,

$$A \cap B = A \cap C = A \cap D = B \cap C = B \cap D = C \cap D = \emptyset.$$

We can diagram this:



This should remind you of a partition! (Definition 0.6)

Example 0.40. Let $\mathcal{B} = \{A, B, C, D\}$ where A, B, C, and D are defined as in (1). Since the elements of $\mathcal{B}$ are disjoint, and they cover $\mathbb{Z}$, we conclude that $\mathcal{B}$ is a partition of $\mathbb{Z}$.

A more subtle property is at work here: division has actually produced for us an equivalence relation on the integers.

Theorem 0.41. Let $d \in \mathbb{Z} \setminus \{0\}$, and define a relation $\equiv_d$ in the following way: for any $m, n \in \mathbb{Z}$, we say that $m \equiv_d n$ if and only if they have the same remainder after division by d. This is an equivalence relation.

Proof. We have to prove that $\equiv_d$ is reflexive, symmetric, and transitive.

Reflexive? Let $n \in \mathbb{Z}$. The Division Theorem tells us that the remainder of division of n by d is unique, so $n \equiv_d n$.

Symmetric? Let $m, n \in \mathbb{Z}$, and assume that $m \equiv_d n$. This tells us that m and n have the same remainder after division by d. It obviously doesn't matter whether we state m first or n first; we can just as well say that n and m have the same remainder after division by d. That is, $n \equiv_d m$.

Transitive? Let $\ell, m, n \in \mathbb{Z}$, and assume that $\ell \equiv_d m$ and $m \equiv_d n$. This tells us that ℓ and m have the same remainder after division by d, and m and n have the same remainder after division by d. The Division Theorem tells us that the remainder of division of n by d is unique, so ℓ and n have the same remainder after division by d. That is, $\ell \equiv_d n$. □

We have seen that division induces both a partition and an equivalence relation. Do equivalence relations always coincide with partitions? Surprisingly, yes!

Theorem 0.42. An equivalence relation partitions a set, and any partition of a set defines an equivalence relation.

Actually, it isn't so surprising if you understand the proof, or even if you just think about the meaning of an equivalence relation. The reflexive property implies that every element is in relation with itself, and the other two properties help ensure that no element can be related to two elements that are not themselves related. The proof provides some detail.

Proof. *Does any partition of any set define an equivalence relation?* Let S be a set, and $\mathcal{B}$ a partition of S. Define a relation $\sim$ on S in the following way: $x \sim y$ if and only if x and y are in the same element of $\mathcal{B}$. That is, if $X \in \mathcal{B}$ is the set such that $x \in X$, then $y \in X$ as well.

We claim that $\sim$ is an equivalence relation. It is reflexive because a partition covers the set; that is, $S = \bigcup_{B \in \mathcal{B}}$, so for any $x \in S$, we can find $B \in \mathcal{B}$ such that $x \in B$, which means the statement that "x is in the same element of $\mathcal{B}$ as itself" ($x \sim x$) actually makes sense. The relation is symmetric because $x \sim y$ means that x and y are in the same element of $\mathcal{B}$, which is equivalent to saying that y and x are in the same element of $\mathcal{B}$; after all, set membership is not affected by which element we list first. So, if $x \sim y$, then $y \sim x$. Finally, the relation is transitive because distinct elements of a partition are disjoint. Let $x, y, z \in S$, and assume $x \sim y$ and $y \sim z$. Choose $X, Z \in \mathcal{B}$ such that $x \in X$ and $z \in Z$. The symmetric property tells us that $z \sim y$, and the definition of the relation implies that $y \in X$ and $y \in Z$. The fact that they share a common element tells us that X and Z are not disjoint ($X \cap Z \neq \emptyset$). By the definition of a partition, X and Z are not distinct.

Does an equivalence relation partition a set? Let S be a set, and $\sim$ an equivalence relation on S. If S is empty, the claim is vacuously true, so assume S is non-empty. Let $x \in S$. Notice that $[x] \neq \emptyset$, since the reflexive property of an equivalence relation guarantees that $x \sim x$, which implies that $x \in [x]$.

Let $\mathcal{B}$ be the set of all equivalence classes of elements of x; that is, $\mathcal{B} = \{[x] : x \in S\}$. We have already seen that every $x \in S$ appears in its own equivalence class, so $\mathcal{B}$ covers S. Are distinct equivalence classes also disjoint?

Let $X, Y \in \mathcal{B}$ and assume that assume that $X \cap Y \neq \emptyset$; this means that we can choose $z \in X \cap Y$. By definition, $X = [x]$ and $Y = [y]$ for some $x, y \in S$. By definition of $X = [x]$ and $Y = [y]$, we know that $x \sim z$ and $y \sim z$. Now let $w \in X$ be arbitary; by definition, $x \sim w$; by the symmetric property of an equivalence relation, $w \sim x$ and $z \sim y$; by the transitive property of an equivalence relation, $w \sim z$, and by the same reasoning, $w \sim y$. Since w was an arbitrary element of X, *every* element of X is related to y; in other words, every element of X is in $[y] = Y$, so $X \subseteq Y$.

A similar argument shows that $X \supseteq Y$. By definition of set equality, $X = Y$. We took two arbitrary equivalence classes of S, and showed that if they were not disjoint, then they were not distinct. The contrapositive states that if they are distinct, then they are disjoint. Since the elements of $\mathcal{B}$ are equivalence classes of S, we conclude that distinct elements of $\mathcal{B}$ are disjoint. They also cover S, so as claimed, $\mathcal{B}$ is a partition of S induced by the equivalence relation. □

Exercises.

Exercise 0.43. Identify the quotient and remainder when dividing:
(a) 10 by -5;
(b) -5 by 10;
(c) -10 by -4.

Exercise 0.44. Prove that if $a \in \mathbb{Z}$, $b \in \mathbb{N}^+$, and $a \mid b$, then $a \leq b$.

Exercise 0.45. Show that $a \leq |a|$ for all $a \in \mathbb{Z}$.

Exercise 0.46. Show that divisibility is transitive for the integers; that is, if $a, b, c \in \mathbb{Z}$, $a \mid b$, and $b \mid c$, then $a \mid c$.

Exercise 0.47. Extend the definition of $<$ so that we can order rational numbers. That is, find a criterion on $a, b, c, d \in \mathbb{Z}$ that tells us when $a/b < c/d$.

Definition 0.48. We define lcm, the **least common multiple** of two integers, as

$$\operatorname{lcm}(a, b) = \min\{n \in \mathbb{N}^+ : a \mid n \text{ and } b \mid n\}.$$

This is a precise definition of the least common multiple that you should already be familiar with: it's the smallest (min) positive ($n \in \mathbb{N}^+$) multiple of a and b ($a \mid n$, and $b \mid n$).

Exercise 0.49.
(a) Fill in each blank of Figure 0.4 with the justification.
(b) One part of the proof claims that "A similar argument shows that $b \mid r$." State this argument in detail.

Exercise 0.50. Define a relation $\equiv$ on $\mathbb{Q}$, the set of real numbers, in the following way:

$$a \equiv b \text{ if and only if } a - b \in \mathbb{Z}.$$

Let $a, b, c \in \mathbb{Z}$.
Claim: If a and b both divide c, then $\text{lcm}(a,b)$ also divides c.
Proof:

1. Let $d = \text{lcm}(a,b)$. By _____, we can choose q, r such that $c = qd + r$ and $0 \leq r < d$.
2. By definition of _____, both a and b divide d.
3. By definition of _____, we can find $x, y \in \mathbb{Z}$ such that $c = ax$ and $d = ay$.
4. By _____, $ax = q(ay) + r$.
5. By _____, $r = a(x - qy)$.
6. By definition of _____, $a \mid r$. A similar argument shows that $b \mid r$.
7. We have shown that a and b divide r. Recall that $0 \leq r < d$, and _____. By definition of lcm, $r = 0$.
8. By _____, $c = qd = q\text{lcm}(a,b)$.
9. By definition of _____, $\text{lcm}(a,b)$ divides c.

Figure 0.4. Material for Exercise 0.49

(a) Give some examples of rational numbers that are related. Include examples where a and b are not themselves integers.
(b) Show that that $a \equiv b$ if they have the same *fractional part*. That is, if we write a and b in decimal form, we see exactly the same numbers on the right hand side of the decimal point, in exactly the same order. (You may assume, without proof, that we can write any rational number in decimal form.)
(c) Is $\equiv$ an equivalence relation?

For any $a \in \mathbb{Q}$, let S_a be the set of all rational numbers b such that $a \equiv b$. We'll call these new sets **classes**.

(d) Is every $a \in \mathbb{Q}$ an element of some class? Why?
(e) Show that if $S_a \neq S_b$, then $S_a \cap S_b = \emptyset$.

Exercise 0.51.
(a) Fill in each blank of Figure 0.5 with the justification.
(b) Why would someone writing a proof of the claim think to look at $n - qd$?
(c) Why would this person want to find a value of q?

Exercise 0.52. Let X and Y on the lattice $L = \mathbb{Z} \times \mathbb{Z}$. Let's say that addition is performed as with vectors:

$$X + Y = (x_1 + y_1, x_2 + y_2),$$

multiplication is performed by this *very odd* definition:

$$X \cdot Y = (x_1 y_1 - x_2 y_2, x_1 y_2 + x_2 y_1),$$

and the magnitude of a point is devided by the usual Euclidean metric,

$$||X|| = \sqrt{x_1^2 + x_2^2}.$$

(a) Suppose $D = (3,1)$. Calculate $(c,0) \cdot D$ for several different values of c. How would you describe the results geometrically?

Let $n, d \in \mathbb{Z}$, where $d \in \mathbb{N}^+$. Define $M = \{n - qd : q \in \mathbb{Z}\}$.
Claim: $M \cap \mathbb{N} \neq \emptyset$.
Proof: We consider two cases.

1. First suppose $n \in \mathbb{N}$.
   (a) Let $q =$_____. By definition of $\mathbb{Z}$, $q \in \mathbb{Z}$.
   (You can reverse-engineer this answer if you look down a little.)
   (b) By properties of arithmetic, $qd =$_____.
   (c) By _____, $n - qd = n$.
   (d) By hypothesis, $n \in$_____.
   (e) By _____, $n - qd \in \mathbb{Z}$.
2. It's possible that $n \notin \mathbb{N}$, so now let's assume that, instead.
   (a) Let $q =$_____. By definition of $\mathbb{Z}$, $q \in \mathbb{Z}$.
   (Again, you can reverse-engineer this answer if you look down a little.)
   (b) By substitution, $n - qd =$_____.
   (c) By _____, $n - qd = -n(d-1)$.
   (d) By _____, $n \notin \mathbb{N}$, but it is in $\mathbb{Z}$. Hence, $-n \in \mathbb{N}^+$.
   (e) Also by _____, $d \in \mathbb{N}^+$, so arithmetic tells us that $d - 1 \in \mathbb{N}$.
   (f) Arithmetic now tells us that $-n(d-1) \in \mathbb{N}$. (pos×natural=natural)
   (g) By _____, $n - qd \in \mathbb{Z}$.
3. In both cases, we showed that $n - qd \in \mathbb{N}$. By definition of _____, $n - qd \in M$.
4. By definition of _____, $n - qd \in M \cap \mathbb{N}$.
5. By definition of _____, $M \cap \mathbb{N} \neq \emptyset$.

Figure 0.5. Material for Exercise 0.51

(b) With the same value of D, calculate $(0, c)\,D$ for several different values of c. How would you describe the results geometrically?
(c) Suppose $N = (10, 4)$, $D = (3, 1)$, and $R = N - (3, 0) \cdot D$. Show that $||R|| < ||D||$.
(d) Suppose $N = (10, 4)$, $D = (1, 3)$, and $R = N - (3, -3) \cdot D$. Show that $||R|| < ||D||$.
(e) Use the results of (a) and (b) to provide a geometric description of how N, D, and R are related in (c) and (d).
(f) Suppose $N = (10, 4)$ and $D = (2, 2)$. Find Q such that if $R = N - Q \cdot D$, then $||R|| < ||D||$. Try to build on the geometric ideas you gave in (e).
(g) Show that for any $N, D \in L$ with $D \neq (0, 0)$, you can find $Q, R \in L$ such that $N \cdot D + R$ and $||R|| < ||D||$. Again, try to build on the geometric ideas you gave in (e).

0.3: Linear algebra

Linear algebra is the study of algebraic objects related to linear polynomials. It includes not only matrices and operations on matrices, but vector spaces, bases, and linear transformations. For the most part, we will focus on matrices and on linear transformations.

Matrices

Definition 0.53. An $m \times n$ **matrix** is a list of m lists (**rows**) of n numbers. If $m = n$, we call the matrix **square**, and say that the **dimension** of the matrix is m.

Notation 0.54. We write the jth element of row i of the matrix A as a_{ij}. If $a_{ij} = 0$ and we are especially lazy, then we often omit writing it in the matrix. If the dimension of A is m, then we write $\dim A = m$.

Example 0.55. If

$$A = \begin{pmatrix} 1 & & 1 \\ & 1 & \\ & 5 & 1 \end{pmatrix},$$

then $a_{21} = 0$ while $a_{32} = 5$. Notice that A is a 3×3 matrix; or, $\dim A = 3$.

Definition 0.56. The **transpose** of a matrix A is the matrix B satisfying $b_{ij} = a_{ji}$. In other words, the jth element of row i of B is the ith element of row j of A. A **column** of a matrix is a row of its transpose.

Notation 0.57. We often write A^T for the transpose of A.

Example 0.58. If A is the matrix of the previous example, then

$$A^T = \begin{pmatrix} 1 & & \\ & 1 & 5 \\ 1 & & 1 \end{pmatrix}.$$

While non-square matrices are important, we consider mostly square matrices in this class, with the exception of $m \times 1$ matrices, which are also called **column vectors**. It is easy to define three operations for matrices:

We *add* matrices by adding entries in the same row and column. That is, if A and B are $m \times n$ matrices and $C = A + B$, then $c_{ij} = a_{ij} + b_{ij}$ for all $1 \leq i \leq m$ and all $1 \leq j \leq n$. Notice that C is also an $m \times n$ matrix.

We *subtract* matrices in the same way.

We *multiply* matrices a little differently. If A is an $m \times r$ matrix, B is an $r \times n$ matrix, and $C = AB$, then C is the $m \times n$ matrix whose entries satisfy

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj};$$

that is, the jth element in row i of C is the sum of the products of corresponding elements of row i of A and column j of B.

Example 0.59. If A is the matrix of the previous example and

$$B = \begin{pmatrix} 1 & 5 & -1 \\ & 1 & \\ & -5 & 1 \end{pmatrix},$$

then

$$AB = \begin{pmatrix} 1\cdot 1+0\cdot 0+1\cdot 0 & 1\cdot 5+0\cdot 1+1\cdot -5 & 1\cdot -1+0\cdot 0+1\cdot 1 \\ 0\cdot 1+1\cdot 0+0\cdot 0 & 0\cdot 5+1\cdot 1+0\cdot -5 & 0\cdot -1+1\cdot 0+0\cdot 1 \\ 0\cdot 1+5\cdot 0+1\cdot 0 & 0\cdot 5+5\cdot 1+1\cdot -5 & 0\cdot -1+5\cdot 0+1\cdot 1 \end{pmatrix}$$
$$= \begin{pmatrix} 1 & & \\ & 1 & \\ & & 1 \end{pmatrix}.$$

If we take the matrices of the previous example and let $I = AB$, then something interesting happens:

$$AI = IA = A \quad \text{and} \quad BI = IB = B.$$

The pattern of this matrix ensures that the property remains true for *any* matrix, as long as you're working in the correct dimension. That is, I is an "identity" matrix. In particular, it's the identity of **multiplication**. Is there another identity matrix? Certainly there is for addition; you can probably guess that one yourself; just let the matrix contain only zeros.

Can there be another identity matrix for *multiplication*? In fact, there *cannot*. Rather than show this directly, however, we will wait until Section 1.1. For now, we'll consolidate our current gains. First, some notation.

Notation 0.60.

- We write $\mathbb{0}$ (that's a **bold zero**) for any matrix whose elements are all zero; that is, $a_{ij} = \mathbb{0}$ for all $1 \leq i, j \leq \dim \mathbb{0}$.
- We write I_n for the matrix of dimension n satisfying
  - $a_{ii} = 1$ for any $i = 1, 2, \ldots, n$; and
  - $a_{ij} = \mathbb{0}$ for any $i \neq j$.

Now, a formal statement of the result.

> **Theorem 0.61.** The zero matrix $\mathbb{0}$ is an identity for matrix addition. The matrix I_n is an identity for matrix multiplication.

Notice that there's a bit of imprecision in this statement. You have to infer from the statement that $n \in \mathbb{N}^+$, $\mathbb{0}$ is an $n \times n$ matrix, and we mean that $\mathbb{0}$ is an identity for addition when we're talking about other matrices of dimension n. We should *not* infer that the statement means that $\mathbb{0}$ is an identity for matrices of dimension $m + 2$; that would be silly, as the addition would be undefined. When reading theorems, you sometimes have to read between the lines.

Proof. Let A be a matrix of dimension n. By definition, the jth element in row i of $A + \mathbb{0}$ is $a_{ij} + \mathbb{0} = a_{ij}$. This is true regardless of the values of i and j, so $A + \mathbb{0} = A$. A similar argument shows that $\mathbb{0} + A = A$. Since A is arbitrary, $\mathbb{0}$ really is an additive identity.

As for I_n, we point out that the jth element of row i of AI_n is (by definition of multiplication)

$$\sum_{\substack{k=1,\ldots,m \\ k \neq j}} a_{ik} \cdot \mathbb{0} + a_{ij} \cdot 1.$$

Simplifying this gives us a_{ij}. This is true regardless of the values of i and j, so $AI_n = A$. A similar argument shows that $I_n A = A$. Since A is arbitrary, I_n really is a multiplicative identity. □

Given a matrix A, an **inverse** of A is any matrix B such that $A+B=0$ (if B is an **additive inverse**) and $AB=I_n$ (if B is a multiplicative inverse). Additive inverses always exist, and it is easy to construct them. Multiplicative inverses *do not* exist for some matrices, even when the matrix is square. Because of this we call a matrix is **invertible** if it has a multiplicative matrix.

Notation 0.62. We write the additive inverse of a matrix A and $-A$, and the multiplicative inverse of A as A^{-1}.

Example 0.63. The matrices A and B of the previous example are inverses; that is, $A=B^{-1}$ and $B=A^{-1}$.

We want one more property before we move on.

Theorem 0.64. Matrix multiplication is associative if the entries of the matrices are associative under multiplication and commutative under addition. That is, if A, B, and C are matrices with those properties, then $A(BC)=(AB)C$.

Proof. Let A be an $m\times r$ matrix, B an $r\times s$ matrix, and C an $s\times n$ matrix. By definition, the ℓth element in row i of AB is

$$(AB)_{i\ell}=\sum_{k=1}^{r}a_{ik}b_{k\ell}.$$

Likewise, the jth element in row i of $(AB)C$ is

$$((AB)C)_{ij}=\sum_{\ell=1}^{s}(AB)_{i\ell}\,c_{\ell j}=\sum_{\ell=1}^{s}\left[\left(\sum_{k=1}^{r}a_{ik}b_{k\ell}\right)c_{\ell j}\right].$$

Notice that $c_{\ell j}$ is multiplied to a sum; we can distribute it and obtain

$$((AB)C)_{ij}=\sum_{\ell=1}^{s}\sum_{k=1}^{r}(a_{ik}b_{k\ell})\,c_{\ell j}. \tag{2}$$

We turn to the other side of the equation. By definition, the jth element in row k of BC is

$$(BC)_{kj}=\sum_{\ell=1}^{s}b_{k\ell}c_{\ell j}.$$

Likewise, the jth element in row i of $A(BC)$ is

$$(A(BC))_{ij}=\sum_{k=1}^{r}\left(a_{ik}\sum_{\ell=1}^{s}b_{k\ell}c_{\ell j}\right).$$

This time, a_{ik} is multiplied to a sum; we can distribute it and obtain

$$(A(BC))_{ij}=\sum_{k=1}^{r}\sum_{\ell=1}^{s}a_{ik}\left(b_{k\ell}c_{\ell j}\right).$$

By the associative property of the entries,

$$(A(BC))_{ij} = \sum_{k=1}^{r} \sum_{\ell=1}^{s} (a_{ik} b_{k\ell})\, c_{\ell j}. \tag{3}$$

The only difference between equations (2) and (3) is in the order of the summations: whether we add up the k's first or the ℓ's first. That is, the sums have the same terms, but those terms appear in different orders! We assumed the entries of the matrices were commutative under addition, so the order of the terms does not matter; we have

$$((AB)C)_{ij} = (A(BC))_{ij} .$$

We chose arbitrary i and j, so this is true for all entries of the matrices. The matrices are equal, which means $(AB)C = A(BC)$, □

Linear transformations

We can view matrices as a special sort of function over other matrices. A common example of this is to consider the set D of $n \times 1$ column vectors. If M is an $n \times n$ matrix, we can define a function $f_M : D \to D$ by

$$f_M(\mathbf{x}) = M\mathbf{x}.$$

Read this as, "f_M maps $\mathbf{x}$ to the product of M and $\mathbf{x}$."

Example 0.65. If

$$M = \begin{pmatrix} 1 & & 1 \\ & 1 & \\ 5 & & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{x} = \begin{pmatrix} 1 \\ 3 \\ -2 \end{pmatrix},$$

then

$$f_M(\mathbf{x}) = M\mathbf{x} = \begin{pmatrix} -1 \\ 3 \\ 13 \end{pmatrix}.$$

This function is a special example of what we call a *linear transformation*. To define it precisely, we have to use the term **vector space**. If you do not remember that term, or never learned it, first go slap whomever taught you linear algebra, then content yourself with the knowledge that, in this class, it will be enough to know that any set D of all possible column vectors with n rows is a vector space for any $n \in \mathbb{N}^+$. Whatever that is. Then go slap your former linear algebra teacher again.

Definition 0.66. Let V be a vector space over the real numbers $\mathbb{R}$, and f a function on V. We say that f is a **linear transformation** if it preserves

- *scalar multiplication*, that is, $f(av) = af(v)$ for any $a \in \mathbb{R}$ and any $v \in V$, and
- *vector addition*, that is, $f(u+v) = f(u) + f(v)$ for any $u, v \in V$.

Eventually, you will learn about a special kind of function that works very similarly to linear

transformations, called a **homomorphism**. For now, let's look at the classic example of a linear transformation, a matrix.

Example 0.67. Recall M and $\mathbf{x}$ from Example 0.65. Let

$$\mathbf{y} = \begin{pmatrix} 3 \\ 0 \\ 2 \end{pmatrix}.$$

Using the definitions of matrix addition and matrix multiplication, you can verify that

$$M(\mathbf{x}+\mathbf{y}) = \begin{pmatrix} 4 \\ 3 \\ 15 \end{pmatrix},$$

and also

$$M\mathbf{x}+M\mathbf{y} = \begin{pmatrix} -1 \\ 3 \\ 17 \end{pmatrix} + \begin{pmatrix} 5 \\ 0 \\ -2 \end{pmatrix} = \begin{pmatrix} 4 \\ 3 \\ 15 \end{pmatrix}.$$

Now let $a = 4$. Using the definitions of matrix and scalar multiplication, you can verify that

$$M(a\mathbf{x}) = \begin{pmatrix} -4 \\ 12 \\ 68 \end{pmatrix},$$

and also

$$aM\mathbf{x} = 4\begin{pmatrix} -1 \\ 3 \\ 17 \end{pmatrix} = \begin{pmatrix} -4 \\ 12 \\ 68 \end{pmatrix}.$$

The example does *not* show that f_M is a linear transformation, because we tested M only with particular vectors $\mathbf{x}$ and $\mathbf{y}$, and with a particular scalar a. To show that f_M is a linear trasnformation, you'd have to show that f_M preserves scalar multiplication and vector addition on *all* scalars and vectors. Who has time for that? There are infinitely many of them, after all! Better to knock it off with a theorem whose proof relies on symbolic, or "generic", structure.

Theorem 0.68. For any matrix A of dimension n, the function f_A on all $n \times 1$ column vectors is a linear transformation.

Proof. Let A be a matrix of dimension n.

First we show that f_A preserves scalar multiplication. Let $c \in \mathbb{R}$ and $\mathbf{x}$ be an $n \times 1$ column vector. By definition of scalar multiplication, the element in row i of $c\mathbf{x}$ is cx_i. By definition of matrix multiplication, the element in row i of $A(c\mathbf{x})$ is

$$\sum_{k=1}^{m} \left[a_{ik}(cx_k)\right].$$

Apply the commutative, associative, and distributive properties of the field to rewrite this as

$$c\sum_{k=1}^{m} a_{ik}x_k.$$

On the other hand, the element in row i of $A\mathbf{x}$ is, by definition of matrix multiplication,

$$\sum_{k=1}^{m} a_{ik}x_k.$$

If we multiply it by c, we find that $A(c\mathbf{x}) = cA\mathbf{x}$, as claimed.

We leave it to you to show that f_A preserves vector addition; see Exercise 0.84. □

An important aspect of a linear transformation is the kernel.

Definition 0.69. The **kernel** of a linear transformation f is the set of vectors that are mapped to 0. In other words, the kernel is the set

$$\{v \in V : f(v) = 0\}.$$

Notation 0.70. We write $\ker f$ for the kernel of f. We also write $\ker M$ when we mean $\ker f_M$.

Example 0.71. Let

$$M = \begin{pmatrix} 1 & 0 & 5 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Let

$$\mathbf{x} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \quad \text{and} \quad \mathbf{y} = \begin{pmatrix} -5 \\ 0 \\ 1 \end{pmatrix}.$$

Since

$$M\mathbf{x} = \begin{pmatrix} 6 \\ 2 \\ 0 \end{pmatrix} \quad \text{and} \quad M\mathbf{y} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = 0,$$

we see that $\mathbf{x}$ is not in the kernel of M, but $\mathbf{y}$ is. In fact, it can be shown (you will do so in the exercises) that

$$\ker M = \left\{ v \in V : v = \begin{pmatrix} -5c \\ 0 \\ c \end{pmatrix} \exists c \in \mathbb{F} \right\}.$$

The kernel has a lot of important and fascinating properties, but exploring them goes well beyond the scope of this course.

Determinants

An important property of a square matrix A is its determinant, denoted by $\det A$. We won't explain *why* it's important here, beyond saying that it has the property of being **invariant** when

you rewrite the matrix in certain ways (see, for example, Theorem 0.77). We don't even define it terribly precisely; we simply summarize what you ought to know:

- to every matrix, we can associate a unique scalar, called its **determinant**;
- we can compute the determinant using a technique called *expansion by minors* along any row or column; and
- the determinant enjoys a number of useful properties, some of which are listed below.

Example 0.72. Recall the matrix A from Example (0.55). If we expand by minors on the first row, we find that

$$\begin{aligned}\det A &= 1\cdot(-1)^{1+1}\begin{vmatrix}1&0\\5&1\end{vmatrix}+0\cdot(-1)^{1+2}\begin{vmatrix}0&0\\0&1\end{vmatrix}+1\cdot(-1)^{1+3}\begin{vmatrix}0&1\\0&5\end{vmatrix}\\ &= 1.\end{aligned}$$

We call a matrix **singular** if its determinant is zero, and **nonsingular** otherwise. The matrix A in the example above is nonsingular.

We now summarize the properties of the determinant. One caveat: these properties are not necessarily true if the entries of the matrices do not come from $\mathbb{R}$. In many cases, they *are* true when the entries come from other sets, but to go into the details requires more work than we have time for here. One particular property that we state without proof is:

Proposition 0.73. The determinant of a matrix is invariant with respect to the choice of row or column for the expansion by cofactors. That is, it doesn't matter which row or column of a matrix you choose; you always get the same answer for that matrix.

Proving Proposition 0.73 would take a lot of time, and isn't really useful for this course. Any half-decent textbook on linear algebra will have the proof, so you can look it up there, if you like.

Notation 0.74. We write $\mathbf{a}_i$ for the ith row of matrix A, and $A_{\hat{i}\hat{j}}$ for the submatrix of A formed by removing row i and column j.

For the remaining properties, the proof is either an exercise, or appears in an appendix to this section after the exercises.

Theorem 0.75. If B is the same as the square matrix A, except that row i has been multiplied by a scalar c, then $\det B = c\det A$.

Proof. See page 30. □

Theorem 0.76. For any square matrix A, $\det A = \det A^T$.

Proof. You do it! See Exercise 0.88. □

The next theorem requires some lesser properties, which we will relegate to the status of "lemmas", as they aren't quite so important, though they are interesting on their own. First, we state the theorem.

Theorem 0.77. If A is a square matrix and B is a matrix found by adding a multiple of one row of A to another, then $\det A = \det B$.

Now, we state and prove each of the special properties we will need.

Lemma 0.78. If B is the same as the square matrix A, except that row i has been exchanged with row j, then $\det B = -\det A$.

Proof. See page 30. □

Lemma 0.79. If the square matrix A has two identical rows, then $\det A = 0$.

Proof. See page 31. □

Lemma 0.80. Let $b_1, \ldots, b_n \in \mathbb{R}$. If

$$A = \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} a_{11} + b_1 & a_{12} + b_2 & \cdots & a_{1n} + b_n \\ & \mathbf{a}_2 & & \\ & \vdots & & \\ & \mathbf{a}_n & & \end{pmatrix},$$

then

$$\det B = \det A + \det \begin{pmatrix} b_1 & b_2 & \cdots & b_n \\ & \mathbf{a}_2 & & \\ & \vdots & & \\ & \mathbf{a}_n & & \end{pmatrix}.$$

Proof. See page 31. □

Theorem 0.81. A square matrix A is singular if and only if we can write its first row as a **linear combination** of the others. That is, if we write $\mathbf{a}_i$ for the ith row of A and $\dim A = n$, then we can find $c_2, \ldots, c_n \in \mathbb{R}$ such that

$$\mathbf{a}_1 = c_2\mathbf{a}_2 + \cdots + c_n\mathbf{a}_n.$$

Proof. You do it! See Exercise 0.81. □

Theorem 0.82. For any two matrices A and B of dimension n, $\det(AB) = \det A \cdot \det B$.

Proof. See page 32. □

Theorem 0.83. An inverse exists of a matrix A exists if and only if $\det A \neq 0$; that is, if and only if A is nonsingular.

Proof. You do it! See Exercise 0.83. □

Exercises.

Exercise 0.84. Show that matrix multiplication distributes over a sum of vectors. In other words, complete the proof of Theorem 0.68.

Exercise 0.85. Let

$$M = \begin{pmatrix} 1 & & 1 \\ & 1 & \\ 5 & -1 & \end{pmatrix} \quad \text{and} \quad N = \begin{pmatrix} 1 & 0 & 5 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Show that

$$\ker M = \{0\},$$

but

$$\ker N = \left\{ v \in V : v = \begin{pmatrix} -5c \\ 0 \\ c \end{pmatrix} \exists c \in \mathbb{F} \right\}.$$

Exercise 0.86. Use Theorem 0.77 to prove Theorem 0.81. That is, show that a matrix is singular if and only if we can write its first row as a linear combination of the others.

Exercise 0.87. Use Theorems 0.77 and 0.82 to prove Theorem 0.83. That is, show that a matrix has an inverse if and only if its determinant is nonzero.

Exercise 0.88. Prove Theorem 0.76. That is, show that for any matrix A, $\det A = \det A^T$.

Exercise 0.89. Show that $\det A^{-1} = (\det A)^{-1}$.
Note: In the first, we have the inverse of a matrix; in the second, we have the inverse of a number!

Exercise 0.90. Let i be the imaginary number such that $i^2 = -1$, and let Q_8 be the set of **quaternions**, defined by the matrices $\{\pm\mathbf{1}, \pm\mathbf{i}, \pm\mathbf{j}, \pm\mathbf{k}\}$ where

$$\mathbf{1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mathbf{i} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix},$$
$$\mathbf{j} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \mathbf{k} = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}.$$

(a) Show that $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -\mathbf{1}$.
(b) Show that $\mathbf{ij} = \mathbf{k}$, $\mathbf{jk} = \mathbf{i}$, and $\mathbf{ik} = -\mathbf{j}$.
(c) Show that $\mathbf{xy} = -\mathbf{yx}$ as long as $\mathbf{x}, \mathbf{y} \neq \pm\mathbf{1}$.

Exercise 0.91. A matrix A is **orthogonal** if its transpose is also its inverse. Let $n \in \mathbb{N}^+$ and $\mathrm{O}(n)$ be the set of all orthogonal $n \times n$ matrices.

(a) Show that this matrix is orthogonal:

$$\begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}.$$

(b) Suppose A is orthogonal. Show that $\det A = \pm 1$.

Proofs of some properties of determinants.

Proof of Theorem 0.75. Let A and B satisfy the hypotheses. Write

$$A = \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_{i-1} \\ \mathbf{a}_i \\ \mathbf{a}_{i+1} \\ \vdots \\ \mathbf{a}_n \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_{i-1} \\ c\mathbf{a}_i \\ \mathbf{a}_{i+1} \\ \vdots \\ \mathbf{a}_n \end{pmatrix}.$$

Expand the determinants of both matrices along row i; then

$$\det A = \sum_{j=1}^{n} a_{ij} (-1)^{i+j} \det A_{\hat{i}\hat{j}},$$

while

$$\det B = \sum_{j=1}^{n} \left(c a_{ij}\right) (-1)^{i+j} \det A_{\hat{i}\hat{j}}.$$

Apply the distributive property to factor out the common c, and we have

$$\det B = c \sum_{j=1}^{n} a_{ij} (-1)^{i+j} \det A_{\hat{i}\hat{j}} = c \det A.$$

□

Proof of Lemma 28. We prove the lemma for the case $i = 1$ and $j = 2$; the other cases are similar. We proceed by induction on the dimension n of the matrices.

For the *inductive base*, we consider $n = 2$; we have

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} c & d \\ a & b \end{pmatrix}.$$

Expansion by cofactors gives us $\det A = ad - bc$ and $\det B = bc - ad$. In other words, $\det A = -\det B$.

For the *inductive hypothesis,* we assume that for all matrices of dimension smaller than n, exchanging the first two rows negates the determinant.

For the *inductive step*, expand $\det A$ along column 1. By definition,

$$\det A = \sum_{i=1}^{n} a_{i1} (-1)^{i+1} \det A_{\hat{i}\hat{1}}.$$

Rewrite so that the first two elements are not part of the sum:

$$\begin{aligned}\det A &= a_{11} (-1)^{1+1} \det A_{\hat{1}\hat{1}} + a_{21} (-1)^{2+1} \det A_{\hat{2}\hat{1}} + \sum_{i=3}^{n} a_{i1} (-1)^{i+1} \det A_{\hat{i}\hat{1}} \\ &= a_{11} \det A_{\hat{1}\hat{1}} - a_{21} \det A_{\hat{2}\hat{1}} + \sum_{i=3}^{n} a_{i1} (-1)^{i+1} \det A_{\hat{i}\hat{1}}.\end{aligned}$$

In a similar way, we find that

$$\det B = b_{11} \det B_{\hat{1}\hat{1}} - b_{21} \det B_{\hat{2}\hat{1}} + \sum_{i=3}^{n} b_{i1} (-1)^{i+1} \det B_{\hat{i}\hat{1}}.$$

Recall that the difference between A and B is that we exchanged the first two rows of A to obtain B. Thus, $b_{11} = a_{21}$, $b_{21} = a_{11}$, $B_{\hat{1}\hat{1}} = A_{\hat{2}\hat{1}}$, and $B_{\hat{2}\hat{1}} = A_{\hat{1}\hat{1}}$ (it may take a moment to see why the matrices have that relationship, but it's not hard to see, in the end). For $i \geq 3$, however, $b_{i1} = a_{i1}$, while $B_{\hat{i}\hat{1}}$ is *almost* the same as $A_{\hat{i}\hat{1}}$ — the difference except that the first two rows, $\mathbf{a}_1$ and $\mathbf{a}_2$, are exchanged! The dimensions of these matrices are $n-1$, so the inductive hypothesis applies, and $\det B_{\hat{i}\hat{1}} = -\det A_{\hat{i}\hat{1}}$.Making the appropriate substitutions, we find that

$$\begin{aligned}\det B &= a_{21} \det A_{\hat{2}\hat{1}} - a_{11} \det A_{\hat{1}\hat{1}} + \sum_{i=3}^{n} a_{i1} (-1)^{i+1} \left(-\det A_{\hat{i}\hat{1}}\right) \\ &= -\left[a_{11} \det A_{\hat{1}\hat{1}} + a_{21} \det A_{\hat{2}\hat{1}} + \sum_{i=3}^{n} a_{i1} (-1)^{i+1} \left(\det A_{\hat{i}\hat{1}}\right)\right] \\ &= -\det A.\end{aligned}$$

□

Proof of Lemma 28. Without loss of generality, we assume that the first two rows of the square matrix A are identical; the other cases are similar. Construct a second matrix B by exchanging the first two rows of A. We can write

$$A = \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_1 \\ \mathbf{a}_3 \\ \vdots \\ \mathbf{a}_n \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_1 \\ \mathbf{a}_3 \\ \vdots \\ \mathbf{a}_n \end{pmatrix}.$$

Notice that $A = B$! By substitution, $\det A = \det B$. On the other hand, Lemma 0.78 implies that $\det B = -\det A$. Thus, $\det A = -\det A$, so $2 \det A = 0$, so $\det A = 0$. □

Proof of Lemma 28. Expand the determinant of B along its first row to see that

$$\det B = \sum_{j=1}^{n} \left(a_{1j} + b_j\right) (-1)^{1+j} \det B_{\hat{1}\hat{j}}.$$

The distributive, associative, and commutative properties allow us to rewrite this equation as

$$\det B = \sum_{j=1}^{n} a_{1j} (-1)^{1+j} \det B_{\hat{1}\hat{j}} + \sum_{j=1}^{n} b_j (-1)^{1+j} \det B_{\hat{1}\hat{j}}.$$

If you look at A and B, you will see that $A_{\hat{1}\hat{j}} = B_{\hat{1}\hat{j}}$ for every $j = 1, \ldots, n$: after all, the only difference between A and B lies in the first row, which is by definition excluded from $A_{\hat{1}\hat{j}}$ and $B_{\hat{1}\hat{j}}$. By substitution, then,

$$\det B = \det A + \det \begin{pmatrix} b_1 & b_2 & \cdots & b_n \\ \mathbf{a}_2 & & & \\ \vdots & & & \\ \mathbf{a}_n & & & \end{pmatrix},$$

as claimed. □

Proof of Theorem 0.77. Without loss of generality, we may assume that we constructed B from A by adding a multiple of the second row to the first. That is,

$$A = \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} \mathbf{a}_1 + c\mathbf{a}_2 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{pmatrix}.$$

By Lemma 0.80,

$$\det B = \det \begin{pmatrix} \mathbf{a}_1 + c\mathbf{a}_2 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{pmatrix} = \det \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{pmatrix} + \det \begin{pmatrix} c\mathbf{a}_2 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{pmatrix}.$$

Now apply Theorem 0.75 and Lemma 0.79 to see that

$$\det B = \det A + c \det \begin{pmatrix} \mathbf{a}_2 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{pmatrix} = \det A + c \cdot 0 = \det A.$$

□

Proof of Theorem 0.82. If $\det A = 0$, then Theorem 0.81 tells us that we can find real numbers

$c_2, \ldots, c_n$ such that $\mathbf{a}_1 = \sum_{k=2}^{n} c_k \mathbf{a}_k$. By properties of matrix multiplication,

$$\mathbf{a}_1 B = \left(\sum_{k=2}^{n} c_k \mathbf{a}_k \right) B = \sum_{k=2}^{n} c_k \left(\mathbf{a}_k B \right).$$

Notice that $\mathbf{a}_i B$ is the ith row of AB, so this new equation shows that the first row of AB is a linear combination of the other rows. Theorem 0.81 again implies that $\det(AB) = 0$.

Now suppose $\det A \neq 0$. A fact of linear algebra that we do not repeat here is that we can write

$$A = E_1 E_2 \cdots E_m,$$

where we constrsuct each E_i by applying one of the operations of Theorem 0.75, Lemma 0.78, or Lemma 0.80 to I_n. Thus,

$$\det(AB) = \det\left(E_1 \cdots E_m B \right).$$

Let $C = E_2 \cdots E_m B$; we have $\det(AB) = \det(E_1 C)$. We now consider three possible values of E_1.

Case 1: If E_1 is the result of swapping two rows of I_n, then $\det E_1 = -1$. On the other hand, $E_1 C$ is the same as C, except that two rows of C are swapped — the same two rows as in E_1, in fact. So $\det(E_1 C) = -\det C = \det E_1 \cdot \det C$.

Case 2: If E_1 is the result of multiplying a row of I_n by a constant $c \in \mathbb{R}$, then $\det E_1 = c$. On the other hand, $E_1 C$ is the same as C, except that a row of C has been multiplied by a constant $c \in \mathbb{R}$ — the same row as in E_1, in fact. So $\det(E_1 C) = c \det C = \det E_1 \cdot \det C$.

Case 3: If E_1 is the result of adding a multiple of a row of I_n to another row, then $\det E_1 = \det I_n = 1$. On the other hand, $E_1 C$ is the same as C, except that a multiple of a row of C has been added to another row of C — the same two rows as E_1, in fact, and the same multiple. So $\det(E_1 C) = \det C = \det E_1 \det C$.

In each case, we found that $\det(E_1 C) = \det E_1 \det C$. Thus, $\det(AB) = \det E_1 \cdot \det(E_2 \cdots E_m B)$. We now repeat this process for each of the E_i, obtaining

$$\det(AB) = \det E_1 \cdots \det E_m \det B = \det A \det B.$$

□

Part I
Monoids and groups

Chapter 1:
Monoids

Algebra was created to solve problems. Like other branches of mathematics, it started off solving very applied problems of a certain type; that is, polynomial equations. When studying algebra the last few years, you have focused on techniques necessary for solving the simplest examples of polynomial equations: for example, factoring, isolating a variable, and taking roots.

These techniques work well for linear equations, and if you massage the problem a bit, they work well for quadratic equations, too. It's quite hard to apply these techniques to polynomials of degree three and four, however, and impossible to apply them to all polynomials of degree five or higher. You might say that these techniques do not scale well. Because of this, algebra took a radically different turn in the 19th century (pun intended), one that develops not just techniques, but structures and viewpoints that can be used to solve a vast array of problems, many of which are surprisingly different.

This chapter introduces some new, but important algebraic ideas. We will try to be intuitive, but don't confuse "intuitive" with "vague"; we will maintain precision. We will use very concrete examples. True, these examples are probably not as concrete as you might like, but believe me when I tell you that the examples I will use are more concrete than the usual presentation. One goal is to get you to use these examples when thinking about the more general ideas later on. It will be important not only that you reproduce what you read here, but that you explore and play with the ideas and examples, specializing or generalizing them as needed to attack new problems.

Success in this course will require you to balance these inductive and deductive approaches.

1.1: From integers and monomials to monoids

We now move from one set that you may consider to be "arithmetical" to another that you will definitely recognize as "algebraic". In doing so, we will notice a similarity in the mathematical structure. That similarity will motivate our first steps into modern algebra, with monoids.

Monomials

Let x represent an unknown quantity. The set of "univariate monomials in x" is

$$\mathbb{M} = \{x^a : a \in \mathbb{N}\}, \tag{4}$$

where x^a, a "monomial", represents precisely what you'd think: the product of a copies of x. In other words,

$$x^a \quad = \quad \prod_{i=1}^{a} x \quad = \quad \underbrace{x \cdot x \cdot \cdots \cdot x}_{n \text{ times}}.$$

We can extend this notion. Let x_1, x_2, …, x_n represent unknown quantities. The set of "multivariate monomials in x_1, x_2, …, x_n" is

$$\mathbb{M}_n = \left\{ \prod_{i=1}^{m} \left(x_1^{a_{i1}} x_2^{a_{i2}} \cdots x_n^{a_{in}} \right) : m, a_{ij} \in \mathbb{N} \right\}. \tag{5}$$

("Univariate" means "one variable"; "multivariate" means "many variables".) For monomials, we allow neither coefficients nor negative exponents. The definition of $\mathbb{M}_n$ indicates that any of its elements is a "product of products".

Example 1.1. The following are monomials:

$$x^2, \quad 1 = x^0 = x_1^0 x_2^0 \cdots x_n^0, \quad x^2 y^3 x y^4.$$

Notice from the last product that the variables need not commute under multiplication; that depends on what they represent. This is consistent with the definition of $\mathbb{M}_n$, each of whose elements is a product of products. We could write $x^2y^3xy^4$ in those terms as

$$\left(x^2y^3\right)\left(xy^4\right) = \prod_{i=1}^{m}\left(x_1^{a_{i1}} x_2^{a_{i2}}\right)$$

with $m = 2$, $a_{11} = 2$, $a_{12} = 3$, $a_{21} = 1$, and $a_{22} = 4$.

The following are not monomials:

$$x^{-1} = \frac{1}{x}, \quad \sqrt{x} = x^{\frac{1}{2}}, \quad \sqrt[3]{x^2} = x^{\frac{2}{3}}.$$

Similarities between $\mathbb{M}$ *and* $\mathbb{N}$

We are interested in similarities between $\mathbb{N}$ and $\mathbb{M}$. Why? Suppose that we can identify a structure common to the two sets. If we make the obvious properties of this structure precise, we can determine non-obvious properties that must be true about $\mathbb{N}$, $\mathbb{M}$, and any other set that adheres to the structure.

If we can prove a fact about a structure,
then we don't have to re-prove that fact for all its elements.
This ***saves*** *time and* ***increases*** *understanding.*

It is harder at first to think about general structures rather than concrete objects, but time, effort, and determination bring agility.

To begin with, what operation(s) should we normally associate with $\mathbb{M}$? We normally associate addition and multiplication with the natural numbers, but the monomials are *not* closed under addition. After all, $x^2 + x^4$ is a *polynomial*, not a monomial. On the other hand, $x^2 \cdot x^4$ is a monomial, and in fact $x^a x^b \in \mathbb{M}$ for any choice of $a, b \in \mathbb{N}$. This is true about monomials in any number of variables.

Lemma 1.2. Let $n \in \mathbb{N}^+$. Both $\mathbb{M}$ and $\mathbb{M}_n$ are closed under multiplication.

Proof for $\mathbb{M}$. Let $t, u \in \mathbb{M}$. By definition, there exist $a, b \in \mathbb{N}$ such that $t = x^a$ and $u = x^b$. By definition of monomial multiplication, we see that

$$tu = x^{a+b}.$$

Since addition is closed in $\mathbb{N}$, the expression $a + b$ simplifies to a natural number. Call this number c. By substitution, $tu = x^c$. This has the form of a univariate monomial; compare it

with the description of a monomial in equation (4). So, $tu \in \mathbb{M}$. Since we chose t and u to be arbitrary elements of $\mathbb{M}$, and found their product to be an element of $\mathbb{M}$, we conclude that $\mathbb{M}$ is closed under multiplication. □

Easy, right? We won't usually state all those steps explicitly, but we want to do so at least once.

What about $\mathbb{M}_n$? The lemma claims that multiplication is closed there, too, but we haven't proved that yet. I wanted to separate the two, to show how operations you take for granted in the univariate case don't work so well in the multivariate case. The problem here is that the variables might not commute under multiplication. If we knew that they did, we could write something like,

$$tu = x_1^{a_1+b_1} \cdots x_n^{a_n+b_n},$$

so long as the a's and the b's were defined correctly. Unfortunately, if we assume that the vairables *are* commutative, then we don't prove the statement for everything that we would like. This requires a little more care in developing the argument. Sometimes, it's just a game of notation, as it will be here.

Proof for $\mathbb{M}_n$. Let $t, u \in \mathbb{M}_n$. By definition, we can write

$$t = \prod_{i=1}^{m_t} \left(x_1^{a_{i1}} \cdots x_n^{a_{in}} \right) \quad \text{and} \quad u = \prod_{i=1}^{m_u} \left(x_1^{b_{i1}} \cdots x_n^{b_{in}} \right).$$

(We give subscripts to m_t and m_u because t and u might have a different number of elements in their product. Since m_t and m_u are not the same symbol, it's possible they have a different value.) By substitution,

$$tu = \left(\prod_{i=1}^{m_t} \left(x_1^{a_{i1}} \cdots x_n^{a_{in}} \right) \right) \left(\prod_{i=1}^{m_u} \left(x_1^{b_{i1}} \cdots x_n^{b_{in}} \right) \right).$$

Intuitively, you want to declare victory; we've written tu as a product of variables, right? All we see are variables, organized into two products.

Unfortunately, we're not quite there yet. To show that $tu \in \mathbb{M}_n$, we must show that we can write it as *one* product of a list of products, rather than two. This turns out to be as easy as making the symbols do what your head is telling you: two lists of products of variables, placed side by side, make one list of products of variables. To show that it's one list, we must identify explicitly how many "small products" are in the "big product". There are m_t in the first, and m_u in the second, which makes $m_t + m_u$ in all. So we know that we should be able to write

$$tu = \prod_{i=1}^{m_t+m_u} \left(x_1^{c_{i1}} \cdots x_n^{c_{in}} \right) \tag{6}$$

for appropriate choices of c_{ij}. The hard part now is identifying the correct values of c_{ij}.

In the list of products, the first few products come from t. How many? There are m_t from t. The rest are from u. We can specify this precisely using a piecewise function:

$$c_{ij} = \begin{cases} a_{ij}, & 1 \le i \le m_t \\ b_{ij}, & m_t < i. \end{cases}$$

Specifying c_{ij} this way justifies our claim that tu has the form shown in equation (6). That satisfies the requirements of $\mathbb{M}_n$, so we can say that $tu \in \mathbb{M}_n$. Since t and u were chosen arbitrarily from $\mathbb{M}_n$, it is closed under multiplication. □

You can see that life is a little harder when we don't have all the assumptions we would like to make; it's easier to prove that $\mathbb{M}_n$ is closed under multiplication if the variables commute under multiplication; we can simply imitate the proof for $\mathbb{M}$. You will do this in one of the exercises.

As with the proof for $\mathbb{M}$, we were somewhat pedantic here; don't expect this level of detail all the time. Pedantry has the benefit that you don't have to read between the lines. That means you don't have to think much, only recall previous facts and apply very basic logic. However, pedantry also makes proofs long and boring. While you could shut down much of your brain while reading a pedantic proof, that would be counterproductive. Ideally, you want to reader to *think* while reading a proof, so shutting down the brain is bad. Thus, a good proof does not recount every basic definition or result for the reader, but requires her to make basic recollections and inferences.

Let's look at two more properties.

Lemma 1.3. Let $n \in \mathbb{N}^+$. Multiplication in $\mathbb{M}$ satifies the commutative property. Multiplication in both $\mathbb{M}$ and $\mathbb{M}_n$ satisfies the associative property.

Proof. We show this to be true for $\mathbb{M}$; the proof for $\mathbb{M}_n$ we will omit (but it can be done as it was above). Let $t, u, v \in \mathbb{M}$. By definition, there exist $a, b, c \in \mathbb{N}$ such that $t = x^a$, $u = x^b$, and $v = x^c$. By definition of monomial multiplication and by the commutative property of addition in $\mathbb{M}$, we see that

$$tu = x^{a+b} = x^{b+a} = ut.$$

As t and u were arbitrary, multiplication of univariate monomials is commutative.

By definition of monomial multiplication and by the associative property of addition in $\mathbb{N}$, we see that

$$\begin{aligned} t(uv) &= x^a \left(x^b x^c\right) = x^a x^{b+c} \\ &= x^{a+(b+c)} = x^{(a+b)+c} \\ &= x^{a+b} x^c = (tu)v. \end{aligned}$$

□

You might ask yourself, *Do I have to show* every *step?* That depends on what the reader needs to understand the proof. In the equation above, it *is* essential to show that the commutative and associative properties of multiplication in $\mathbb{M}$ depend strictly on the commutative and associative properties of addition in $\mathbb{N}$. Thus, the steps

$$x^{a+b} = x^{b+a} \quad \text{and} \quad x^{a+(b+c)} = x^{(a+b)+c},$$

with the parentheses as indicated, are absolutely crucial, and cannot be omitted from a good proof.[7]

[7] Of course, a professional mathematician would not even prove these things in a paper, because they are well-known

Another property the natural numbers have is that of an identity: both additive and multiplicative. Since we associate only multiplication with the monomials, we should check whether they have a multiplicative identity. I hope this one doesn't surprise you!

Lemma 1.4. Both $\mathbb{M}$ and $\mathbb{M}_n$ have $1 = x^0 = x_1^0 x_2^0 \cdots x_n^0$ as a multiplicative identity.

We won't bother proving this one, but leave it to the exercises.

Monoids

There are quite a few other properties that the integers and the monomials share, but the three properties we have mentioned here are already quite interesting, and as such are precisely the ones we want to highlight. This motivates the following definition.

Definition 1.5. Let M be a set, and $\circ$ an operation on M. We say that the pair $(M, \circ)$ is a **monoid** if it satisfies the following properties:

(closed) for any $x, y \in M$, we have $x \circ y \in M$;
(associative) for any $x, y, z \in M$, we have $(x \circ y) \circ z = x \circ (y \circ z)$; and
(identity) there exists an **identity element** $e \in M$ such that for any $x \in M$, we have $e \circ x = x \circ e = x$.

We may also say that M is a **monoid under** $\circ$.

So far, then, we know the following:

Theorem 1.6. $\mathbb{N}$ is a monoid under both addition and multiplication, while $\mathbb{M}$ and $\mathbb{M}_n$ are monoids under multiplication.

Proof. For $\mathbb{N}$, this is part of its definition. For $\mathbb{M}$ and $\mathbb{M}_n$, see Lemmas 1.2, 1.3, and 1.4. □

Generally, we don't write the operation in conjunction with the set; we write the set alone, leaving it to the reader to infer the operation. In some cases, this might lead to ambiguity; after all, both $(\mathbb{N}, +)$ and $(\mathbb{N}, \times)$ are monoids, so which should we prefer? We will prefer $(\mathbb{N}, +)$ as the usual monoid associated with $\mathbb{N}$. Thus, we can write that $\mathbb{N}$, $\mathbb{M}$, and $\mathbb{M}_n$ are examples of monoids: the first under addition, the others under multiplication.

What other mathematical objects are examples of monoids?

Example 1.7. Let $m, n \in \mathbb{N}^+$. The set of $m \times n$ matrices with integer entries, written $\mathbb{Z}^{m \times n}$, satisfies properties that make it a monoid under addition:

- closure is guaranteed by the definition;
- the associative property is guaranteed by the associative property of its elements; and
- the additive identity is $\mathbb{0}$, the zero matrix, by Theorem 0.61;

Example 1.8. The set of square matrices with integer entries $\mathbb{Z}^{m \times m}$ satisfies properties that make it a monoid under multiplication:for multiplication,

- closure is guaranteed by the definition;

and easy. On the other hand, a good professional mathematician *would* feel compelled to include in a proof steps that include novel and/or difficult information.

- the associative property is guaranteed by Theorem 0.64; and
- the multiplicative identity is I_n, by Theorem 0.61.

Your professor almost certainly didn't *call* the set of square matrices a monoid at the time.

Here's an example you probably *haven't* seen before.

Example 1.9. Let S be a set, and let F_S be the set of all functions mapping S to itself, with the proviso that for any $f \in F_S$, $f(s)$ is defined for every $s \in S$. We can show that F_S is a monoid under composition of functions, since

- for any $f, g \in F_S$, we also have $f \circ g \in F_S$, where $f \circ g$ is the function h such that for any $s \in S$,
$$h(s) = (f \circ g)(s) = f(g(s))$$
(notice how important it was that $g(s)$ have a defined value regardless of the value of s);
- for any $f, g, h \in F_S$, we have $(f \circ g) \circ h = f \circ (g \circ h)$, since for any $s \in S$,
$$((f \circ g) \circ h)(s) = (f \circ g)(h(s)) = f(g(h(s)))$$
and
$$(f \circ (g \circ h))(s) = f((g \circ h)(s)) = f(g(h(s)));$$
- if we consider the function $\iota \in F_S$ where $\iota(s) = s$ for all $s \in S$, then for any $f \in F_S$, we have $\iota \circ f = f \circ \iota = f$, since for any $s \in S$,
$$(\iota \circ f)(s) = \iota(f(s)) = f(s)$$
and
$$(f \circ \iota)(s) = f(\iota(s)) = f(s)$$
(we can say that $\iota(f(s)) = f(s)$ because $f(s) \in S$).

Although monoids are useful, they don't capture all the properties that interest us. Not all the properties we found for $\mathbb{N}$ will hold for $\mathbb{M}$, let alone for all monoids. After all, monoids characterize the properties of a set with respect to *only one* operation. Because of this, they cannot describe properties based on two operations.

For example, the Division Theorem requires *two* operations: multiplication (by the quotient) and addition (of the remainder). So, there is no "Division Theorem for Monoids"; it simply doesn't make sense in the context. If we want to generalize the Division Theorem to other sets, we will need a more specialized structure. We will actually meet one later! (in Section 7.4.)

Here is one useful property that we can prove already. A natural question to ask about monoids is whether the identity of a monoid is unique. (We asked it about the matrices, back in Section 0.3.) It isn't hard to show that it is.

Theorem 1.10. Suppose that M is a monoid, and there exist $e, i \in M$ such that $ex = x$ and $xi = x$ for all $x \in M$. Then $e = i$, so that the identity of a monoid is unique.

"Unique" in mathematics means *exactly one*. To prove uniqueness of an object x, you consider a generic object y that shares all the properties of x, then reason to show that $x = y$. This is not a

contradiction, because we didn't assume that $x \neq y$ in the first place; we simply wondered about a generic y. We did the same thing with the Division Theorem (Theorem 0.34 on page 13).

Proof. Suppose that e is a left identity, and i is a right identity. Since i is a right identity, we know that

$$e = ei.$$

Since e is a left identity, we know that

$$ei = i.$$

By substitution,

$$e = i.$$

We chose an arbitrary left identity of M and an arbitrary right identity of M, and showed that they were in fact the same element. Hence left identities are also right identities. This implies in turn that there is only one identity: any identity is both a left identity and a right identity, so the argument above shows that any two identities are in fact identical. □

Exercises.

Exercise 1.11. Is $\mathbb{N}$ a monoid under:
(a) subtraction?
(b) division?
Be sure to explain your answer.

Exercise 1.12. Is $\mathbb{Z}$ a monoid under:
(a) addition?
(b) subtraction?
(c) multiplication?
(d) division?
Be sure to explain your answer.

Exercise 1.13. Consider the set $B = \{F, T\}$ with the operation $\vee$ where

$$F \vee F = F$$
$$F \vee T = T$$
$$T \vee F = T$$
$$T \vee T = T.$$

This operation is called **Boolean or**.

Is $(B, \vee)$ a monoid? If so, explain how it justifies each property.

Exercise 1.14. Consider the set $B = \{F, T\}$ with the operation $\oplus$ where

$$F \oplus F = F$$
$$F \oplus T = T$$
$$T \oplus F = T$$
$$T \oplus T = F.$$

This operation is called **Boolean exclusive or**, or **xor** for short.

Is $(B, \oplus)$ a monoid? If so, explain how it justifies each property.

Exercise 1.15. Suppose multiplication of x and y commutes. Show that multiplication in $\mathbb{M}_n$ is both closed and associative.

Exercise 1.16.

(a) Show that $\mathbb{N}[x]$, the ring of polynomials in one variable with integer coefficients, is a monoid under addition.

(b) Show that $\mathbb{N}[x]$ is also a monoid if the operation is multiplication.

(c) Explain why we can replace $\mathbb{N}$ by $\mathbb{Z}$ and the argument would remain valid. (*Hint:* think about the *structure* of these sets.)

Exercise 1.17. Recall the lattice L from Exercise 0.52.

(a) Show that L is a monoid under the addition defined in that exercise.

(b) Show that L is a monoid under the multiplication defined in that exercise.

Exercise 1.18. Let A be a set of symbols, and L the set of all finite sequences that can be constructed using elements of A. Let $\circ$ represent *concatenation of lists.* For example, $(a, b) \circ (c, d, e, f) = (a, b, c, d, e, f)$. Show that $(L, \circ)$ is a monoid.

Definition 1.19. For any set S, let $P(S)$ denote the set of all subsets of S. We call this the **power set** of S.

Exercise 1.20.

(a) Suppose $S = \{a, b\}$. Compute $P(S)$, and show that it is a monoid under $\cup$ (union).

(b) Let S be *any* set. Show that $P(S)$ is a monoid under $\cup$ (union).

Exercise 1.21.

(a) Suppose $S = \{a, b\}$. Compute $P(S)$, and show that it is a monoid under $\cap$ (intersection).

(b) Let S be *any* set. Show that $P(S)$ is a monoid under $\cap$ (intersection).

Exercise 1.22.

(a) Fill in each blank of Figure 1.1 with the justification.

(b) Is $(\mathbb{N}, \mathrm{lcm})$ also a monoid? If so, do we have to change anything about the proof? If not, which property fails?

Exercise 1.23. Recall the usual ordering $<$ on $\mathbb{M}$: $x^a < x^b$ if $a < b$. Show that this is a well-ordering.

Remark 1.24. While we can define a well-ordering on $\mathbb{M}_n$, it is a much more complicated proposition, which we take up in Section 11.2.

Exercise 1.25. In Exercise 0.46, you showed that divisibility is transitive in the integers.

(a) Show that divisibility is transitive in *any* monoid; that is, if M is a monoid, $a, b, c \in M$, $a \mid b$, and $b \mid c$, then $a \mid c$.

Claim: $(\mathbb{N}^+, \text{lcm})$ is a monoid. Note that the operation here looks unusual: instead of something like $x \circ y$, you're looking at $\text{lcm}(x, y)$.
Proof:

1. First we show closure.
   (a) Let $a, b \in$_____, and let $c = \text{lcm}(a, b)$.
   (b) By definition of _____, $c \in \mathbb{N}$.
   (c) By definition of _____, $\mathbb{N}$ is closed under lcm.
2. Next, we show the associative property. This is one is a bit tedious...
   (a) Let $a, b, c \in$_____.
   (b) Let $m = \text{lcm}(a, \text{lcm}(b, c))$, $n = \text{lcm}(\text{lcm}(a, b), c)$, and $\ell = \text{lcm}(b, c)$. By _____, we know that $\ell, m, n \in \mathbb{N}$.
   (c) We claim that $\text{lcm}(a, b)$ divides m.
      i. By definition of _____, both a and $\text{lcm}(b, c)$ divide m.
      ii. By definition of _____, we can find x such that $m = ax$.
      iii. By definition of _____, both b and c divide m.
      iv. By definition of _____, we can find y such that $m = by$
      v. By definition of _____, both a and b divide m.
      vi. By Exercise _____, $\text{lcm}(a, b)$ divides m.
   (d) Recall that _____ divides m. Both $\text{lcm}(a, b)$ and _____ divide m. (Both blanks expect the same answer.)
   (e) By definition of _____, $n \leq m$.
   (f) A similar argument shows that $m \leq n$; by Exercise _____, $m = n$.
   (g) By _____, $\text{lcm}(a, \text{lcm}(b, c)) = \text{lcm}(\text{lcm}(a, b), c)$.
   (h) Since $a, b, c \in \mathbb{N}$ were arbitrary, we have shown that lcm is associative.
3. Now, we show the identity property.
   (a) Let $a \in$_____.
   (b) Let $\iota =$_____.
   (c) By arithmetic, $\text{lcm}(a, \iota) = a$.
   (d) By definition of _____, ι is the identity of $\mathbb{N}$ under lcm.
4. We have shown that $(\mathbb{N}, \text{lcm})$ satisfies the properties of a monoid.

Figure 1.1. Material for Exercise 1.22

(b) In fact, you don't need all the properties of a monoid for divisibility to be transitive! Which properties *do* you need?

1.2: Isomorphism

We've seen that several important sets share the monoid structure. In particular, $(\mathbb{N}, +)$ and $(\mathbb{M}, \times)$ are very similar. Are they in fact identical *as monoids?* If so, the technical word for this is *isomorphism*. How can we determine whether two monoids are isomorphic? We will look for a way to determine whether their operations behave the same way.

Imagine two offices. How would you decide if the offices were equally suitable for a certain job? First, you would need to know what tasks have to be completed, and what materials you need for those tasks. For example, if your job required you to keep books for reference, you

would want to find a bookshelf in the office. If it required you to write, you would need a desk, and perhaps a computer. If it required you to communicate with people in other locations, you might need a phone. Having made such a list, you would then want to compare the two offices. If they both had the equipment you needed, you'd think they were both suitable for the job at hand. It wouldn't really matter how the offices satisfied the requirements; if one had a desk by the window, and the other had it on the side opposite the window, that would be okay. If one office lacked a desk, however, it wouldn't be up to the required job.

Deciding whether two sets are isomorphic is really the same idea. First, you decide what structure the sets have, which you want to compare. (So far, we've only studied monoids, so for now, we care only whether the sets have the same monoid structure.) Next, you compare how the sets satisfy those structural properties. If you're looking at finite monoids, an exhaustive comparison might work, but exhaustive methods tend to become exhausting, and don't scale well to large sets. Besides, we deal with infinite sets like $\mathbb{N}$ and $\mathbb{M}$ often enough that we need a non-exhaustive way to compare their structure. Functions turn out to be just the tool we need.

How so? Let S and T be any two sets. Recall that a **function** $f : S \to T$ is a relation that sends every input $x \in S$ to precisely one value in T, the output $f(x)$. You have probably heard the geometric interpretation of this: f passes the "vertical line test." You might suspect at this point that we are going to generalize the notion of function to something more general, just as we generalized $\mathbb{Z}$, $\mathbb{M}$, etc. to monoids. To the contrary; we will *specialize* the notion of a function in a way that tells us important information about a monoid.

Suppose M and N are monoids. If they are isomorphic, their monoid structure is identical, so we ought to be able to build a function that maps elements with a certain behavior in M to elements with the same behavior in N. (Table to table, phone to phone.) What does that mean? Let $x, y, z \in M$ and $a, b, c \in N$. Suppose that $f(x) = a$, $f(y) = b$, $f(z) = c$, and $xy = z$. If M and N have the same structure as monoids, then:

- since $xy = z$,
- we want $ab = c$, or

$$f(x) f(y) = f(z)$$

Substituting xy for z suggests that we want the property

$$f(x) f(y) = f(xy).$$

Of course, we would also want to preserve the identity: f ought to be able to map the identity of M to the identity of N. In addition, just as we only need one table in the office, we want to make sure that there is a one-to-one correspondence between the elements of the monoids. If we're going to reverse the function, it needs to be onto. That more or less explains why we define isomorphism in the following way:

Definition 1.26. Let $(M,\times)$ and $(N,+)$ be monoids. If there exists a function $f : M \longrightarrow N$ such that

- $f(1_M) = 1_N$ *(f preserves the identity)*

and

- $f(xy) = f(x) + f(y)$ for all $x, y \in M$, *(f preserves the operation)*

then we call f a **homomorphism**. If f is also a bijection, then we say that M is **isomorphic** to N, write $M \cong N$, and call f an **isomorphism**.[a] (A **bijection** is a function that is both one-to-one and onto.)

[a] The word *homomorphism* comes from the Greek words for *same* and *shape;* the word *isomorphism* comes from the Greek words for *identical* and *shape*. The *shape* is the effect of the operation on the elements of the group. Isomorphism shows that the group operation behaves the same way on elements of the range as on elements of the domain.

If you do not remember the definitions of one-to-one and onto, see Definition 0.32 on page 12. Another way of saying that a function $f : S \to U$ is onto is to say that $f(S) = U$; that is, the **image** of S is *all* of U, or that *every* element of U corresponds via f to some element of S.

We used $(M,\times)$ and $(N,+)$ in the definition partly to suggest our goal of showing that $\mathbb{M}$ and $\mathbb{N}$ are isomorphic, but also because they could stand for *any* monoids. You will see in due course that not all monoids are isomorphic, but first let's see about $\mathbb{M}$ and $\mathbb{N}$.

Example 1.27. We claim that $(\mathbb{M},\times)$ is isomorphic to $(\mathbb{N},+)$. To see why, let $f : \mathbb{M} \longrightarrow \mathbb{N}$ by

$$f(x^a) = a.$$

First we show that f is a bijection.

To see that it is one-to-one, let $t, u \in \mathbb{M}$, and assume that $f(t) = f(u)$. By definition of $\mathbb{M}$, $t = x^a$ and $u = x^b$ for $a, b \in \mathbb{N}$. Susbtituting this into $f(t) = f(u)$, we find that $f(x^a) = f\left(x^b\right)$. The definition of f allows us to rewrite this as $a = b$. In this case, $x^a = x^b$, so $t = u$. We assumed that $f(t) = f(u)$ for arbitrary $t, u \in \mathbb{M}$, and showed that $t = u$; that proves f is one-to-one.

To see that f is onto, let $a \in \mathbb{N}$. We need to find $t \in \mathbb{M}$ such that $f(t) = a$. Which t should we choose? We want $f\left(x^?\right) = a$, and $f\left(x^?\right) = ?$, so the "natural" choice seems to be $t = x^a$. That would certainly guarantee $f(t) = a$, but can we actually find such an object t in $\mathbb{M}$? Since $x^a \in \mathbb{M}$, we can in fact make this choice! We took an arbitrary element $a \in \mathbb{N}$, and showed that f maps some element of $\mathbb{M}$ to a; that proves f is onto.

So f is a bijection. Is it also an isomorphism? First we check that f preserves the operation. Let $t, u \in \mathbb{M}$.[8] By definition of $\mathbb{M}$, $t = x^a$ and $u = x^b$ for $a, b \in \mathbb{N}$. We now manipulate $f(tu)$ using definitions and substitutions to show that the operation is preserved:

$$\begin{aligned} f(tu) &= f\left(x^a x^b\right) = f\left(x^{a+b}\right) \\ &= a + b \\ &= f(x^a) + f\left(x^b\right) = f(t) + f(u). \end{aligned}$$

[8] The definition uses the variables x and y, but those are just letters that stand for arbitrary elements of M. Here $M = \mathbb{M}$ and we can likewise choose any two letters we want to stand in place of x and y. It would be a very bad idea to use x when talking about an arbitrary element of $\mathbb{M}$, because there *is* an element of $\mathbb{M}$ called x. So we choose t and u instead.

Does f also preserve the identity? We usually write the identity of $M = \mathbb{M}$ as the symbol 1, but recall that this is a convenient stand-in for x^0. On the other hand, the identity (under addition) of $N = \mathbb{N}$ is the number 0. We use this fact to verify that f preserves the identity:

$$f(1_M) = f(1) = f\left(x^0\right) = 0 = 1_N.$$

(We don't usually write 1_M and 1_N, but I'm doing it here to show explicitly how this relates to the definition.)

We have shown that there exists a bijection $f : \mathbb{M} \longrightarrow \mathbb{N}$ that preserves the operation and the identity. We conclude that $\mathbb{M} \cong \mathbb{N}$.

On the other hand, is $(\mathbb{N}, +) \cong (\mathbb{N}, \times)$? You might think this is easier to verify, since the sets are the same. Let's see what happens.

Example 1.28. Suppose there *does* exist an isomorphism $f : (\mathbb{N}, +) \to (\mathbb{N}, \times)$. What would have to be true about f? Let $a \in \mathbb{N}$ such that $f(1) = a$; after all, f has to map 1 to *something!* An isomorphism must preserve the operation, so

$$\begin{aligned} &f(2) = f(1+1) = f(1) \times f(1) = a^2 \text{ and} \\ &f(3) = f(1+(1+1)) = f(1) \times f(1+1) = a^3, \text{ so that} \\ &f(n) = \cdots = a^n \text{ for } \textit{any } n \in \mathbb{N}. \end{aligned}$$

So f sends *every* integer in $(\mathbb{N}, +)$ to a power of a.

Think about what this implies. For f to be a bijection, it would have to be onto, so *every* element of $(\mathbb{N}, \times)$ would *have* to be an integer power of a. ***This is false!*** After all, 2 is not an integer power of 3, and 3 is not an integer power of 2.

The claim was correct: $(\mathbb{N}, +) \ncong (\mathbb{N}, \times)$.

Exercises.

Exercise 1.29. Show that the monoids "Boolean or" and "Boolean xor" from Exercises 1.13 and 1.14 are *not* isomorphic.

Exercise 1.30. Let $(M, \times)$, $(N, +)$, and $(P, \sqcap)$ be monoids.

(a) Show that the identity function $\iota(x) = x$ is an isomorphism on M.

(b) Suppose that we know $(M, \times) \cong (N, +)$. That means there is an isomorphism $f : M \to N$. One of the requirements of isomorphism is that f be a bijection. Recall from previous classes that this means f has an inverse *function*, $f^{-1} : N \to M$. Show that f^{-1} is an isomorphism.

(c) Suppose that we know $(M, \times) \cong (N, +)$ and $(N, +) \cong (P, \sqcap)$. As above, we know there exist isomorphisms $f : M \to N$ and $g : N \to P$. Let $h = g \circ f$; that is, h is the composition of the functions g and f. Explain why $h : M \to P$, and show that h is also an isomorphism.

(d) Explain how (a), (b), and (c) prove that isomorphism is an equivalence relation.

1.3: Direct products

It might have occurred to you that a multivariate monomial is really a vector of univariate monomials. (Pat yourself on the back if so.) If not, here's an example:

$$x_1^6 x_2^3 \text{ looks an awful lot like } \left(x^6, x^3\right).$$

So, we can view any element of $\mathbb{M}_n$ as a list of n elements of $\mathbb{M}$. In fact, if you multiply two multivariate monomials, you would have a corresponding result to multiplying two vectors of univariate monomials componentwise:

$$\left(x_1^6 x_2^3\right)\left(x_1^2 x_2\right) = x_1^8 x_2^4 \quad \text{and} \quad \left(x^6, x^3\right) \times \left(x^2, x\right) = \left(x^8, x^4\right).$$

Last section, we showed that $(\mathbb{M}, \times) \cong (\mathbb{N}, +)$, so it should make sense that we can simplify this idea even further:

$$x_1^6 x_2^3 \text{ looks an awful lot like } (6,3), \text{ and in fact } (6,3) + (2,1) = (8,4).$$

We can do this with other sets, as well.

Definition 1.31. Let $r \in \mathbb{N}^+$ and $S_1, S_2, \ldots, S_r$ be sets. The **Cartesian product** of $S_1, \ldots, S_r$ is the set of all lists of r elements where the ith entry is an element of S_i; that is,

$$S_1 \times \cdots \times S_r = \{(s_1, s_2, \ldots, s_n) : s_i \in S_i\}.$$

Example 1.32. We already mentioned a Cartesian product of two sets in the introduction to this chapter. Another example would be $\mathbb{N} \times \mathbb{M}$; elements of $\mathbb{N} \times \mathbb{M}$ include $\left(2, x^3\right)$ and $\left(0, x^5\right)$. In general, $\mathbb{N} \times \mathbb{M}$ is the set of all ordered pairs where the first entry is a natural number, and the second is a monomial.

If we can preserve the structure of the underlying sets in a Cartesian product, we call it a *direct product*.

Definition 1.33. Let $r \in \mathbb{N}^+$ and $M_1, M_2, \ldots, M_r$ be monoids. The **direct product** of $M_1, \ldots, M_r$ is the pair

$$(M_1 \times \cdots \times M_r, \otimes)$$

where $M_1 \times \cdots \times M_r$ is the usual Cartesian product, and $\otimes$ is the "natural" operation on $M_1 \times \cdots \times M_r$.

What do we mean by the "natural" operation on $M_1 \times \cdots \times M_r$? Let $x, y \in M_1 \times \cdots \times M_r$; by definition, we can write

$$x = (x_1, \ldots, x_r) \quad \text{and} \quad y = (y_1, \ldots, y_r)$$

where each x_i and each y_i is an element of M_i. Then

$$x \otimes y = (x_1 y_1, x_2 y_2, \ldots, x_r y_r)$$

where each product $x_i y_i$ is performed according to the operation that makes the corresponding M_i a monoid.

Example 1.34. Recall that $\mathbb{N} \times \mathbb{M}$ is a Cartesian product; if we consider the monoids $(\mathbb{N}, \times)$ and $(\mathbb{M}, \times)$, we can show that the direct product is a monoid, much like $\mathbb{N}$ and $\mathbb{M}$! To see why, we check each of the properties.

(closure) Let $t, u \in \mathbb{N} \times \mathbb{M}$. By definition, we can write $t = (a, x^\alpha)$ and $u = \left(b, x^\beta\right)$ for appropriate $a, \alpha, b, \beta \in \mathbb{N}$. Then

$$\begin{aligned} tu &= (a, x^\alpha) \otimes \left(b, x^\beta\right) \\ &= \left(ab, x^\alpha x^\beta\right) && \text{(def. of } \otimes) \\ &= \left(ab, x^{\alpha+\beta}\right) \in \mathbb{N} \times \mathbb{M}. \end{aligned}$$

We took two arbitrary elements of $\mathbb{N} \times \mathbb{M}$, multiplied them according to the new operation, and obtained another element of $\mathbb{N} \times \mathbb{M}$; the operation is therefore closed.

(associativity) Let $t, u, v \in \mathbb{N} \times \mathbb{M}$. By definition, we can write $t = (a, x^\alpha)$, $u = \left(b, x^\beta\right)$, and $v = (c, x^\gamma)$ for appropriate $a, \alpha, b, \beta, c, \gamma \in \mathbb{N}$. Then

$$\begin{aligned} t\,(uv) &= (a, x^\alpha) \otimes \left[\left(b, x^\beta\right) \otimes (c, x^\gamma)\right] \\ &= (a, x^\alpha) \otimes \left(bc, x^\beta x^\gamma\right) \\ &= \left(a\,(bc), x^\alpha \left(x^\beta x^\gamma\right)\right). \end{aligned}$$

To show that this equals $(tu)\,v$, we have to rely on the associative properties of $\mathbb{N}$ and $\mathbb{M}$:

$$\begin{aligned} t\,(uv) &= \left((ab)\,c, \left(x^\alpha x^\beta\right) x^\gamma\right) \\ &= \left(ab, x^\alpha x^\beta\right) \otimes (c, x^\gamma) \\ &= \left[(a, x^\alpha) \otimes \left(b, x^\beta\right)\right] \otimes (c, x^\gamma) \\ &= (tu)\,v. \end{aligned}$$

We took three elements of $\mathbb{N} \times \mathbb{M}$, and showed that the operation was associative for them. Since the elements were arbitrary, the operation is associative.

(identity) We claim that the identity of $\mathbb{N} \times \mathbb{M}$ is $(1,1) = \left(1, x^0\right)$. To see why, let $t \in \mathbb{N} \times \mathbb{M}$. By definition, we can write $t = (a, x^\alpha)$ for appropriate $a, \alpha \in \mathbb{N}$. Then

$$\begin{aligned} (1,1) \otimes t &= (1,1) \otimes (a, x^\alpha) && \text{(subst.)} \\ &= (1 \times a, 1 \times x^\alpha) && \text{(def. of } \otimes) \\ &= (a, x^\alpha) = t \end{aligned}$$

and similarly $t \otimes (1,1) = t$. We took an arbitrary element of $\mathbb{N} \times \mathbb{M}$, and showed that $(1,1)$ acted as an identity under the operation $\otimes$ with that element. Since the element was arbitrary, $(1,1)$ must be *the* identity for $\mathbb{N} \times \mathbb{M}$.

Interestingly, if we had used $(\mathbb{N}, +)$ *instead* of $(\mathbb{N}, \times)$ in the previous example, we *still* would

have obtained a direct product! Indeed, the direct product of monoids is *always* a monoid!

Theorem 1.35. The direct product of monoids $M_1, \ldots, M_r$ is itself a monoid. Its identity element is $(e_1, e_2, \ldots, e_r)$, where each e_i denotes the identity of the corresponding monoid M_i.

Proof. You do it! See Exercise 1.38. □

We finally turn our attention the question of whether $\mathbb{M}_n$ and $\mathbb{M}^n$ are the same.

Admittedly, the two are not identical: $\mathbb{M}_n$ is the set of *products* of powers of *n distinct* variables, whereas $\mathbb{M}^n$ is a set of *lists* of powers of *one* variable. In addition, if the variables are *not* commutative (remember that this can occur), then $\mathbb{M}_n$ and $\mathbb{M}^n$ are not at all similar. Think about $(xy)^4 = xyxyxyxy$; if the variables are commutative, we can combine them into x^4y^4, which looks likes $(4,4)$. If the variables are not commutative, however, it is not at *all* clear how we could get $(xy)^4$ to correspond to an element of $\mathbb{N} \times \mathbb{N}$.

That leads to the following result.

Theorem 1.36. The variables of $\mathbb{M}_n$ are commutative if and only if $\mathbb{M}_n \cong \mathbb{M}^n$.

Proof. Assume the variables of $\mathbb{M}_n$ are commutative. Let $f : \mathbb{M}_n \longrightarrow \mathbb{M}^n$ by

$$f\left(x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}\right) = \left(x^{a_1}, x^{a_2}, \ldots, x^{a_n}\right).$$

The fact that we cannot combine a_i and a_j if $i \neq j$ shows that f is one-to-one, and any element $\left(x^{b_1}, \ldots, x^{b_n}\right)$ of $\mathbb{M}^n$ has a preimage $x_1^{b_1} \cdots x_n^{b_n}$ in $\mathbb{M}_n$; thus f is a bijection.

Is it also an isomorphism? To see that it is, let $t, u \in \mathbb{M}_n$. By definition, we can write $t = x_1^{a_1} \cdots x_n^{a_n}$ and $u = x_1^{b_1} \cdots x_n^{b_n}$ for appropriate $a_1, b_1 \ldots, a_n, b_n \in \mathbb{N}$. Then

$$\begin{aligned}
f(tu) &= f\left(\left(x_1^{a_1} \cdots x_n^{a_n}\right)\left(x_1^{b_1} \cdots x_n^{b_n}\right)\right) && \text{(substitution)} \\
&= f\left(x_1^{a_1+b_1} \cdots x_n^{a_n+b_n}\right) && \text{(commutative)} \\
&= \left(x^{a_1+b_1}, \ldots, x^{a_n+b_n}\right) && \text{(definition of } f\text{)} \\
&= \left(x^{a_1}, \ldots, x^{a_n}\right) \otimes \left(x^{b_1}, \ldots, x^{b_n}\right) && \text{(def. of } \otimes\text{)} \\
&= f(t) \otimes f(u). && \text{(definition of } f\text{)}
\end{aligned}$$

Hence f is an isomorphism, and we conclude that $\mathbb{M}_n \cong \mathbb{M}^n$.

Conversely, suppose $\mathbb{M}_n \cong \mathbb{M}^n$. By Exercise 1.30, $\mathbb{M}^n \cong \mathbb{M}_n$. By definition, there exists a bijection $f : \mathbb{M}^n \longrightarrow \mathbb{M}_n$ satisfying Definition 1.26. Let $t, u \in \mathbb{M}^n$; by definition, we can find $a_i, b_j \in \mathbb{N}$ such that $t = x_1^{a_1} \cdots x_n^{a_n}$ and $u = x_1^{b_1} \cdots x_n^{b_n}$. Since f preserves the operation, $f(tu) = f(t) \otimes f(u)$. Now, $f(t)$ and $f(u)$ are elements of $\mathbb{M}^n$, which is commutative by Exercise 1.39 (with the $S_i = \mathbb{M}$ here). Hence $f(t) \otimes f(u) = f(u) \otimes f(t)$, so that $f(tu) = f(u) \otimes f(t)$. Using the fact that f preserves the operation again, only in reverse, we see that $f(tu) = f(ut)$. Recall that f, as a bijection, is one-to-one! Thus $tu = ut$, and $\mathbb{M}^n$ is commutative. □

Notation 1.37. Although we used $\otimes$ in this section to denote the operation in a direct product, this is not standard; I was trying to emphasize that the product is different for the direct product than for the monoids that created it. In general, the product $x \otimes y$ is written simply as xy. Thus, the last line of the proof above would have $f(t) f(u)$ instead of $f(t) \otimes f(u)$.

Exercises.

Exercise 1.38. Prove Theorem 1.35. Use Example 1.34 as a guide.

Exercise 1.39. Suppose M_1, M_2, ..., and M_n are *commutative* monoids. Show that the direct product $M_1 \times M_2 \times \cdots \times M_n$ is also a commutative monoid.

Exercise 1.40. Show that $\mathbb{M}^n \cong \mathbb{N}^n$. What does this imply about $\mathbb{M}_n$ and $\mathbb{N}^n$?

Exercise 1.41. Recall the lattice L from Exercise 0.52. Exercise 1.17 shows that this is both a monoid under addition and a monoid under multiplication, as defined in that exercise. Is either monoid isomorphic to $\mathbb{N}^2$?

Exercise 1.42. Let $\mathbb{T}_S^n$ denote the set of terms in n variables whose coefficients are elements of the set S. For example, $2xy \in \mathbb{T}_{\mathbb{Z}}^2$ and $\pi x^3 \in \mathbb{T}_{\mathbb{R}}^1$.

(a) Show that if S is a monoid, then so is $\mathbb{T}_S^n$.

(b) Show that if S is a monoid, then $\mathbb{T}_S^n \cong S \times \mathbb{M}_n$.

Exercise 1.43. We define the **kernel** of a monoid homomorphism $\varphi : M \to N$ as

$$\ker \varphi = \{(x, y) \in M \times M : \varphi(x) = \varphi(y)\}.$$

Recall from this section that $M \times M$ is a monoid.

(a) Show that $\ker \varphi$ is a "submonoid" of $M \times M$; that is, it is a subset that is also a monoid.

(b) Fill in each blank of Figure 1.2 with the justification.

(c) Denote $K = \ker \varphi$, and define M/K in the following way.

> A **coset** xK is the set S of all $y \in M$ such that $(x, y) \in K$, and M/K is the set of all such cosets.

Show that

(i) every $x \in M$ appears in at least one coset;

(ii) M/K is a partition of M.

Suppose we try to define an operation on the cosets in a "natural" way:

$$(xK) \circ (yK) = (xy)K.$$

It can happen that two cosets X and Y can each have different representations: $X = xK = wK$, and $Y = yK = zK$. It often happens that $xy \neq wz$, which could open a can of worms:

$$XY = (xK)(yK) = (xy)K \neq (wz)K = (wK)(zK) = XY.$$

Obviously, we'd rather that not happen, so

(iii) Fill in each blank of Figure 1.3 with the justification.

Claim: $\ker\varphi$ is an equivalence relation on M. That is, if we define a relation $\sim$ on M by $x \sim y$ if and only if $(x,y) \in \ker\varphi$, then $\sim$ satisfies the reflective, symmetric, and transitive properties.

1. We prove the three properties in turn.
2. The reflexive property:
   (a) Let $m \in M$.
   (b) By _____, $\varphi(m) = \varphi(m)$.
   (c) By _____, $(m,m) \in \ker\varphi$.
   (d) Since _____, every element of M is related to itself by $\ker\varphi$.
3. The symmetric property:
   (a) Let $a, b \in M$. Assume a and b are related by $\ker\varphi$.
   (b) By _____, $\varphi(a) = \varphi(b)$.
   (c) By _____, $\varphi(b) = \varphi(a)$.
   (d) By _____, b and a are related by $\ker\varphi$.
   (e) Since _____, this holds for all pairs of elements of M.
4. The transitive property:
   (a) Let $a, b, c \in M$. Assume a and b are related by $\ker\varphi$, and b and c are related by $\ker\varphi$.
   (b) By _____, $\varphi(a) = \varphi(b)$ and $\varphi(b) = \varphi(c)$.
   (c) By _____, $\varphi(a) = \varphi(c)$.
   (d) By _____, a and c are related by $\ker\varphi$.
   (e) Since _____, this holds for any selection of three elements of M.
5. We have shown that a relation defined by $\ker\varphi$ satisfies the reflexive, symmetric, and transitive properties. Thus, $\ker\varphi$ is an equivalence relation on M.

Figure 1.2. Material for Exercise 1.43(b)

Once you've shown that the operation is well defined, show that
(iv) M/K is a monoid with this operation.
This means that we can use monoid morphisms to create new monoids.

1.4: Absorption and the Ascending Chain Condition

We conclude our study of monoids by introducing a new object, and a fundamental notion.

Absorption

Definition 1.44. Let M be a monoid, and $A \subseteq M$. If $ma \in A$ for every $m \in M$ and $a \in A$, then A **absorbs from** M. We also say that A is an **absorbing subset**, or that satisfies the **absorption property**.

Notice that if A absorbs from M, then A is closed under multiplication: if $x, y \in A$, then $A \subseteq M$ implies that $x \in M$, so by absorption, $xy \in A$, as well. Unfortunately, that doesn't make A a monoid, as 1_M might not be in A.

Let M and N be monoids, φ a homomorphism from M to N, and $K = \ker\varphi$.
Claim: The "natural" operation on cosets of K is well defined.
Proof:

1. Let $X, Y \in$_____. That is, X and Y are cosets of K.
2. By _____, there exist $x, y \in M$ such that $X = xK$ and $Y = yK$.
3. Assume there exist $w, z \in$_____ such that $X = wK$ and $Y = zK$. We must show that $(xy)K = (wz)K$.
4. Let $a \in (xy)K$.
5. By definition of coset, _____$\in K$.
6. By _____, $\varphi(xy) = \varphi(a)$.
7. By _____, $\varphi(x)\varphi(y) = \varphi(a)$.
8. We claim that $\varphi(x) = \varphi(w)$ and $\varphi(y) = \varphi(z)$.
   (a) To see why, recall that by _____, $xK = X = wK$ and $yK = Y = zK$.
   (b) By part _____ of this exercise, $(x, x) \in K$ and $(w, w) \in K$.
   (c) By _____, $x \in xK$ and $w \in wK$.
   (d) By _____, $w \in xK$.
   (e) By _____, $(x, w) \in \ker\varphi$.
   (f) By _____, $\varphi(x) = \varphi(w)$. A similar argument shows that $\varphi(y) = \varphi(z)$.
9. By _____, $\varphi(w)\varphi(z) = \varphi(a)$.
10. By _____, $\varphi(wz) = \varphi(a)$.
11. By definition of coset, _____$\in K$.
12. By _____, $a \in (wz)K$.
13. By _____, $(xy)K \subseteq (wz)K$. A similar argument shows that $(xy)K \supseteq (wz)K$.
14. By definition of equality of sets, _____.
15. We have see that the representations of _____ and _____ do not matter; the product is the same regardless. Coset multiplication is well defined.

Figure 1.3. Material for Exercise 1.43

Example 1.45. Write $2\mathbb{Z}$ for the set of even integers. By definition, $2\mathbb{Z} \subsetneq \mathbb{Z}$. Notice that $2\mathbb{Z}$ is *not* a monoid, since $1 \notin 2\mathbb{Z}$. On the other hand, any $a \in 2\mathbb{Z}$ has the form $a = 2z$ for some $z \in \mathbb{Z}$. Thus, for any $m \in \mathbb{Z}$, we have

$$ma = m(2z) = 2(mz) \in 2\mathbb{Z}.$$

Since a and m were arbitrary, $2\mathbb{Z}$ absorbs from $\mathbb{Z}$.

The set of integer multiples of an integer is important enough that it inspires notation.

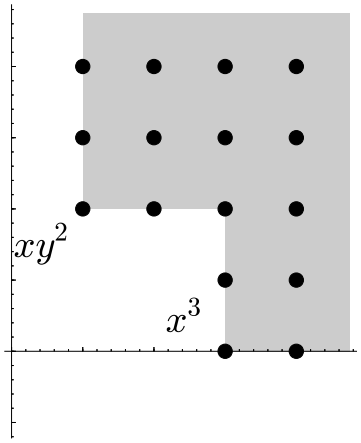
Notation 1.46. We write $d\mathbb{Z}$ for the set of integer multiples of d.

So $2\mathbb{Z} = \{\ldots, -2, 0, 2, 4, \ldots\}$ is the set of integer multiples of 2; $5\mathbb{Z}$ is the set of integer multiples of 5; and so forth. You will show in Exercise 1.56 that $d\mathbb{Z}$ absorbs multiplication from $\mathbb{Z}$, but *not* addition.

The monomials provide another important example of absorption.

Example 1.47. Let A be an absorbing subset of $\mathbb{M}_2$. Suppose that $xy^2, x^3 \in A$, but none of their factors is in A. Since A absorbs from $\mathbb{M}_2$, all the monomial multiples of xy^2 and x^3 are also in A.

We can illustrate this with a **monomial diagram**:



Every dot represents a monomial in A; the dot at $(1,2)$ represents the monomial xy^2, and the dots above it represent xy^3, xy^4, ... Notice that multiples of xy^2 and x^3 lie *above and to the right* of these monomials.

The diagram suggests that we can identify special elements of subsets that absorb from the monomials.

Definition 1.48. Suppose A is an absorbing subset of $\mathbb{M}_n$, and $t \in A$. If no other $u \in A$ divides t, then we call t a **generator** of A.

In the diagram above, xy^2 and x^3 are the generators of an ideal corresponding to the monomials covered by the shaded region, extending indefinitely upwards and rightwards. The name "generator" is apt, because every monomial multiple of these two xy^2 and x^3 is also in A, but nothing "smaller" is in A, in the sense of divisibility.

This leads us to a remarkable result.

Dickson's Lemma and the Ascending Chain Condition

Theorem 1.49 (Dickson's Lemma)**.** Every absorbing subset of $\mathbb{M}_n$ has a finite number of generators.

(Actually, Dickson proved a similar result for a similar set, but is more or less the same.) The proof is a little complicated, so we'll illustrate it using some monomial diagrams. In Figure 1.4(A), we see an absorbing subset A. (The same as you saw before.) Essentially, the argument *projects* A down one dimension, as in Figure 1.4(B). In this smaller dimension, an argument by induction allows us to choose a finite number of generators, which correspond to elements of A, illustrated in Figure 1.4(C). These corresponding elements of A are always generators of A, but they might not be *all* the generators of A, shown in Figure 1.4(C) by the red circle. In that case, we take the remaining generators of A, use them to construct a new absorbing subset, and project again to obtain new generators, as in Figure 1.4(D). The thing to notice is that, in Figures 1.4(C) and 1.4(D), the y-values of the new generators decrease with each projection. This cannot continue indefinitely, since $\mathbb{N}$ is well-ordered, and we are done.

Proof. Let A be an absorbing subset of $\mathbb{M}_n$. We proceed by induction on the dimension, n.

For the *inductive base,* assume $n = 1$. Let S be the set of exponents of monomials in A. Since $S \subseteq \mathbb{N}$, it has a minimal element; call it a. By definition of S, $x^a \in A$. We claim that x^a is, in fact,

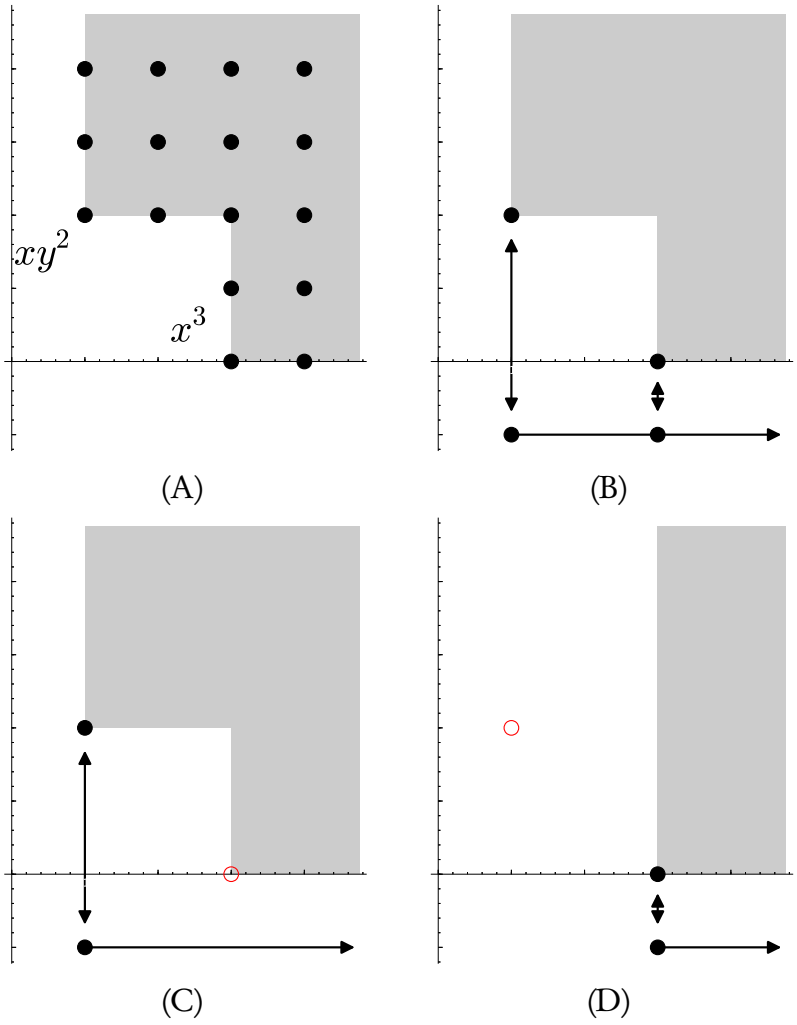


Figure 1.4. Illustration of the proof of Dickson's Lemma.

the one generator of A. To see why, let $u \in A$. Suppose that $u \mid x^a$; by definition of monomial divisibility, $u = x^b$ and $b \leq a$. Since $u \in A$, it follows that $b \in S$. Since a is the *minimal* element of S, $a \leq b$. We already knew that $b \leq a$, so it must be that $a = b$. The claim is proved: no other element of A divides x^a. Thus, x^a is a generator, and since $n = 1$, the generator is unique.

For the *inductive hypothesis*, assume that any absorbing subset of $\mathbb{M}_{n-1}$ has a finite number of generators.

For the *inductive step*, we use A to construct a sequence of absorbing subsets of $\mathbb{M}_{n-1}$ in the following way.

- Let B_1 be the set of all monomials in $\mathbb{M}_{n-1}$ such that $t \in B_1$ implies that $t x_n^a \in A$ for some $a \in \mathbb{N}$. We call this a **projection** of A onto $\mathbb{M}_{n-1}$.
  We claim that B_1 absorbs from $\mathbb{M}_{n-1}$. To see why, let $t \in B_1$, and let $u \in \mathbb{M}_{n-1}$ be any monomial multiple of t. By definition, there exists $a \in \mathbb{N}$ such that $t x_n^a \in A$. Since A absorbs from $\mathbb{M}_n$, and $u \in \mathbb{M}_{n-1} \subsetneq \mathbb{M}_n$, absorption implies that $u\left(t x_n^a\right) \in A$. The associative property tells us that $(ut)\, x_n^a \in A$, and the definition of B_1 tells us that $ut \in B_1$. Since t_1 is an arbitrary element of B_1, u is an arbitrary multiple of t, and we found that $u \in B_1$, we can conclude that B_1 absorbs from $\mathbb{M}_{n-1}$.

This result is important! By the inductive hypothesis, B_1 has a finite number of generators; call them $\{t_1, \ldots, t_m\}$. Each of these generators corresponds to an element of A. Let $T_1 = \left\{t_1 x_n^{a_1}, \ldots, t_m x_n^{a_m}\right\} \subsetneq A$ such that a_1 is the *smallest* element of $\mathbb{N}$ such that $t_1 x_n^{a_1} \in A$, …, a_m is the *smallest* element of $\mathbb{N}$ such that $t_m x_n^{a_m} \in A$. (Such a smallest element must exist on account of the well-ordering of $\mathbb{N}$.)

We now claim that T_1 is a list of some of the generators of A. To see this, assume by way of contradiction that we can find some $u \in T_1$ that is not a generator of A. The definition of a generator means that there exists some other $v \in A$ that divides u. We can write $u = t x_n^a$ and $v = t' x_n^b$ for some $a, b \in \mathbb{N}$; then $t, t' \in B_1$. Here, things fall apart! After all, t' also divides t, contradicting the assumption that t' is a generator of B_1.

- If T_1 is a complete list of the generators of A, then we are done. Otherwise, let $A^{(1)}$ be the absorbing subset whose elements are multiples of the generators of A that are *not* in T_1. Let B_2 be the projection of $A^{(1)}$ onto $\mathbb{M}_{n-1}$. As before, B_2 absorbs from $\mathbb{M}_{n-1}$, and the inductive hypothesis implies that it has a finite number of generators, which correspond to a set T_2 of generators of $A^{(1)}$.
- As long as T_i is not a complete list of the generators of A, we continue building
  - an absorbing subset $A^{(i)}$ whose elements are multiples of the generators of A that are *not* in T_i;
  - an absorbing subset B_{i+1} whose elements are the projections of $A^{(i)}$ onto $\mathbb{M}_{n-1}$, and
  - sets T_{i+1} of generators of A that correspond to generators of B_{i+1}.

Can this process continue indefinitely? No, it cannot. First, if $t \in T_{i+1}$, then write it as $t = t' x_n^a$. On the one hand,

$$t \in A^{(i)} \subsetneq A^{(i-1)} \subsetneq \cdots A^{(1)} \subsetneq A,$$

so t' was an element of every B_j such that $j \leq i$. That means that for each j, t' was divisible by at least one generator u'_j of B_j. However, t was *not* in the absorbing subsets generated by $T_1, \ldots, T_i$. So the $u_j \in T_j$ corresponding to u'_j does *not* divide t. Write $t = x_1^{a_1} \cdots x_n^{a_1}$ and $u = x_1^{b_1} \cdots x_n^{b_n}$. Since $u' \mid t'$, $b_k \leq a_k$ for each $k = 1, \ldots, n-1$. Since $u \nmid t$, $b_n > a_n$.

In other words, the minimal degree of x_n is decreasing in T_i as i increases. This gives us a strictly decreasing sequence of natural numbers. By the well-ordering property, such a sequence cannot continue indefinitely. Thus, we cannot create sets T_i containing new generators of A indefinitely; there are only finitely many such sets. In other words, A has a finite number of generators. □

This fact leads us to an important concept, that we will exploit greatly, starting in Chapter 8.

Definition 1.50. Let M be a monoid. Suppose that, for any ideals A_1, A_2, … of M, we can guarantee that if $A_1 \subseteq A_2 \subseteq \cdots$, then there is some $n \in \mathbb{N}^+$ such that $A_n = A_{n+1} = \cdots$. In this case, we say that M satisfies the **ascending chain condition**,, or that M is **Noetherian**.

A look back at the Hilbert-Dickson game

We conclude with two results that will, I hope, delight you. There is a technique for counting the number of elements *not* shaded in the monomial diagram.

Definition 1.51. Let A be an absorbing subset of $\mathbb{M}_n$. The **Hilbert Function** $H_A(d)$ counts the number of monomials of total degree d and *not* in A. The **Affine Hilbert Function** $H_A^{\text{aff}}(d)$ is the sum of the Hilbert Function for degree no more than d; that is, $H_A^{\text{aff}}(d) = \sum_{i=0}^{d} H_A(d)$.

Example 1.52. In the diagram of Example 1.47, $H(0) = 1$, $H(1) = 2$, $H(2) = 3$, $H(3) = 2$, and $H(d) = 1$ for all $d \geq 4$. On the other hand, $H^{\text{aff}}(4) = 9$.

The following result is immediate.

Theorem 1.53. Suppose that A is the absorbing subset generated by the moves chosen in a Hilbert-Dickson game, and let $d \in \mathbb{N}$. The number of moves (a,b) possible in a Hilbert-Dickson game with $a+b \leq d$ is $H_A^{\text{aff}}(d)$.

Corollary 1.54. Every Hilbert-Dickson game must end in a finite number of moves.

Proof. Every ith move in a Hilbert-Dickson game corresponds to the creation of a new absorbing subset A_i of $\mathbb{M}_2$. Let A be the union of these A_i; you will show in Exercise 1.57 that A also absorbs from $\mathbb{M}_2$. By Dickson's Lemma, A has finitely many generators; call them $t_1, \ldots, t_m$. Each t_j appears in A, and the definition of union means that each t_j must appear in some A_{i_j}. Let k be the largest such i_j; that is, $k = \max\{i_1, \ldots, i_m\}$. Practically speaking, "largest" means "last chosen", so each t_i has been chosen at this point. Another way of saying this in symbols is that $t_1, \ldots, t_m \in \bigcup_{i=1}^{k} A_i$. All the generators of A are in this union, so no element of A can be absent! So $A = \bigcup_{i=1}^{k} A_i$; in other words, the ideal is generated after finitely many moves. □

Dickson's Lemma is a perfect illustration of the Ascending Chain Condition. It also illustrates a relationship between the Ascending Chain Condition and the well-ordering of the integers: we used the well-ordering of the integers repeatedly to prove that $\mathbb{M}_n$ is Noetherian. You will see this relationship again in the future.

Exercises.

Exercise 1.55. Is $2\mathbb{Z}$ an absorbing subset of $\mathbb{Z}$ under addition? Why or why not?

Exercise 1.56. Let $d \in \mathbb{Z}$ and $A = d\mathbb{Z}$. Show that A is an absorbing subset of $\mathbb{Z}$.

Exercise 1.57. Fill in each blank of Figure 1.5 with its justification.

Exercise 1.58. Let L be the lattice defined in Exercise 0.52. Exercise 1.17 shows that L is a monoid under its strange multiplication. Let $P = (3,1)$ and A be the absorbing subset generated by P. Sketch L and P, distinguishing the elements of P from those of L using different colors, or an X, or some similar distinguishing mark.

Suppose $A_1, A_2, \ldots$ absorb from a monoid M, and $A_i \subseteq A_{i+1}$ for each $i \in \mathbb{N}^+$.
Claim: Show that $A = \bigcup_{i=1}^{\infty} A_i$ also absorbs from M.

1. Let $m \in M$ and $a \in A$.
2. By _____, there exists $i \in \mathbb{N}^+$ such that $a \in A_i$.
3. By _____, $ma \in A_i$.
4. By _____, $A_i \subseteq A$.
5. By _____, $ma \in A$.
6. Since _____, this is true for all $m \in M$ and all $a \in A$.
7. By _____, A also absorbs from M.

Figure 1.5. Material for Exercise 1.57

Chapter 2:
Groups

In Chapter 1, we described monoids. In this chapter, we study a *group,* which is a special kind of monoid. What motivates us is the observation that the set of integers is a monoid, but also more than a monoid.

How? The natural numbers are closed under addition: for any two $a, b \in \mathbb{N}$, we know that $a + b \in \mathbb{N}$ also. This is also true about the integers: for any $a, b \in \mathbb{Z}$, $a + b \in \mathbb{Z}$. However, the integers are also closed under subtraction, *while the natural numbers are not!* Even though $3, 5 \in \mathbb{N}$, $3 - 5 = -2 \notin \mathbb{N}$!

That is, groups are special in that every element in the group has an *inverse element.* It is not entirely wrong to say that groups actually have two operations. You will see in a few moments that the integers are a group under addition: not only does it satisfy the properties of a monoid, but each of its elements also has an additive inverse in $\mathbb{Z}$. Stated a different way, $\mathbb{Z}$ has a second operation, *subtraction*. However, the conditions on this second operation are so restrictive (it has to "undo" the first operation) that most mathematicians won't consider groups to have two operations; they prefer to say that a property of the group operation is that every element has an inverse element.

This property is essential to a large number of mathematical phenomena. We describe a special class of groups called the cyclic groups (Section 2.3) and then look at two groups related to important mathematical problems. The first, D_3, describes symmetries of a triangle using groups (Section 2.2). The second, Ω_n, consists of the roots of unity (Section 2.4).

2.1: Groups

This first section looks only at some very basic properties of groups, and some very basic examples.

Precise definition, first examples

Definition 2.1. Let G be a set, and $\circ$ a binary operation on G. We say that the pair $(G, \circ)$ is a **group** if it satisfies the following properties.

(closure) for any $x, y \in G$, we have $x \circ y \in G$;

(associative) for any $x, y, z \in G$, we have $(x \circ y) \circ z = x \circ (y \circ z)$;

(identity) there exists an **identity element** $e \in G$; that is, for any $x \in G$, we have $x \circ e = e \circ x = x$; and

(inverses) each element of the group has an **inverse**; that is, for any $x \in G$ we can find $y \in G$ such that $x \circ y = y \circ x = e$.

We may also say that G is a **group under** $\circ$. We say that $(G, \circ)$ is an **abelian group** if it also satisfies

(commutative) the operation is commutative; that is, $xy = yx$ for all $x, y \in G$.

Notation 2.2. If the operation is addition, we may refer to the group as an **additive group** or a **group under addition**. We also write $-x$ instead of x^{-1}, and $x+(-y)$ or even $x-y$ instead of $x+y^{-1}$, keeping with custom. Additive groups are normally abelian.

If the operation is multiplication, we may refer to the group as a **multiplicative group** or a **group under multiplication**. The operation is usually understood from context, so we typically write G rather than $(G,+)$ or $(G,\times)$ or $(G,\circ)$. We will write $(G,+)$ when we want to emphasize that the operation is addition.

Example 2.3. Certainly $\mathbb{Z}$ is an additive group; in fact, it is abelian. Why?

- We know it is a monoid under addition.
- Every integer has an additive inverse *in* $\mathbb{Z}$.
- Addition of integers is commutative.

However, while $\mathbb{N}$ is a monoid under addition, it is not a group. Why not? The problem is with inverses. We know that every natural number has an additive inverse; after all, $2+(-2)=0$. Nevertheless, the inverse property is *not* satisfied because $-2\notin\mathbb{N}$! It's not enough to have an inverse in *some* set; *the inverse be in the same set!* For this reason, $\mathbb{N}$ is not a group.

Example 2.4. In addition to $\mathbb{Z}$, the following sets are groups under addition.

- the set $\mathbb{Q}$ of **rational numbers**;
- the set $\mathbb{R}$ of **real numbers**; and
- if $S=\mathbb{Z},\mathbb{Q}$, or $\mathbb{R}$, the set $S^{m\times n}$ of $m\times n$ matrices whose elements are in S. (It's important here that the operation is *addition*.)

However, none of them is a group under multiplication. On the other hand, the set of invertible $n\times n$ matrices with elements in $\mathbb{Q}$ or $\mathbb{R}$ is a multiplicative group. We leave the proof to the exercises, but this fact builds on properties you learned in linear algebra, such as those described in Section 0.3.

Definition 2.5. We call the set of invertible $n\times n$ matrices with elements in $\mathbb{R}$ the **general linear group of degree** n, and write $\mathrm{GL}_n(\mathbb{R})$ for this set.

Order of a group, Cayley tables

Mathematicians of the 20th century invested substantial effort in an attempt to classify all *finite, simple groups*. (You will learn later what makes a group "simple".) Replicating that achievement is far, far beyond the scope of these notes, but we can take a few steps in this area.

Definition 2.6. Let S be any set. We write $|S|$ to indicate the number of elements in S, and say that $|S|$ is the **size** or **cardinality** of S. If there is an infinite number of elements in S, then we write $|S|=\infty$. We also write $|S|<\infty$ to indicate that $|S|$ is finite, if we don't want to state a precise number.

For any group G, the **order of** G is the size of G. A group has finite order if $|G|<\infty$ and infinite order if $|G|=\infty$.

Here are three examples of finite groups; in fact, they are all of order 2.

Example 2.7. The sets

$$\{1,-1\}, \quad \left\{\begin{pmatrix}1&0\\0&1\end{pmatrix},\begin{pmatrix}-1&0\\0&1\end{pmatrix}\right\},$$
$$\text{and} \quad \left\{\begin{pmatrix}1&0\\0&1\end{pmatrix},\begin{pmatrix}1&0\\0&-1\end{pmatrix}\right\}$$

are all groups under multiplication:

- In the first group, the identity is 1, and -1 is its own inverse; closure is obvious, and you know from arithmetic that the associative property holds.
- In the second and third groups, the identity is the identity matrix; each matrix is its own inverse; closure is easy to verify, and you know from linear algebra that the associative property holds.

I will now make an extraordinary claim:

Claim 1. For all intents and purposes, there is only one group of order two.

This claim may seem preposterous on its face; after all, the example above has three completely different groups of order two. In fact, the claim is quite vague, because we're using vague language. After all, what is meant by the phrase, "for all intents and purposes"? Basically, we meant that:

- group theory cannot distinguish between the groups *as groups;* or,
- their multiplication table (or addition table, or whatever-operation table) has the same structure.

If you read the second characterization and think, "he means they're isomorphic!", then pat yourself on the back. Unfortunately, we won't look at this notion seriously until Chapter 4, but Chapter 1 gave you a rough idea of what that meant: the groups are identical *as groups*.

We will prove the claim above in a "brute force" manner, by looking at the table generated by the operation of the group. Now, "the table generated by the operation of the group" is an ungainly phrase, and quite a mouthful. Since the name of the table depends on the operation (multiplication table, addition table, etc.), we have a convenient phrase that describes all of them.

Definition 2.8. The table listing all results of the operation of a monoid or group is its **Cayley table**.

Since groups are monoids, we can call their table a Cayley table, too.

Back to our claim. We want to build a Cayley table for a "generic" group of order two. We will show that there is only one possible way to construct such a table. As a consequence, regardless of the set and its operation, every group of order 2 behaves exactly the same way. ***It does not matter one whit*** what the elements of G are, or the fancy name we use for the operation, or the convoluted procedure we use to simplify computations in the group. If there are only two elements, and it's a group, then ***it always works the same***. Why?

Example 2.9. Let G be an arbitrary group of order two. By definition, it has an identity, so write $G = \{e, a\}$ where e represents the known identity, and a the other element.

We did *not* say that e represents the *only* identity. For all we know, a might also be an identity; is that possible? In fact, it is not possible; why? Remember that a group is a monoid. We showed

in Proposition 2.12 that the identity of a monoid is unique; thus, the identity of a group is unique; thus, there can be only one identity, e.

Now we build the addition table. We *have* to assign $a \circ a = e$. *Why?*

- To satisfy the identity property, we must have $e \circ e = e$, $e \circ a = a$, and $a \circ e = a$.
- To satisfy the inverse property, a must have an additive inverse. We know the inverse can't be e, since $a \circ e = a$; so the only inverse possible is a itself! That is, $a^{-1} = a$. (Read that as, "the inverse of a is a.") So $a \circ a^{-1} = a \circ a = e$.

So the Cayley table of our group looks like:

$\circ$	e	a
e	e	a
a	a	e

The only assumption we made about G is that it was a group of order two. That means this table applies to *any* group of order two, and we have determined the Cayley table of *all* groups of order two!

In Definition 2.1 and Example 2.9, the symbol $\circ$ is a placeholder for any operation. We assumed nothing about its actual behavior, so it can represent addition, multiplication, or other operations that we have not yet considered. Behold the power of abstraction!

Other elementary properties of groups

Notation 2.10. We adopt the following convention:

- If we know only that G is a group under some operation, we write $\circ$ for the operation and proceed as if the group were multiplicative, so that xy is shorthand for $x \circ y$.
- If we know that G is a group and a symbol is provided for its operation, we *usually* use that symbol for the group, *but not always.* Sometimes we treat the group as if it were multiplicative, writing xy instead of the symbol provided.
- We reserve the symbol $+$ exclusively for additive groups.

The following fact looks obvious—but remember, we're talking about elements of *any* group, not merely the sets you have worked with in the past.

Proposition 2.11. Let G be a group and $x \in G$. Then $\left(x^{-1}\right)^{-1} = x$. If G is additive, we write instead that $-(-x) = x$.

Proposition 2.11 says that the inverse of the inverse of x is x itself; that is, if y is the inverse of x, then x is the inverse of y.

Proof. You prove it! See Exercise 2.15. □

Proposition 2.12. The identity of a group is both two-sided and unique; that is, every group has exactly one identity. Also, the inverse of an element is both two-sided and unique; that is, every element has exactly one inverse element.

Proof. Let G be a group. We already pointed out that, since G is a monoid, and the identity of a monoid is both two-sided and unique, the identity of G is unique.

We turn to the question of the inverse. First we show that any inverse is two-sided. Let $x \in G$. Let w be a left inverse of x, and y a right inverse of x. Since y is a right inverse,

$$xy = e.$$

By the identity property, we know that $ex = x$. So, substitution and the associative property give us

$$\begin{aligned}(xy)\,x &= ex\\ x\,(yx) &= x.\end{aligned}$$

Since w is a left inverse, $wx = e$, so substitution, the associative property, the identity property, and the inverse property give

$$\begin{aligned}w\,(x\,(yx)) &= wx\\ (wx)\,(yx) &= wx\\ e\,(yx) &= e\\ yx &= e.\end{aligned}$$

Hence y is a left inverse of x. We already knew that it was a right inverse of x, so right inverses are in fact two-sided inverses. A similar argument shows that left inverses are two-sided inverses.

Now we show that inverses are unique. Suppose that $y, z \in G$ are both inverses of x. Since y is an inverse of x,

$$xy = e.$$

Since z is an inverse of x,

$$xz = e.$$

By substitution,

$$xy = xz.$$

Multiply both sides of this equation on the left by y to obtain

$$y\,(xy) = y\,(xz)\,.$$

By the associative property,

$$(yx)\,y = (yx)\,z,$$

and by the inverse property,

$$ey = ez.$$

Since e is the identity of G,

$$y = z.$$

We chose two arbitrary inverses of x, and showed that they were the same element. Hence the inverse of x is unique. □

In Example 2.9, the structure of a group compelled certain assignments for the operation. We can infer a similar conclusion for any group of finite order.

Theorem 2.13. Let G be a group of finite order, and let $a, b \in G$. Then a appears exactly once in any row or column of the Cayley table that is headed by b.

It might surprise you that this is *not* necessarily true for a monoid; see Exercise 2.23.

Proof. First we show that a cannot appear more than once in any row or column headed by b. In fact, we show it only for a row; the proof for a column is similar.

The element a appears in a row of the Cayley table headed by b any time there exists $c \in G$ such that $bc = a$. Let $c, d \in G$ such that $bc = a$ and $bd = a$. (We have *not* assumed that $c \neq d$.) Since $a = a$, substitution implies that $bc = bd$. Thus

$$\begin{aligned} c &\underset{\text{id.}}{=} ec \underset{\text{inv.}}{=} \left(b^{-1}b\right)c \underset{\text{ass.}}{=} b^{-1}(bc) \\ &\underset{\text{subs.}}{=} b^{-1}(bd) \underset{\text{ass.}}{=} \left(b^{-1}b\right)d \underset{\text{inv.}}{=} ed \underset{\text{id.}}{=} d. \end{aligned}$$

By the transitive property of equality, $c = d$. This shows that if a appears in one column of the row headed by b, then that column is unique; a does not appear in a different column.

We still have to show that a appears in at least one row of the addition table headed by b. This follows from the fact that each row of the Cayley table contains $|G|$ elements. What applies to a above applies to the other elements, so each element of G can appear at most once. Thus, if we do not use a, then only $n-1$ pairs are defined, which contradicts either the definition of an operation (bx must be defined for all $x \in G$) or closure (that $bx \in G$ for all $x \in G$). Hence a must appear at least once. □

Definition 2.14. Let $G_1, \ldots, G_n$ be groups. The **direct product** of $G_1, \ldots, G_n$ is the cartesian product $G_1 \times \cdots \times G_n$ together with the operation $\otimes$ such that for any $(g_1, \ldots, g_n)$ and $(h_1, \ldots, h_n)$ in $G_1 \times \cdots \times G_n$,

$$(g_1, \ldots, g_n) \otimes (h_1, \ldots, h_n) = (g_1 h_1, \ldots, g_n h_n),$$

where each product $g_i h_i$ is performed according to the operation of G_i. In other words, the direct product of *groups* generalizes the direct product of *monoids*.

You will show in the exercises that the direct product of groups is also a group.

Exercises.

Exercise 2.15.

(a) Fill in each blank of Figure 2.1 with the appropriate justification or statement.

(b) Why should someone think to look at the product of x and x^{-1} in order to show that $\left(x^{-1}\right)^{-1} = x$?

Exercise 2.16. Explain why $(\mathbb{M}, \times)$ is not a group.

Exercise 2.17. Is $\left(\mathbb{N}^+, \text{lcm}\right)$ a group? (See Exercise 1.22.)

Let G be a group, and $x \in G$.
Claim: $\left(x^{-1}\right)^{-1} = x$; or, if the operation is addition, $-(-x) = x$.
Proof:
1. By _____, $x \cdot x^{-1} = e$ and $x^{-1} \cdot x = e$.
2. By _____, $\left(x^{-1}\right)^{-1} = x$.
3. Negative are merely how we express opposites when the operation is addition, so $-(-x) = x$.

Figure 2.1. Material for Exercise 2.15

Exercise 2.18. Let G be a group, and $x, y, z \in G$. Show that if $xz = yz$, then $x = y$; or if the operation is addition, that if $x + z = y + z$, then $x = y$.

Exercise 2.19. Show in detail that $\mathbb{R}^{2\times 2}$ is an additive group.

Exercise 2.20. Recall the Boolean-or monoid $(B, \vee)$ from Exercise 1.13. Is it a group? If so, is it abelian? Explain how it justifies each property. If not, explain why not.

Exercise 2.21. Recall the Boolean-xor monoid $(B, \oplus)$ from Exercise 1.14. Is it a group? If so, is it abelian? Explain how it justifies each property. If not, explain why not.

Exercise 2.22. In Section 1.1, we showed that F_S, the set of all functions, is a monoid for any S.
(a) Show that $F_{\mathbb{R}}$, the set of all functions on the real numbers $\mathbb{R}$, is *not* a group.
(b) Describe a subset of $F_{\mathbb{R}}$ that *is* a group. Another way of looking at this question is: what restriction would you have to impose on any function $f \in F_S$ to fix the problem you found in part (a)?

Exercise 2.23. Indicate a monoid you have studied that does not satisfy Theorem 2.13. That is, find a monoid M such that (i) M is finite, and (ii) there exist $a, b \in M$ such that in the the Cayley table, a appears at least twice in a row or column headed by b.

Exercise 2.24. Show that the Cartesian product

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

is a group under the direct product's notion of addition; that is,

$$x + y = (a + c, b + d).$$

Exercise 2.25. Let $(G, \circ)$ and $(H, *)$ be groups, and define

$$G \times H = \{(a, b) : a \in G,\ b \in H\}.$$

Define an operation $\dagger$ on $G \times H$ in the following way. For any $x, y \in G \times H$, write $x = (a, b)$ and $y = (c, d)$; we say that

$$x \dagger y = (a \circ c, b * d).$$

(a) Show that $(G \times H, \dagger)$ is a group.

(b) Show that if G and H are both abelian, then so is $G \times H$.

Exercise 2.26. Let $n \in \mathbb{N}^+$. Let $G_1, G_2, \ldots, G_n$ be groups, and consider

$$\begin{aligned}\prod_{i=1}^{n} G_i &= G_1 \times G_2 \times \cdots \times G_n \\ &= \{(a_1, a_2, \ldots, a_n) : a_i \in G_i \ \forall i = 1, 2, \ldots, n\}\end{aligned}$$

with the operation $\dagger$ where if $x = (a_1, a_2, \ldots, a_n)$ and $y = (b_1, b_2, \ldots, b_n)$, then

$$x \dagger y = (a_1 b_1, a_2 b_2, \ldots, a_n b_n),$$

where each product $a_i b_i$ is performed according to the operation of the group G_i. Show that $\prod_{i=1}^{n} G_i$ is a group, and notice that this shows that the direct product of groups is a group, as claimed above. (We used $\otimes$ instead of $\dagger$ there, though.)

Exercise 2.27. Let $m \in \mathbb{N}^+$.

(a) Show in detail that $\mathbb{R}^{m \times m}$ is a group under addition.

(b) Show by counterexample that $\mathbb{R}^{m \times m}$ is *not* a group under multiplication.

Exercise 2.28. Let $m \in \mathbb{N}^+$. Explain why $\mathrm{GL}_m(\mathbb{R})$ satisfies the identity and inverse properties of a group.

Exercise 2.29. Let $\mathbb{R}^+ = \{x \in \mathbb{R} : x > 0\}$, and $\times$ the ordinary multiplication of real numbers. Show that $\left(\mathbb{R}^+, \times\right)$ is a group.

Exercise 2.30. Define $\mathbb{Q}^*$ to be the set of non-zero rational numbers; that is,

$$\mathbb{Q}^* = \left\{ \frac{a}{b} : a, b \in \mathbb{Z} \text{ where } a \neq 0 \text{ and } b \neq 0 \right\}.$$

Show that $\mathbb{Q}^*$ is a multiplicative group.

Exercise 2.31. Show that every group of order 3 has the same structure.

Exercise 2.32. *Not* every group of order 4 has the same structure, because there are two Cayley tables with different structures. One of these groups is the **Klein four-group**, where each element is its own inverse; the other is called a **cyclic group** of order 4, where not every element is its own inverse. Determine the Cayley tables for each group.

Exercise 2.33. Let G be a group, and $x, y \in G$. Show that $xy^{-1} \in G$.

Exercise 2.34.

(a) Let $m \in \mathbb{N}^+$ and $G = \mathrm{GL}_m(\mathbb{R})$. Show that there exist $a, b \in G$ such that $(ab)^{-1} \neq a^{-1}b^{-1}$.

(b) Suppose that H is an arbitrary group.

(i) Explain why we cannot assume that for every $a, b \in H$, $(ab)^{-1} = a^{-1}b^{-1}$.

(ii) Fill in the blanks of Figure 2.2 with the appropriate justification or statement.

Claim: Any two elements a, b of any group G satisfy $(ab)^{-1} = b^{-1}a^{-1}$.
Proof:

1. Let _____.
2. By the _____, _____, and _____ properties of groups,

$$(ab)\, b^{-1}a^{-1} = a\left(b \cdot b^{-1}\right) a^{-1} = aea^{-1} = aa^{-1} = e.$$

3. We chose _____ arbitrarily, so this holds for all elements of all groups, as claimed.

Figure 2.2. Material for Exercise 2.34

Exercise 2.35. Let $\circ$ denote the ordinary composition of functions, and consider the following functions that map any point $P = (x, y) \in \mathbb{R}^2$ to another point in $\mathbb{R}^2$:

$$\begin{aligned} I(P) &= P, \\ F(P) &= (y, x), \\ X(P) &= (-x, y), \\ Y(P) &= (x, -y). \end{aligned}$$

(a) Let $P = (2,3)$. Label the points P, $I(P)$, $F(P)$, $X(P)$, $Y(P)$, $(F \circ X)(P)$, $(X \circ Y)(P)$, and $(F \circ F)(P)$ on an x-y axis. (Some of these may result in the same point; if so, label the point twice.)
(b) Show that $F \circ F = X \circ X = Y \circ Y = I$.
(c) Show that $G = \{I, F, X, Y\}$ is *not* a group.
(d) Find the smallest group $\overline{G}$ such that $G \subset \overline{G}$. While you're at it, construct the Cayley table for $\overline{G}$.
(e) Is $\overline{G}$ abelian?

Definition 2.36. Let G be any group.

1. For all $x, y \in G$, define the **commutator of x and y** to be $x^{-1}y^{-1}xy$. We write $[x, y]$ for the commutator of x and y.
2. For all $z, g \in G$, define the **conjugation of g by z** to be zgz^{-1}. We write g^z for the conjugation of g by z.

Exercise 2.37. (a) Explain why $[x, y] = e$ iff x and y commute.
(b) Show that $[x, y]^{-1} = [y, x]$; that is, the inverse of $[x, y]$ is $[y, x]$.
(c) Show that $(g^z)^{-1} = \left(g^{-1}\right)^z$; that is, the inverse of conjugation of g by z is the conjugation of the inverse of g by z.
(d) Fill in each blank of Figure 2.3 with the appropriate justification or statement.

2.2: The symmetries of a triangle

In this section, we show that the symmetries of an equilateral triangle form a group. We call this group D_3. This group *is not abelian.* You already know that groups of order 2, 3, and 4 are

Claim: $[x,y]^z = [x^z, y^z]$ for all $x, y, z \in G$.
Proof:

1. Let _____.
2. By _____, $[x^z, y^z] = \left[zxz^{-1}, zyz^{-1}\right]$.
3. By _____, $\left[zxz^{-1}, zyz^{-1}\right] = \left(zxz^{-1}\right)^{-1}\left(zyz^{-1}\right)^{-1}\left(zxz^{-1}\right)\left(zyz^{-1}\right)$.
4. By Exercise _____,
$$\left(zxz^{-1}\right)^{-1}\left(zyz^{-1}\right)^{-1}\left(zxz^{-1}\right)\left(zyz^{-1}\right) = \\ = \left(zx^{-1}z^{-1}\right)\left(zy^{-1}z^{-1}\right)\left(zxz^{-1}\right)\left(zyz^{-1}\right).$$
5. By _____,
$$\left(zx^{-1}z^{-1}\right)\left(zy^{-1}z^{-1}\right)\left(zxz^{-1}\right)\left(zyz^{-1}\right) = \\ \left(zx^{-1}\right)\left(z^{-1}z\right)y^{-1}\left(z^{-1}z\right)x\left(z^{-1}z\right)\left(yz^{-1}\right).$$
6. By _____,
$$\left(zx^{-1}\right)\left(z^{-1}z\right)y^{-1}\left(z^{-1}z\right)x\left(z^{-1}z\right)\left(yz^{-1}\right) = \\ = \left(zx^{-1}\right)ey^{-1}exe\left(yz^{-1}\right).$$
7. By _____, $\left(zx^{-1}\right)ey^{-1}exe\left(yz^{-1}\right) = \left(zx^{-1}\right)y^{-1}x\left(yz^{-1}\right)$.
8. By _____, $\left(zx^{-1}\right)y^{-1}x\left(yz^{-1}\right) = z\left(x^{-1}y^{-1}xy\right)z^{-1}$.
9. By _____, $z\left(x^{-1}y^{-1}xy\right)z^{-1} = z\,[x,y]\,z^{-1}$.
10. By _____, $z\,[x,y]\,z^{-1} = [x,y]^z$.
11. By _____, $[x^z, y^z] = [x,y]^z$.

Figure 2.3. Material for Exercise 2.37(c)

abelian; in Section 3.3 you will learn why a group of order 5 must also be abelian. Thus, D_3 is the smallest non-abelian group.

Intuitive development of D_3

To describe D_3, start with an equilateral triangle in $\mathbb{R}^2$, with its center at the origin. We want to look at its group of symmetries. Intuitively, a "symmetry" is a transformation of the plane that leaves the *triangle* in the same location, even if its *points* are in different locations. "Transformations" include actions like rotation, reflection (flip), and translation (shift). Translating the plane in some direction certainly won't leave the triangle intact, but rotation and reflection can. Two obvious symmetries of an equilateral triangle are a 120° rotation through the origin, and a reflection through the y-axis. We'll call the first of these ρ, and the second φ. See Figure 2.4.

It is helpful to observe two important properties.

Theorem 2.38. If φ and ρ are as specified, then $\varphi\rho = \rho^2\varphi$.

For now, we consider intuitive proofs only. Detailed proofs appear later in the section.

Intuitive proof. The expression $\varphi\rho$ means to apply ρ first, then φ. It'll help if you sketch what

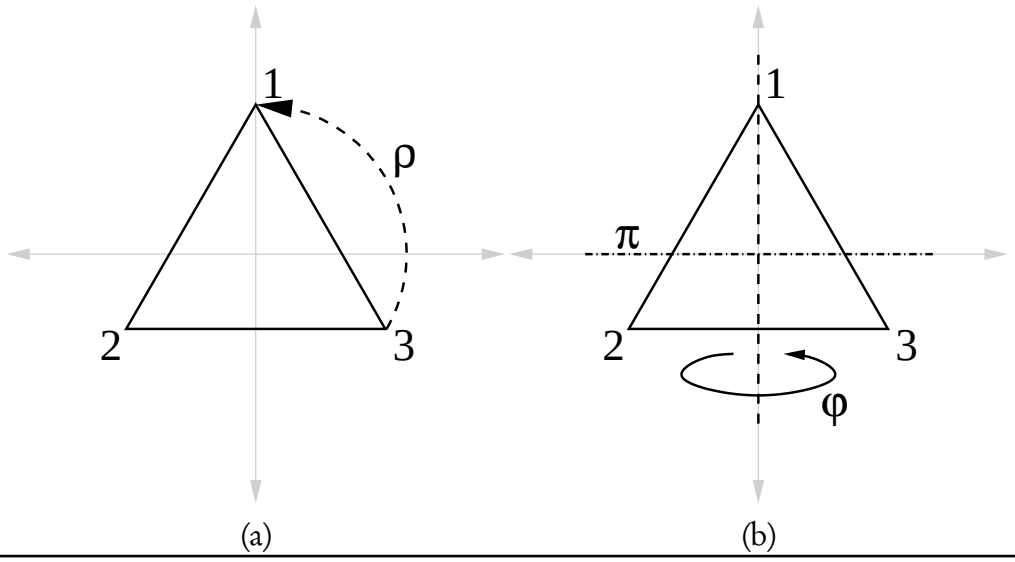


Figure 2.4. Rotation and reflection of the triangle

takes place here. Rotating 120° moves vertex 1 to vertex 2, vertex 2 to vertex 3, and vertex 3 to vertex 1. Flipping through the y-axis leaves the top vertex in place; since we performed the rotation first, the top vertex is now vertex 3, so vertices 1 and 2 are the ones swapped. Thus, vertex 1 has moved to vertex 3, vertex 3 has moved to vertex 1, and vertex 2 is in its original location.

On the other hand, $\rho^2\varphi$ means to apply φ first, then apply ρ twice. Again, it will help to sketch what follows. Flipping through the y-axis swaps vertices 2 and 3, leaving vertex 1 in the same place. Rotating twice then moves vertex 1 to the lower right position, vertex 3 to the top position, and vertex 2 to the lower left position. This is the same arrangement of the vertices as we had for $\varphi\rho$, which means that $\varphi\rho = \rho^2\varphi$. □

You might notice that there's a gap in our reasoning: we showed that the *vertices* of the triangle ended up in the same place, but not the *points in between*. That requires a little more work, which is why we provide detailed proofs later.

By the way, did you notice something interesting about Corollary 2.38? It implies that the operation in D_3 is non-commutative! We have $\varphi\rho = \rho^2\varphi$, and a little logic shows that $\rho^2\varphi \neq \rho\varphi$: thus $\varphi\rho \neq \rho\varphi$. After all, $\rho\varphi$

Another "obvious" symmetry of the triangle is the transformation where you *do nothing* – or, if you prefer, where you effectively *move every point back to itself,* as in a 360° rotation, say. We'll call this symmetry ι. It gives us the last property we need to specify the group, D_3.

Theorem 2.39. In D_3, $\rho^3 = \varphi^2 = \iota$.

Intuitive proof. Rotating 120° three times is the same as rotating 360°, which is the same as not rotating at all! Likewise, φ moves any point (x,y) to $(x,-y)$, and applying φ again moves $(x,-y)$ back to (x,y), which is the same as not flipping at all!

We are now ready to specify D_3. □

Definition 2.40. Let $D_3 = \{\iota, \varphi, \rho, \rho^2, \rho\varphi, \rho^2\varphi\}$.

Theorem 2.41. D_3 is a group under composition of functions.

Proof. To prove this, we will show that all the properties of a group are satisfied. We will start the proof, and leave you to finish it in Exercise 2.45.

Closure: In Exercise 2.45, you will compute the Cayley table of D_3. There, you will see that every composition is also an element of D_3.

Associative: Way back in Section 1.1, we showed that F_S, the set of functions over a set S, was a monoid under composition for *any* set S. To do that, we had to show that composition of functions was associative. There's no point in repeating that proof here; doing it once is good enough for a sane person. Symmetries are functions; after all, they map any point in $\mathbb{R}^2$ to another point in $\mathbb{R}^2$, with no ambiguity about where the point goes. So, we've already proved this.

Identity: We claim that ι is the identity function. To see this, let $\sigma \in D_3$ be any symmetry; we need to show that $\iota\sigma = \iota$ and $\sigma\iota = \sigma$. For the first, apply σ to the triangle. Then apply ι. Since ι effectively leaves everything in place, all the points are in the same place they were after we applied σ. In other words, $\iota\sigma = \sigma$. The proof that $\sigma\iota = \sigma$ is similar.

Alternately, you could look at the result of Exercise 2.45; you will find that $\iota\sigma = \sigma\iota = \sigma$ for every $\sigma \in D_3$.

Inverse: Intuitively, rotation and reflection are one-to-one-functions: after all, if a point P is mapped to a point R by either, it doesn't make sense that another point Q would also be mapped to R. Since one-to-one functions have inverses, every element σ of D_3 must have an inverse function σ^{-1}, which undoes whatever σ did. But is $\sigma^{-1} \in D_3$, also? Since σ maps every point of the triangle onto the triangle, σ^{-1} will undo that map: every point of the triangle will be mapped back onto itself, as well. So, yes, $\sigma^{-1} \in D_3$.

Here, the intuition is a little too imprecise; it isn't *that* obvious that rotation is a one-to-one function. Fortunately, the result of Exercise 2.45 shows that ι, the identity, appears in every row and column. That means that every element has an inverse. □

Detailed proof that D_3 contains all symmetries of the triangle

To prove that D_3 contains *all* symmetries of the triangle, we need to make some notions more precise. First, what is a symmetry? A **symmetry** of *any* polygon is a distance-preserving function on $\mathbb{R}^2$ that maps points of the polygon back onto itself. Notice the careful wording: the *points* of the polygon can change places, but since they have to be mapped back onto the polygon, the polygon itself has to remain in the same place.

Let's look at the specifics for our triangle. What functions are symmetries of the triangle? To answer this question, we divide it into two parts.

1. What are the distance-preserving functions that map $\mathbb{R}^2$ to itself, and leave the origin undisturbed? Here, distance is measured by the usual metric,

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

(You might wonder why we don't want the origin to move. Basically, if a function α preserves both distances between points and a figure centered at the origin, then the origin *cannot* move, since then its distance to points on the figure would change.)

2. Not all of the functions identitifed by question (1) map points on the triangle back onto the triangle; for example a 45° degree rotation does not. Which ones do?

Lemma 2.42 answers the first question.

Lemma 2.42. Let $\alpha : \mathbb{R}^2 \to \mathbb{R}^2$. If

- α does not move the origin; that is, $\alpha(0,0) = (0,0)$, and
- the distance between $\alpha(P)$ and $\alpha(R)$ is the same as the distance between P and R for every $P, R \in \mathbb{R}^2$,

then α has one of the following two forms:

$$\rho = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \quad \exists t \in \mathbb{R}$$

or

$$\varphi = \begin{pmatrix} \cos t & \sin t \\ \sin t & -\cos t \end{pmatrix} \quad \exists t \in \mathbb{R}.$$

The two values of t may be different.

Proof. Assume that $\alpha(0,0) = (0,0)$ and for every $P, R \in \mathbb{R}^2$ the distance between $\alpha(P)$ and $\alpha(R)$ is the same as the distance between P and R. We can determine α precisely merely from how it acts on two points in the plane!

First, let $P = (1,0)$. Write $\alpha(P) = Q = (q_1, q_2)$; this is the point where α moves Q. The distance between P and the origin is 1. Since $\alpha(0,0) = (0,0)$, the distance between Q and the origin is $\sqrt{q_1^2 + q_2^2}$. Because α preserves distance,

$$1 = \sqrt{q_1^2 + q_2^2},$$

or

$$q_1^2 + q_2^2 = 1.$$

The only values for Q that satisfy this equation are those points that lie on the circle whose center is the origin. Any point on this circle can be parametrized as

$$(\cos t, \sin t)$$

where $t \in [0, 2\pi)$ represents an angle. Hence, $\alpha(P) = (\cos t, \sin t)$.

Let $R = (0,1)$. Write $\alpha(R) = S = (s_1, s_2)$. An argument similar to the one above shows that S also lies on the circle whose center is the origin. Moreover, the distance between P and R is $\sqrt{2}$, so the distance between Q and S is also $\sqrt{2}$. That is,

$$\sqrt{(\cos t - s_1)^2 + (\sin t - s_2)^2} = \sqrt{2},$$

or

$$(\cos t - s_1)^2 + (\sin t - s_2)^2 = 2. \tag{7}$$

We can simplify (7) to obtain

$$-2(s_1 \cos t + s_2 \sin t) + \left(s_1^2 + s_2^2\right) = 1. \tag{8}$$

To solve this, recall that the distance from S to the origin must be the same as the distance from R to the origin, which is 1. Hence

$$\begin{aligned} \sqrt{s_1^2 + s_2^2} &= 1 \\ s_1^2 + s_2^2 &= 1. \end{aligned}$$

Substituting this into (8), we find that

$$\begin{aligned} -2(s_1 \cos t + s_2 \sin t) + s_1^2 + s_2^2 &= 1 \\ -2(s_1 \cos t + s_2 \sin t) + 1 &= 1 \\ -2(s_1 \cos t + s_2 \sin t) &= 0 \\ s_1 \cos t &= -s_2 \sin t. \end{aligned} \tag{9}$$

At this point we can see that $s_1 = \sin t$ and $s_2 = -\cos t$ would solve the problem; so would $s_1 = -\sin t$ and $s_2 = \cos t$. Are there any other solutions?

Recall that $s_1^2 + s_2^2 = 1$, so $s_2 = \pm\sqrt{1 - s_1^2}$. Likewise $\sin t = \pm\sqrt{1 - \cos^2 t}$. Substituting into equation (9) and squaring (so as to remove the radicals), we find that

$$\begin{aligned} s_1 \cos t &= -\sqrt{1 - s_1^2} \cdot \sqrt{1 - \cos^2 t} \\ s_1^2 \cos^2 t &= \left(1 - s_1^2\right)\left(1 - \cos^2 t\right) \\ s_1^2 \cos^2 t &= 1 - \cos^2 t - s_1^2 + s_1^2 \cos^2 t \\ s_1^2 &= 1 - \cos^2 t \\ s_1^2 &= \sin^2 t \\ \therefore s_1 &= \pm \sin t. \end{aligned}$$

Along with equation (9), this implies that $s_2 = \mp \cos t$. Thus there are *two* possible values of s_1 and s_2.

It can be shown (see Exercise 2.48) that α is a linear transformation on the vector space $\mathbb{R}^2$ with the basis $\left\{\vec{P}, \vec{R}\right\} = \{(1,0),(0,1)\}$. Linear algebra tells us that we can describe any linear transformation over a finite-dimensional vector space as a matrix. If $s = (\sin t, -\cos t)$ then

$$\alpha = \begin{pmatrix} \cos t & \sin t \\ \sin t & -\cos t \end{pmatrix};$$

otherwise

$$\alpha = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}.$$

The lemma names the first of these forms φ and the second ρ. □

Before answering the second question, let's consider an example of what the two basic forms of α do to the points in the plane.

Example 2.43. Consider the set of points

$$\mathcal{S} = \{(0,2), (\pm 2, 1), (\pm 1, -2)\};$$

these form the vertices of a (non-regular) pentagon in the plane. Let $t = \pi/4$; then

$$\rho = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} \quad \text{and} \quad \varphi = \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix}.$$

If we apply ρ to every point in the plane, then the points of $\mathcal{S}$ move to

$$\begin{aligned} \rho(\mathcal{S}) &= \{\rho(0,2), \rho(-2,1), \rho(2,1), \rho(-1,-2), \rho(1,-2)\} \\ &= \left\{ \left(-\sqrt{2}, \sqrt{2}\right), \left(-\sqrt{2} - \frac{\sqrt{2}}{2}, -\sqrt{2} + \frac{\sqrt{2}}{2}\right), \right. \\ &\qquad \left(\sqrt{2} - \frac{\sqrt{2}}{2}, \sqrt{2} + \frac{\sqrt{2}}{2}\right), \\ &\qquad \left(-\frac{\sqrt{2}}{2} + \sqrt{2}, -\frac{\sqrt{2}}{2} - \sqrt{2}\right), \\ &\qquad \left. \left(\frac{\sqrt{2}}{2} + \sqrt{2}, \frac{\sqrt{2}}{2} - \sqrt{2}\right) \right\} \\ &\approx \{(-1.4, 1.4), (-2.1, -0.7), (0.7, 2.1), \\ &\qquad (0.7, -2.1), (2.1, -0.7)\}. \end{aligned}$$

This is a 45° ($\pi/4$) counterclockwise rotation in the plane.

If we apply φ to every point in the plane, then the points of $\mathcal{S}$ move to

$$\begin{aligned} \varphi(\mathcal{S}) &= \{\varphi(0,2), \varphi(-2,1), \varphi(2,1), \varphi(-1,-2), \varphi(1,-2)\} \\ &\approx \{(1.4, -1.4), (-0.7, -2.1), (2.1, 0.7), \\ &\qquad \downarrow(-2.1, 0.7), (-0.7, 2.1)\}. \end{aligned}$$

This is shown in Figure 2.5 . The line of reflection for φ has slope $\left(1 - \cos\frac{\pi}{4}\right) / \sin\frac{\pi}{4}$. (You will show this in Exercise 2.50)

The second questions asks which of the matrices described by Lemma 2.42 also preserve the triangle.

- The first solution (ρ) corresponds to a rotation of degree t of the plane. To preserve the triangle, we can only have $t = 0, 2\pi/3, 4\pi/3$ (0°, 120°, 240°). (See Figure 2.4(a).) Let ι

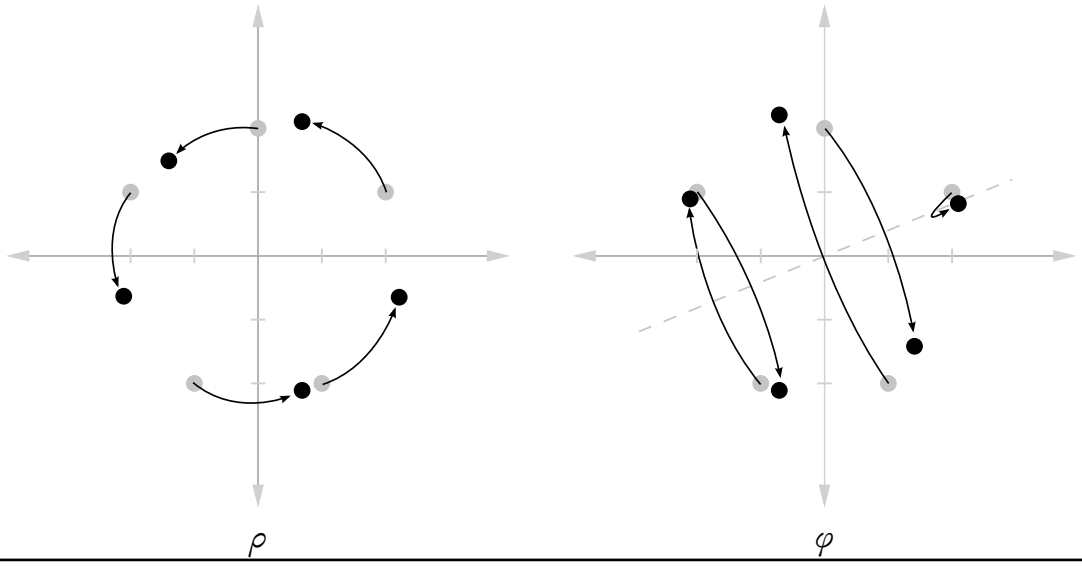


Figure 2.5. Actions of ρ and φ on a pentagon, with $t = \pi/4$

correspond to $t = 0$, the identity rotation; notice that

$$\iota = \begin{pmatrix} \cos 0 & -\sin 0 \\ \sin 0 & \cos 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

which is what we would expect for the identity. We can let ρ correspond to a counterclockwise rotation of 120°, so

$$\rho = \begin{pmatrix} \cos\frac{2\pi}{3} & -\sin\frac{2\pi}{3} \\ \sin\frac{2\pi}{3} & \cos\frac{2\pi}{3} \end{pmatrix} = \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}.$$

A rotation of 240° is the same as rotating 120° twice. We can write that as $\rho \circ \rho$ or ρ^2; matrix multiplication gives us

$$\begin{aligned} \rho^2 &= \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix} \\ &= \begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}. \end{aligned}$$

- The second solution (φ) corresponds to a flip along the line whose slope is

$$m = (1 - \cos t) \,/\, \sin t.$$

One way to do this would be to flip across the y-axis (see Figure 2.4(b)). For this we need the slope to be undefined, so the denominator needs to be zero and the numerator needs to be non-zero. One possibility for t is $t = \pi$ (but not $t = 0$). So

$$\varphi = \begin{pmatrix} \cos\pi & \sin\pi \\ \sin\pi & -\cos\pi \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}.$$

There are two other flips, but we can actually ignore them, because they are combinations of φ and ρ. (Why? See Exercise 2.47.)

We can now give more detailed proofs of Theorems 2.38 and 2.39. We'll prove the first here, and you'll prove the second in the exercises.

Detailed proof of Theorem 2.38. Compare

$$\varphi\rho = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}$$

and

$$\begin{aligned} \rho^2\varphi &= \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}. \end{aligned}$$

□

Exercises.

Unless otherwise specified ρ and φ refer to the elements of D_3.

Exercise 2.44. Show explicitly (by matrix multiplication) that $\rho^3 = \varphi^2 = \iota$.

Exercise 2.45. The multiplication table for D_3 has at least this structure:

$\circ$	ι	φ	ρ	ρ^2	$\rho\varphi$	$\rho^2\varphi$
ι	ι	φ	ρ	ρ^2	$\rho\varphi$	$\rho^2\varphi$
φ	φ		$\rho^2\varphi$			
ρ	ρ	$\rho\varphi$				
ρ^2	ρ^2					
$\rho\varphi$	$\rho\varphi$					
$\rho^2\varphi$	$\rho^2\varphi$					

Complete the multiplication table, writing every element in the form $\rho^m\varphi^n$, never with φ before ρ. *Do not use matrix multiplication;* instead, use Theorems 2.38 and 2.39.

Exercise 2.46. Find a geometric figure (not a polygon) that is preserved by at least one rotation, at least one reflection, *and* at least one translation. Keep in mind that, when we say "preserved", we mean that the points of the figure end up on the figure itself — just as a 120° rotation leaves the triangle on itself.

Exercise 2.47. Two other values of t allow us to define flips for the triangle. Find these values of t, and explain why their matrices are equivalent to the matrices $\rho\varphi$ and $\rho^2\varphi$.

Exercise 2.48. Show that any function α satisfying the requirements of Theorem 2.42 is a linear transformation; that is, for all $P, Q \in \mathbb{R}^2$ and for all $a, b \in \mathbb{R}$, $\alpha(aP + bQ) = a\alpha(P) + b\alpha(Q)$. Use the following steps.

(a) Prove that $\alpha(P) \cdot \alpha(Q) = P \cdot Q$, where $\cdot$ denotes the usual dot product (or inner product) on $\mathbb{R}^2$.

(b) Show that $\alpha(1,0) \cdot \alpha(0,1) = 0$.

(c) Show that $\alpha((a,0) + (0,b)) = a\alpha(1,0) + b\alpha(0,1)$.

(d) Show that $\alpha(aP) = a\alpha(P)$.

(e) Show that $\alpha(P + Q) = \alpha(P) + \alpha(Q)$.

Exercise 2.49. Show that the only stationary point in $\mathbb{R}^2$ for the general ρ is the origin. That is, if $\rho(P) = P$, then $P = (0,0)$. (By "general", we mean any ρ, not just the one in D_3.)

Exercise 2.50. Fill in each blank of Figure 2.6 with the appropriate justification.

Claim: The only stationary points of φ lie along the line whose slope is $(1-\cos t)\,/\,\sin t$, where $t \in [0, 2\pi)$ and $t \neq 0, \pi$. If $t = 0$, only the x-axis is stationary, and for $t = \pi$, only the y-axis.
Proof:

1. Let $P \in \mathbb{R}^2$. By _____, there exist $x, y \in \mathbb{R}$ such that $P = (x, y)$.
2. Assume φ leaves P stationary. By _____,
$$\begin{pmatrix} \cos t & \sin t \\ \sin t & -\cos t \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}.$$
3. By linear algebra,
$$\begin{pmatrix} \_\_\_\_\_ \\ \_\_\_\_\_ \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}.$$
4. By the principle of linear independence, _____$= x$ and _____$= y$.
5. For each equation, collect x on the left hand side, and y on the right, to obtain
$$\begin{cases} x\,(\_\_\_\_\_) = -y\,(\_\_\_\_\_) \\ x\,(\_\_\_\_\_) = \quad y\,(\_\_\_\_\_) \end{cases}.$$
6. If we solve the first equation for y, we find that $y =$_____.
   (a) This, of course, requires us to assume that _____$\neq 0$.
   (b) If that *was* in fact zero, then $t =$_____, _____ (remembering that $t \in [0, 2\pi)$).
7. Put these values of t aside. If we solve the second equation for y, we find that $y =$_____.
   (a) Again, this requires us to assume that _____$\neq 0$.
   (b) If that *was* in fact zero, then $t =$_____. We already put this value aside, so ignore it.
8. Let's look at what happens when $t \neq$_____ and _____.
   (a) Multiply numerator and denominator of the right hand side of the first solution by the denominator of the second to obtain $y =$_____.
   (b) Multiply right hand side of the second with denominator of the first: $y =$_____.
   (c) By _____, $\sin^2 t = 1 - \cos^2 t$. Substitution into the second solution gives the first!
   (d) That is, points that lie along the line $y =$_____ are left stationary by φ.
9. Now consider the values of t we excluded.
   (a) If $t =$_____, then the matrix simplifies to $\varphi =$_____.
   (b) To satisfy $\varphi(P) = P$, we must have _____$= 0$, and _____ free. The points that satisfy this are precisely the _____-axis.
   (c) If $t =$_____, then the matrix simplifies to $\varphi =$_____.
   (d) To satisfy $\varphi(P) = P$, we must have _____$= 0$, and _____ free. The points that satisfy this are precisely the _____-axis.

Figure 2.6. Material for Exercise 2.50

2.3: Cyclic groups and order of elements

Here we re-introduce the familiar notation of exponents, in a manner consistent with what you learned for exponents of real numbers. We use this to describe an important class of groups that recur frequently.

Cyclic groups and generators

Notation 2.51. Let G be a group, and $g \in G$. If we want to perform the operation on g ten times, we could write

$$\prod_{i=1}^{10} g = g \cdot g \cdot g \cdot g \cdot g \cdot g \cdot g \cdot g \cdot g \cdot g$$

but this grows tiresome. Instead we will adapt notation from high-school algebra and write

$$g^{10}.$$

We likewise define g^{-10} to represent

$$\prod_{i=1}^{10} g^{-1} = g^{-1} \cdot g^{-1} \cdot g^{-1} \cdot g^{-1} \cdot g^{-1} \cdot g^{-1} \cdot g^{-1} \cdot g^{-1} \cdot g^{-1} \cdot g^{-1}.$$

Indeed, for any $n \in \mathbb{N}^+$ and any $g \in G$ we adopt the following convention:

- g^n means to perform the operation on n copies of g, so $g^n = \prod_{i=1}^{n} g$;
- g^{-n} means to perform the operation on n copies of g^{-1}, so $g^{-n} = \prod_{i=1}^{n} g^{-1} = \left(g^{-1}\right)^n$;
- $g^0 = e$, and if I want to be annoying I can write $g^0 = \prod_{i=1}^{0} g$.

In additive groups we write instead $ng = \sum_{i=1}^{n} g$, $(-n)\,g = \sum_{i=1}^{n} (-g)$, and $0g = 0$.

Notice that this definition assume n is *positive*.

Definition 2.52. Let G be a group. If there exists $g \in G$ such that every element $x \in G$ has the form $x = g^n$ for some $n \in \mathbb{Z}$, then G is a **cyclic group** and we write $G = \langle g \rangle$. We call g a **generator** of G.

The idea of a cyclic group is that it has the form

$$\left\{\ldots, g^{-2}, g^{-1}, e, g^1, g^2, \ldots\right\}.$$

If the group is additive, we would of course write

$$\{\ldots, -2g, -g, 0, g, 2g, \ldots\}.$$

Example 2.53. $\mathbb{Z}$ is cyclic, since any $n \in \mathbb{Z}$ has the form $n \cdot 1$. Thus $\mathbb{Z} = \langle 1 \rangle$. In addition, n has the form $(-n) \cdot (-1)$, so $\mathbb{Z} = \langle -1 \rangle$ as well. Both 1 and -1 are generators of $\mathbb{Z}$.

You will show in the exercises that $\mathbb{Q}$ is not cyclic.

In Definition 2.52 we referred to g as *a* generator of G, not as *the* generator. There could in fact be more than one generator; we see this in Example 2.53 from the fact that $\mathbb{Z} = \langle 1 \rangle = \langle -1 \rangle$. Here is another example.

Example 2.54. Let

$$G = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \right\} \subsetneq \mathrm{GL}_m(\mathbb{R}).$$

It turns out that G is a group; both the second and third matrices generate it. For example,

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

An important question arises here. Given a group G and an element $g \in G$, define

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

We know that every cyclic group has the form $\langle g \rangle$ for some $g \in G$. Is the converse also true that $\langle g \rangle$ is a group for any $g \in G$? As a matter of fact, yes!

Theorem 2.55. For every group G and for every $g \in G$, $\langle g \rangle$ is an abelian group.

To prove Theorem 2.55, we need to make sure we can perform the usual arithmetic on exponents.

Lemma 2.56. Let G be a group, $g \in G$, and $m, n \in \mathbb{Z}$. Each of the following holds:
(A) $g^m g^{-m} = e$; that is, $g^{-m} = (g^m)^{-1}$.
(B) $(g^m)^n = g^{mn}$.
(C) $g^m g^n = g^{m+n}$.

The proof will justify this argument by applying the notation described at the beginning of this chapter. We have to be careful with this approach, because in the lemma we have $m, n \in \mathbb{Z}$, but the notation was given under the assumption that $n \in \mathbb{N}^+$. To make this work, we'll have to consider the cases where m and n are positive or negative separately. We call this a *case analysis*.

Proof. Each claim follows by case analysis.

(A) If $m = 0$, then $g^{-m} = g^0 = e = e^{-1} = \left(g^0\right)^{-1} = (g^m)^{-1}$.

Otherwise, $m \neq 0$. First assume that $m \in \mathbb{N}^+$. By notation, $g^{-m} = \prod_{i=1}^{m} g^{-1}$. Hence

$$\begin{aligned} g^m g^{-m} &\underset{\text{def.}}{=} \left(\prod_{i=1}^{m} g\right)\left(\prod_{i=1}^{m} g^{-1}\right) \\ &\underset{\text{ass.}}{=} \left(\prod_{i=1}^{m-1} g\right)\left(g \cdot g^{-1}\right)\left(\prod_{i=1}^{m-1} g^{-1}\right) \\ &\underset{\text{id.}}{=} \left(\prod_{i=1}^{m-1} g\right) e \left(\prod_{i=1}^{m-1} g^{-1}\right) \\ &\underset{\text{inv.}}{=} \left(\prod_{i=1}^{m-1} g\right)\left(\prod_{i=1}^{m-1} g^{-1}\right) \\ &\;\;\vdots \\ &= e. \end{aligned}$$

Since the inverse of an element is unique, $g^{-m} = (g^m)^{-1}$.

Now assume that $m \in \mathbb{Z}\backslash\mathbb{N}$. Since m is negative, we cannot express the product using m; the notation discussed on page 76 requires a *positive* exponent. Consider instead $\widehat{m} = |m| \in \mathbb{N}^+$. Since the opposite of a negative number is positive, we can write $-m = \widehat{m}$ and $-\widehat{m} = m$. Since $\widehat{m}$ is positive, we can apply the notation to it directly; $g^{-m} = g^{\widehat{m}} = \prod_{i=1}^{\widehat{m}} g$, while $g^m = g^{-\widehat{m}} = \prod_{i=1}^{\widehat{m}} g^{-1}$. (To see this in a more concrete example, try it with an actual number. If $m = -5$, then $\widehat{m} = |-5| = 5 = -(-5)$, so $g^m = g^{-5} = g^{-\widehat{m}}$ and $g^{-m} = g^5 = g^{\widehat{m}}$.) As above, we have

$$g^m g^{-m} \underset{\text{subs.}}{=} g^{-\widehat{m}} g^{\widehat{m}} \underset{\text{not.}}{=} \left(\prod_{i=1}^{\widehat{m}} g^{-1}\right)\left(\prod_{i=1}^{\widehat{m}} g\right) = e.$$

Hence $g^{-m} = (g^m)^{-1}$.

(B) If $n = 0$, then $(g^m)^n = (g^m)^0 = e$ because *anything* to the zero power is e. Assume first that $n \in \mathbb{N}^+$. By notation, $(g^m)^n = \prod_{i=1}^{n} g^m$. We split this into two subcases.

(B1) If $m \in \mathbb{N}$, we have

$$(g^m)^n \underset{\text{not.}}{=} \prod_{i=1}^{n}\left(\prod_{i=1}^{m} g\right) \underset{\text{ass.}}{=} \prod_{i=1}^{mn} g \underset{\text{not.}}{=} g^{mn}.$$

(B2) Otherwise, let $\widehat{m} = |m| \in \mathbb{N}^+$ and we have

$$\begin{aligned} (g^m)^n &\underset{\text{subs.}}{=} \left(g^{-\widehat{m}}\right)^n \underset{\text{not.}}{=} \prod_{i=1}^{n}\left(\prod_{i=1}^{\widehat{m}} g^{-1}\right) \\ &\underset{\text{ass.}}{=} \prod_{i=1}^{\widehat{m}n} g^{-1} \underset{\text{not.}}{=} \left(g^{-1}\right)^{\widehat{m}n} \\ &\underset{\text{not.}}{=} g^{-\widehat{m}n} \underset{\text{subs.}}{=} g^{mn}. \end{aligned}$$

What if n is negative? Let $\widehat{n} = -n$; by notation, $(g^m)^n = (g^m)^{-\widehat{n}} = \prod_{i=1}^{\widehat{n}} (g^m)^{-1}$. By (A), this becomes $\prod_{i=1}^{\widehat{n}} g^{-m}$. By notation, we can rewrite this as $\left(g^{-m}\right)^{\widehat{n}}$. Since $\widehat{n} \in \mathbb{N}^+$, we can apply case (B1) or (B2) as appropriate, so

$$(g^m)^n = \left(g^{-m}\right)^{\widehat{n}} \underset{\text{(B1) or (B2)}}{=} g^{(-m)\widehat{n}} \underset{\text{integers!}}{=} g^{m(-\widehat{n})} \underset{\text{subst}}{=} g^{mn}.$$

(C) We consider three cases.
If $m = 0$ or $n = 0$, then $g^0 = e$, so $g^{-0} = g^0 = e$.
If m, n have the same sign (that is, $m, n \in \mathbb{N}^+$ or $m, n \in \mathbb{Z}\backslash\mathbb{N}$), then write $\widehat{m} = |m|$, $\widehat{n} = |n|$, $g_m = g^{\frac{\widehat{m}}{m}}$, and $g_n = g^{\frac{\widehat{n}}{n}}$. This effects a really nice trick: if $m \in \mathbb{N}^+$, then $g_m = g$, whereas if m is negative, $g_m = g^{-1}$. This notational trick allows us to write $g^m = \prod_{i=1}^{\widehat{m}} g_m$ and $g^n = \prod_{i=1}^{\widehat{n}} g_n$, where $g_m = g_n$ and $\widehat{m}$ and $\widehat{n}$ are both positive integers. Then

$$\begin{aligned} g^m g^n &= \prod_{i=1}^{\widehat{m}} g_m \prod_{i=1}^{\widehat{n}} g_n = \prod_{i=1}^{\widehat{m}} g_m \prod_{i=1}^{\widehat{n}} g_m \\ &= \prod_{i=1}^{\widehat{m}+\widehat{n}} g_m = (g_m)^{\widehat{m}+\widehat{n}} = g^{m+n}. \end{aligned}$$

Since g and n were arbitrary, the induction implies that $g^n g^{-n} = e$ for all $g \in G$, $n \in \mathbb{N}^+$. Now consider the case where m and n have different signs. In the first case, suppose m is negative and $n \in \mathbb{N}^+$. As in (A), let $\widehat{m} = |m| \in \mathbb{N}^+$; then

$$g^m g^n = \left(g^{-1}\right)^{-m} g^n = \left(\prod_{i=1}^{\widehat{m}} g^{-1}\right)\left(\prod_{i=1}^{n} g\right).$$

If $\widehat{m} \geq n$, we have more copies of g^{-1} than g, so after cancellation,

$$g^m g^n = \prod_{i=1}^{\widehat{m}-n} g^{-1} = g^{-(\widehat{m}-n)} = g^{m+n}.$$

Otherwise, $\widehat{m} < n$, and we have more copies of g than of g^{-1}. After cancellation,

$$g^m g^n = \prod_{i=1}^{n-\widehat{m}} g = g^{n-\widehat{m}} = g^{n+m} = g^{m+n}.$$

The remaining case ($m \in \mathbb{N}^+$, $n \in \mathbb{Z}\backslash\mathbb{N}$) is similar, and you will prove it for homework.

□

These properties of exponent arithmetic allow us to show that $\langle g \rangle$ is a group.

Proof of Theorem 2.55. We show that $\langle g \rangle$ satisfies the properties of an abelian group. Let $x, y, z \in \langle g \rangle$. By definition of $\langle g \rangle$, there exist $a, b, c \in \mathbb{Z}$ such that $x = g^a$, $y = g^b$, and $z = g^c$. We will

use Lemma 2.56 implicitly.

- By substitution, $xy = g^a g^b = g^{a+b} \in \langle g \rangle$. So $\langle g \rangle$ is closed.
- By substitution, $x(yz) = g^a \left(g^b g^c\right)$. These are elements of G by inclusion (that is, $\langle g \rangle \subseteq G$ so $x, y, z \in G$), so the associative property *in* G gives us

$$x(yz) = g^a \left(g^b g^c\right) = \left(g^a g^b\right) g^c = (xy)z.$$

- By definition, $e = g^0 \in \langle g \rangle$.
- By definition, $g^{-a} \in \langle g \rangle$, and $x \cdot g^{-a} = g^a g^{-a} = e$. Hence $x^{-1} = g^{-a} \in \langle g \rangle$.
- Using the fact that $\mathbb{Z}$ is commutative under addition,

$$xy = g^a g^b = g^{a+b} = g^{b+a} = g^b g^a = yx.$$

□

The order of an element

Given an element and an operation, Theorem 2.55 links them to a group. It makes sense, therefore, to link an element to the order of the group that it generates.

Definition 2.57. Let G be a group, and $g \in G$. We say that the **order** of g is $\operatorname{ord}(g) = |\langle g \rangle|$. If $\operatorname{ord}(g) = \infty$, we say that g has **infinite order**.

If the order of a group is finite, then we can write an element in different ways.

Example 2.58. Recall Example 2.54; we can write

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^0 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^4 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^8 = \cdots.$$

Since multiples of 4 give the identity, let's take any power of the matrix, and divide it by 4. The Division Theorem allows us to write any power of the matrix as $4q + r$, where $0 \leq r < 4$. Since there are only four possible remainders, and multiples of 4 give the identity, positive powers of this matrix can generate only four possible matrices:

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^{4q} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^{4q+1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^{4q+2} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix},$$

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^{4q+3} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

We can do the same with negative powers; the Division Theorem still gives us only four possible remainders. Let's write

$$g = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Thus

$$\langle g \rangle = \left\{ I_2, g, g^2, g^3 \right\}.$$

The example suggests that if the order of an element G is $n \in \mathbb{N}$, then we can write

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

This explains why we call $\langle g \rangle$ a *cyclic* group: once they reach $\operatorname{ord}(g)$, the powers of g "cycle".To prove this in general, we have to show that for a generic cyclic group $\langle g \rangle$ with $\operatorname{ord}(g) = n$,

- n is the smallest positive power that gives us the identity; that is, $g^n = e$, and
- for any two integers between 0 and n, the powers of g are different; that is, if $0 \leq a < b < n$, then $g^a \neq g^b$.

Theorem 2.59 accomplishes that, and a bit more as well.

Theorem 2.59. Let G be a group, $g \in G$, and $\operatorname{ord}(g) = n$. Then
(A) for all $a, b \in \mathbb{N}$ such that $0 \leq a < b < n$, we have $g^a \neq g^b$.
In addition, if $n < \infty$, each of the following holds:
(B) $g^n = e$;
(C) n is the smallest positive integer d such that $g^d = e$; and
(D) if $a, b \in \mathbb{Z}$ and $n \mid (a - b)$, then $g^a = g^b$.

Proof. The fundamental assertion of the theorem is (A). The remaining assertions turn out to be corollaries.

(A) By way of contradiction, suppose that there exist $a, b \in \mathbb{N}$ such that $0 \leq a < b < n$ and $g^a = g^b$; then $e = (g^a)^{-1} g^b$. By Lemma 2.56, we can write

$$e = g^{-a} g^b = g^{-a+b} = g^{b-a}.$$

Let $S = \left\{ m \in \mathbb{N}^+ : g^m = e \right\}$. By the well-ordering property of $\mathbb{N}$, there exists a smallest element of S; call it d. Recall that $a < b$, so $b - a \in \mathbb{N}^+$, so $g^{b-a} \in S$. By the choice of d, we know that $d \leq b - a$. By Exercise 0.25, $d \leq b - a < b$, so $0 < d < b < n$.
We can now list d distinct elements of $\langle g \rangle$:

$$g, g^2, g^3, \ldots, g^d = e. \tag{10}$$

Using Lemma 2.56 again, we extrapolate that $g^{d+1} = g$, $g^{d+2} = g^2$, etc., so

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

We see that $|\langle g \rangle| = d$, but this contradicts the assumption that $n = \operatorname{ord}(g) = |\langle d \rangle|$.

(B) Let $S = \left\{ m \in \mathbb{N}^+ : g^m = e \right\}$. Is S non-empty? Since $\langle g \rangle < \infty$, there must exist $a, b \in \mathbb{N}^+$ such that $a < b$ and $g^a = g^b$. Using the inverse property and substitution, $g^0 = e =$

$g^b (g^a)^{-1}$. By Lemma 2.56, $g^0 = g^{b-a}$. By definition, $b-a \in \mathbb{N}^+$. Hence S is non-empty.
By the well-ordering property of $\mathbb{N}$, there exists a smallest element of S; call it d. Since $\langle g \rangle$ contains n elements, $1 < d \leq n$. If $d < n$, that would contradict assertion (A) of this theorem (with $a = 0$ and $b = d$). Hence $d = n$, and $g^n = e$, and we have shown (A).

(C) In (B), S is the set of all positive integers m such that $g^m = e$; we let the smallest element be d, and thus $d \leq n$. On the other hand, (A) tells us that we cannot have $d < n$; otherwise, $g^d = g^0 = e$. Hence, $n \leq d$. We already had $d \leq n$, so the two must be equal.

(D) Let $a, b \in \mathbb{Z}$. Assume that $n \mid (a-b)$. Let $q \in \mathbb{Z}$ such that $nq = a-b$. Then

$$\begin{aligned} g^b &= g^b \cdot e = g^b \cdot e^q \\ &= g^b \cdot (g^n)^q = g^b \cdot g^{nq} \\ &= g^b \cdot g^{a-b} = g^{b+(a-b)} = g^a. \end{aligned}$$

□

We conclude therefore that, at least when they are finite, cyclic groups are aptly named: increasing powers of g generate new elements until the power reaches n, in which case $g^n = e$ and we "cycle around".

Exercises.

Exercise 2.60. Recall from Example 2.54 the matrix

$$A = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Express A as a power of the other non-identity matrices of the group.

Exercise 2.61. Complete the proof of Lemma 2.56(C).

Exercise 2.62. Fill in each blank of Figure 2.7 with the justification or statement.

Exercise 2.63. Show that any group of 3 elements is cyclic.

Exercise 2.64. Is the Klein 4-group (Exercise 2.32 on page 65) cyclic? What about the cyclic group of order 4?

Exercise 2.65. Show that $\mathbb{Q}$ is not cyclic.

Exercise 2.66. Use a fact from linear algebra to explain why $\mathrm{GL}_m(\mathbb{R})$ is not cyclic.

2.4: The roots of unity

One of the major motivations in the development of group theory was to study roots of polynomials. A polynomial, of course, has the form

$$ax+b, \quad ax^2+bx+c, \quad ax^3+bx^2+cx+d, \quad \ldots$$

Let G be a group, and $g \in G$. Let $d, n \in \mathbb{Z}$ and assume $\text{ord}(g) = d$.
Claim: $g^n = e$ if and only if $d \mid n$.
Proof:

1. Assume that $g^n = e$.
   (a) By _____, there exist $q, r \in \mathbb{Z}$ such that $n = qd + r$ and $0 \leq r < d$.
   (b) By _____, $g^{qd+r} = e$.
   (c) By _____, $g^{qd} g^r = e$.
   (d) By _____, $\left(g^d\right)^q g^r = e$.
   (e) By _____, $e^q g^r = e$.
   (f) By _____, $eg^r = e$. By the identity property, $g^r = e$.
   (g) By _____, d is the *smallest* positive integer such that $g^d = e$.
   (h) Since _____, it cannot be that r is positive. Hence, $r = 0$.
   (i) By _____, $g = qd$. By definition, then $d \mid n$.
2. Now we show the converse. Assume that _____.
   (a) By definition of divisibility, _____.
   (b) By substitution, $g^n =$_____.
   (c) By Lemma 2.56, the right hand side of that equation can be rewritten as to _____.
   (d) Recall that $\text{ord}(g) = d$. By Theorem 2.59, $g^d = e$, so we can rewrite the right hand side again as _____.
   (e) A little more simplification turns the right hand side into _____, which obviously simplifies to e.
   (f) By _____, then, $g^n = e$.
3. We showed first that if $g^n = e$, then $d|n$; we then showed that _____. This proves the claim.

Figure 2.7. Material for Exercise 2.62

A **root** of a polynomial $f(x)$ is any a such that $f(a) = 0$. For example, if $f(x) = x^4 - 1$, then 1 and -1 are both roots of f. However, they are not the *only* roots of f! For the full explanation, you'll need to read about polynomial rings and ideals in Chapters 7 and 8, but we can take some first steps in that direction already.

Imaginary and complex numbers

First, notice that f factors as $f(x) = (x-1)(x+1)\left(x^2+1\right)$. The roots 1 and -1 show up in the linear factors, and they're the only possible roots of those factors. So, if f has other roots, we would expect them to be roots of $x^2 + 1$. However, the square of a real number is nonnegative; adding 1 forces it to be positive. So, $x^2 + 1$ has no roots in $\mathbb{R}$.

Let's make a root up, anyway. If it doesn't make sense, we should find out soon enough. Let's call this polynomial $g(x) = x^2 + 1$, and say that g has a root, which we'll call i, for "imaginary". Since i is a root of g, we have the equation

$$0 = g(i) = i^2 + 1,$$

or $i^2 = -1$.

We'll create a new set of numbers by adding i to the set $\mathbb{R}$. Since $\mathbb{R}$ is a monoid under multiplication and a group under addition, we'd like to preserve those properties as well. This

means we have to define multiplication and addition for our new set, and maybe add more objects, too.

We start with $\mathbb{R} \cup \{i\}$. Does multiplication add any new elements? Since $i^2 = -1$, and $-1 \in \mathbb{R}$ already, we're okay there. On the other hand, for any $b \in \mathbb{R}$, we'd like to multiply b and i. Since bi is not already in our new set, we'll have to add it if we want to keep multiplication closed. Our set has now expanded to $\mathbb{R} \cup \{bi : b \in \mathbb{R}\}$.

Let's look at addition. Our new set has real numbers like 1 and "imaginary" numbers like $2i$; if addition is to satisfy closure, we need $1 + 2i$ to be in the set, too. That's not the case yet, so we have to extend our set by $a + bi$ for any $a, b \in \mathbb{R}$. That gives us

$$\mathbb{R} \cup \{bi : b \in \mathbb{R}\} \cup \{a + bi : a, b \in \mathbb{R}\}.$$

If you think about it, the first two sets are in the third; just let $a = 0$ or $b = 0$ and you get bi or a, respectively. So, we can simplify our new set to

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

Do we need anything else?

We haven't checked closure of addition. In fact, we still haven't *defined* addition of complex numbers. We will borrow an idea from polynomials, and add complex numbers by adding like terms; that is, $(a + bi) + (c + di) = (a + c) + (b + d)i$. Closure implies that $a + c \in \mathbb{R}$ and $b + d \in \mathbb{R}$, so this is just another expression in the form already described. In fact, we can also see what additive inverses look like; after all, $(a + bi) + (-a - bi) = 0$. We don't have to add any new objects to our set to maintain the group structure of addition.

We also haven't checked closure of multiplication in this larger set — or even defined it, really. Again, let's borrow an idea from polynomials, and multiply complex numbers using the distributive property; that is,

$$(a + bi)(c + di) = ac + adi + bci + bdi^2.$$

Remember that $i^2 = -1$, and we can combine like terms, so the expression above simplifies to

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

Since $ac - bd \in \mathbb{R}$ and $ad + bc \in \mathbb{R}$, this is just another expression in the form already described. Again, we don't have to add any new objects to our set.

Definition 2.67. The **complex numbers** are the set

$$\mathbb{C} = \left\{a + bi : a, b \in \mathbb{R},\ i^2 = -1\right\}.$$

The **real part** of $a + bi$ is a, and the **imaginary part** is b.

We can now state with confidence that we have found what we wanted to obtain.

Theorem 2.68. $\mathbb{C}$ is a monoid under multiplication, and an abelian group under addition.

Proof. Let $x, y, z \in \mathbb{C}$. Write $x = a + bi$, $y = c + di$, and $z = e + fi$, for some $a, b, c, d, e, f \in \mathbb{R}$. Let's look at multiplication first.

closure? We built $\mathbb{C}$ to be closed under multiplication, so the discussion above suffices.

associative? We need to show that

$$(xy)\,z = x\,(yz)\,. \tag{11}$$

Expanding the product on the left, we have

$$[(a+bi)\,(c+di)]\,(e+fi) = [(ac-bd)+(ad+bc)\,i]\,(e+fi)\,.$$

Expand again, and we get

$$\begin{aligned}[(a+bi)\,(c+di)]\,(e+fi) = {} & [(ac-bd)\,e-(ad+bc)\,f] \\ & +[(ac-bd)\,f+(ad+bc)\,e]\,i.\end{aligned}$$

Now let's look at the product on the right of equation (11). Expanding it, we have

$$(a+bi)\,[(c+di)\,(e+fi)] = (a+bi)\,[(ce-df)+(cf+de)\,i]\,.$$

Expand again, and we get

$$\begin{aligned}(a+bi)\,[(c+di)\,(e+fi)] = {} & [a\,(ce-df)-b\,(cf+de)] \\ & +[a\,(cf+de)+b\,(ce-df)]\,i.\end{aligned}$$

If you look carefully, you will see that both expansions resulted in the same complex number:

$$(ace-bde-adf-bcf)+(acf-bdf+ade+bce)\,i.$$

Thus, multiplication is $\mathbb{C}$ is associative.

identity? We claim that $1 \in \mathbb{R}$ is the multiplicative identity even for $\mathbb{C}$. Recall that we can write $1 = 1 + 0i$. Then,

$$1x = (1+0i)\,(a+bi) = (1a-0b)+(1b+0a)\,i = a+bi = x.$$

Since x was arbitrary in $\mathbb{C}$, it must be that 1 is, in fact, the identity.

We have shown that $\mathbb{C}$ is a monoid under multiplication. What about addition; it is a group? We leave that to the exercises. □

There are a *lot* of wonderful properties of $\mathbb{C}$ that we could discuss. For example, you can see that the roots of x^2+1 lie in $\mathbb{C}$, but what of the roots of x^2+2? It turns out that they're in there, too. In fact, *every* polynomial of degree n with real coefficients has n roots in $\mathbb{C}$! We need a lot more theory to discuss that, however, so we pass over it for the time being. In any case, we can now talk about a group that is both interesting and important.

Remark 2.69. You may wonder if we really *can* just make up some number i, and build a new set by adjoining it to $\mathbb{R}$. Isn't that just a little, oh, *imaginary*? Actually, no, it is quite concrete, and we can provide two very sound justifications.

First, mathematicians typically model the oscillation of a pendulum by a differential equations of the form $y'' + ay = 0$. As any book in the subject explains, we have good reason to solve such *differential* equations by resorting to *auxiliary polynomial* equations of the form $r^2 + a = 0$. The solutions to this equation are $r = \pm i\sqrt{a}$, so unless the oscillation of a pendulum is "imaginary", i is quite "real".

Second, we can construct from the real numbers a set that looks an awful lot like these purported complex numbers, using a very sensible approach, and we can even show that this set is isomorphic to the complex numbers in all the ways that we would like. That's a bit beyond us; you will learn more in Section 8.3.

The complex plane

We can diagram the real numbers along a line. In fact, it's quite easy to argue that what makes real numbers "real" is precisely the fact that they measure location or distance along a line. That's only one-dimensional, and you've seen before that we can do something similar on the plane or in space using $\mathbb{R}^2$ and $\mathbb{R}^3$.

What about the complex numbers? By definition, any complex number is the sum of its real and imaginary parts. We cannot simplify $a + bi$ any further using this representation, much as we cannot simplify the point $(a, b) \in \mathbb{R}^2$ any further. Since $\mathbb{R}^2$ forms a *vector space* over $\mathbb{R}$, does $\mathbb{C}$ also form a vector space over $\mathbb{R}$? In fact, it does! Here's a quick reminder of what makes a vector space:

- addition of vectors must satisfy closure and the associative, commutative, identity, and inverse properties;
- multiplication of vectors *by scalars* must have an identity scalar, must be associative on the scalars, and must satisfy the properties of distribution of scalars to vectors and vice-versa.

The properties for addition of vectors are precisely the properties of a group — and Theorem 2.68 tells us that $\mathbb{C}$ *is* a group under addition! All that remains is to show that $\mathbb{C}$ satisfies the required properties of multiplication. You will do that in Exercise 2.83.

Right now, we are more interested in the *geometric* implications of this relationship. We've already hinted that $\mathbb{C}$ and $\mathbb{R}^2$ have a similar structure. Let's start with the notion of *dimension*. Do you remember what that word means? Essentially, the dimension of a vectors space is the number of *basis vectors* needed to describe a vector space. Do $\mathbb{C}$ and $\mathbb{R}^2$ have the same dimension over $\mathbb{R}$? For that, we need to identify a *basis* of $\mathbb{C}$ over $\mathbb{R}$.

Theorem 2.70. $\mathbb{C}$ is a vector space over $\mathbb{R}$ with basis $\{1, i\}$.

Proof. We have already discussed why $\mathbb{C}$ is a vector space over $\mathbb{R}$; we still have to show that $\{0, i\}$ is a basis of $\mathbb{C}$. This is straightforward from the definition of $\mathbb{C}$, as any element can be written in terms of the basis elements as $a + bi = a \cdot 1 + b \cdot i$. □

We see from Theorem 2.70 that $\mathbb{C}$ and $\mathbb{R}^2$ do have the same dimension! After all, any point of $\mathbb{R}^2$ can be written as $(a, b) = a\,(1, 0) + b\,(0, 1)$, so a basis of $\mathbb{R}^2$ is $\{(1, 0), (0, 1)\}$.

This will hopefully prompt you to realize that $\mathbb{C}$ and $\mathbb{R}^2$ are identical as vector spaces. For our purposes, what matters that we can map any point of $\mathbb{C}$ to a unique point of $\mathbb{R}^2$, and vice-versa.

Theorem 2.71. There is a one-to-one, onto function from $\mathbb{C}$ to $\mathbb{R}^2$ that maps the basis vectors 1 to $(1, 0)$ and i to $(0, 1)$.

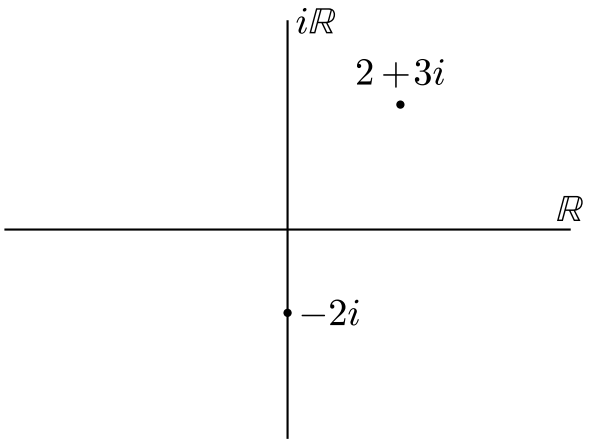


Figure 2.8. Two elements of $\mathbb{C}$, visualized as points on the complex plane

Proof. Let $\varphi : \mathbb{C} \to \mathbb{R}^2$ by $\varphi(a+bi) = (a,b)$. That is, we map a complex number to $\mathbb{R}^2$ by sending the real part to the first entry (the x-ordinate) and the imaginary part to the second entry (the y-ordinate). As desired, $\varphi(1) = (1,0)$ and $\varphi(i) = (0,1)$.

Is this a bijection? We see that φ is one-to-one by the fact that if $\varphi(a+bi) = \varphi(c+di)$, then $(a,b) = (c,d)$; equality of points in $\mathbb{R}^2$ implies that $a = c$ and $b = d$; equality of complex numbers implies that $a+bi = c+di$. We see that φ is onto by the fact that for any $(a,b) \in \mathbb{R}^2$, $\varphi(a+bi) = (a,b)$. □

Since $\mathbb{R}^2$ has a nice, geometric representation as the x-y plane, we can represent complex numbers in the same way. That motivates our definition of the **complex plane**, which is nothing more than a visualization of $\mathbb{C}$ in $\mathbb{R}^2$.

Take a look at Figure 2.8. We have labeled the x-axis as $\mathbb{R}$ and the y-axis as $i\mathbb{R}$. We call the former the **real axis** and the latter the **imaginary axis** of the complex plane. This agrees with our mapping above, which sent the real part of a complex number to the x-ordinate, and the imaginary part to the y-ordinate. Thus, the complex number $2+3i$ corresponds to the point $(2,3)$, while the complex number $-2i$ corresponds to the point $(0,-2)$.

We could say a great deal about the complex plane, but that would distract us from our main goal, which is to proceed further in group theory. Even so, we should not neglect one important and beautiful point.

Roots of unity

Any root of the polynomial $f(x) = x^n - 1$ is called a **root of unity**. These are very important in the study of polynomial roots. At least some of them satisfy a very nice form.

Theorem 2.72. Let $n \in \mathbb{N}^+$. The complex number

$$\omega = \cos\left(\frac{2\pi}{n}\right) + i\sin\left(\frac{2\pi}{n}\right)$$

is a root of $f(x) = x^n - 1$.

To prove Theorem 2.72, we need a different property of ω.

Lemma 2.73. If ω is defined as in Theorem 2.72, then

$$\omega^m = \cos\left(\frac{2\pi m}{n}\right) + i\sin\left(\frac{2\pi m}{n}\right)$$

for every $m \in \mathbb{N}^+$.

Proof. We proceed by induction on m. For the *inductive base*, the definition of ω shows that ω^1 has the desired form. For the *inductive hypothesis*, assume that ω^m has the desired form; in the *inductive step*, we need to show that

$$\omega^{m+1} = \cos\left(\frac{2\pi(m+1)}{n}\right) + i\sin\left(\frac{2\pi(m+1)}{n}\right).$$

To see why this is true, use the trigonometric sum identities $\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$ and $\sin(\alpha+\beta) = \sin\alpha\cos\beta + \sin\beta\cos\alpha$ to rewrite ω^{m+1}, like so:

$$\begin{aligned}
\omega^{m+1} &= \omega^m \cdot \omega \\
&\underset{\text{hyp.}}{=}{}^{\text{ind.}} \left[\cos\left(\frac{2\pi m}{n}\right) + i\sin\left(\frac{2\pi}{n}\right)\right] \\
&\qquad \cdot \left[\cos\left(\frac{2\pi}{n}\right) + i\sin\left(\frac{2\pi}{n}\right)\right] \\
&= \cos\left(\frac{2\pi m}{n}\right)\cos\left(\frac{2\pi}{n}\right) + i\sin\left(\frac{2\pi}{n}\right)\cos\left(\frac{2\pi m}{n}\right) \\
&\quad + i\sin\left(\frac{2\pi m}{n}\right)\cos\left(\frac{2\pi}{n}\right) - \sin\left(\frac{2\pi m}{n}\right)\sin\left(\frac{2\pi}{n}\right) \\
&= \left[\cos\left(\frac{2\pi m}{n}\right)\cos\left(\frac{2\pi}{n}\right) - \sin\left(\frac{2\pi m}{n}\right)\sin\left(\frac{2\pi}{n}\right)\right] \\
&\quad + i\left[\sin\left(\frac{2\pi}{n}\right)\cos\left(\frac{2\pi m}{n}\right)\right. \\
&\qquad \left. + \sin\left(\frac{2\pi m}{n}\right)\cos\left(\frac{2\pi}{n}\right)\right] \\
&= \cos\left(\frac{2\pi(m+1)}{n}\right) + i\sin\left(\frac{2\pi(m+1)}{n}\right).
\end{aligned}$$

□

Once we have Lemma 2.73, proving Theorem 2.72 is spectacularly easy.

Proof of Theorem 2.72. Substitution and the lemma give us

$$\begin{aligned}
\omega^n - 1 &= \left[\cos\left(\frac{2\pi n}{n}\right) + i\sin\left(\frac{2\pi n}{n}\right)\right] - 1 \\
&= \cos 2\pi + i\sin 2\pi - 1 \\
&= (1 + i\cdot 0) - 1 = 0,
\end{aligned}$$

so ω is indeed a root of $x^n - 1$. □

As promised, $\langle\omega\rangle$ gives us a nice group.

Theorem 2.74. The nth roots of unity are $\Omega_n = \{1, \omega, \omega^2, \ldots, \omega^{n-1}\}$, where ω is defined as in Theorem 2.72. They form a cyclic group of order n under multiplication.

The theorem does not claim merely that Ω_n is a list of *some* nth roots of unity; it claims that Ω_n is a list of *all* nth roots of unity. Our proof is going to cheat a little bit, because we don't quite have the machinery to prove that Ω_n is an exhaustive list of the roots of unity. We will eventually, however, and you should be able to follow the general idea now. The idea is called *unique factorization*. Basically, let f be a polynomial of degree n. Suppose that we have n roots of f; call them $\alpha_1, \alpha_2, \ldots, \alpha_n$. The parts you have to take on faith (for now) are twofold. First, $x - \alpha_i$ is a factor of f for each α_i. Each linear factor adds one to the degree of a polynomial, and f has degree n, so the number of linear factors cannot be more than n. Second, and this is not quite so clear, there is only one way to factor f into linear polynomials

(You can see this in the example above with $x^4 - 1$, but Theorem 7.45 on page 222 will have the details. You should have seen that theorem in your precalculus studies, and since it doesn't depend on anything in this section, the reasoning is not circular.)

If you're okay with that, then you're okay with everything else.

Proof. For $m \in \mathbb{N}^+$, we use the associative property of multiplication in $\mathbb{C}$ and the commutative property of multiplication in $\mathbb{N}^+$:

$$(\omega^m)^n - 1 = \omega^{mn} - 1 = \omega^{nm} - 1 = (\omega^n)^m - 1 = 1^m - 1 = 0.$$

Hence ω^m is a root of unity for any $m \in \mathbb{N}^+$. If $\omega^m = \omega^\ell$, then

$$\cos\left(\frac{2\pi m}{n}\right) = \cos\left(\frac{2\pi \ell}{n}\right) \quad \text{and} \quad \sin\left(\frac{2\pi m}{m}\right) = \sin\left(\frac{2\pi \ell}{n}\right),$$

and we know from trigonometry that this is possible only if

$$\begin{aligned} \frac{2\pi m}{n} &= \frac{2\pi \ell}{n} + 2\pi k \\ \frac{2\pi}{n}(m - \ell) &= 2\pi k \\ m - \ell &= kn. \end{aligned}$$

That is, $m - \ell$ is a multiple of n. Since Ω_n lists only those powers from 0 to $n-1$, the powers must be distinct, so Ω_n contains n distinct roots of unity. (See also Exercise 2.82.) As there can be at most n distinct roots, Ω_n is a complete list of nth roots of unity.

Now we show that Ω_n is a cyclic group.

(closure) Let $x, y \in \Omega_n$; you will show in Exercise 2.79 that $xy \in \Omega_n$.

(associativity) The complex numbers are associative under multiplication; since $\Omega_n \subseteq \mathbb{C}$, the elements of Ω_n are also associative under multiplication.

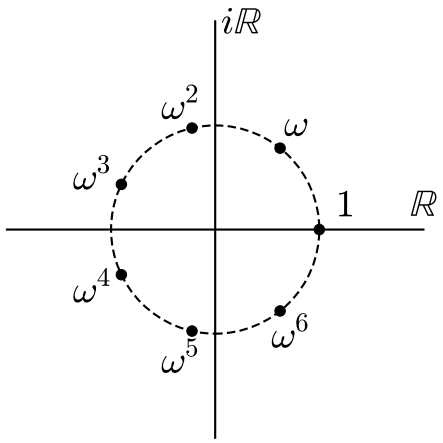


Figure 2.9. The seventh roots of unity, on the complex plane

(identity) The multiplicative identity in $\mathbb{C}$ is 1. This is certainly an element of Ω_n, since $1^n = 1$ for all $n \in \mathbb{N}^+$.

(inverses) Let $x \in \Omega_n$; you will show in Exercise 2.80 that $x^{-1} \in \Omega_n$.

(cyclic) Theorem 2.72 tells us that $\omega \in \Omega_n$; the remaining elements are powers of ω. Hence $\Omega_n = \langle \omega \rangle$.

□

Combined with the explanation we gave earlier of the complex plane, Theorem 2.74 gives us a wonderful symmetry for the roots of unity.

Example 2.75. We'll consider the case where $n = 7$. According to the theorem, the 7th roots of unity are $\Omega_7 = \{1, \omega, \omega^2, \ldots, \omega^6\}$ where

$$\omega = \cos\left(\frac{2\pi}{7}\right) + i \sin\left(\frac{2\pi}{7}\right).$$

According to Lemma 2.73,

$$\omega^m = \cos\left(\frac{2\pi m}{7}\right) + i \sin\left(\frac{2\pi m}{7}\right),$$

where $m = 0, 1, \ldots, 6$. By substitution, the angles we are looking at are

$$0, \frac{2\pi}{7}, \frac{4\pi}{7}, \frac{6\pi}{7}, \frac{8\pi}{7}, \frac{10\pi}{7}, \frac{12\pi}{7}.$$

Recall that in the complex plane, any complex number $a + bi$ corresponds to the point (a, b) on $\mathbb{R}^2$. The Pythagorean identity $\cos^2\alpha + \sin^2\alpha = 1$ tells us that the coordinates of the roots of unity lie on the unit circle. Since the angles are at equal intervals, they divide the unit circle into seven equal arcs! See Figure 2.9.

Although we used $n = 7$ in this example, we used no special properties of that number in the argument. That tells us that this property is true for any n: the nth roots of unity divide the unit

circle of the complex plane into n equal arcs!

Here's an interesting question: is ω is the only generator of Ω_n? In fact, no. A natural follow-up: are *all* the elements of Ω_n generators of the group? Likewise, no. Well, which ones are? We are not yet ready to give a precise criterion that signals which elements generate Ω_n, but they do have a special name.

Definition 2.76. We call any generator of Ω_n a **primitive nth root of unity**.

Exercises.

Unless stated otherwise, $n \in \mathbb{N}^+$ and ω is a primitive n-th root of unity.

Exercise 2.77. Show that $\mathbb{C}$ is a group under addition.

Exercise 2.78.

(a) Find all the primitive square roots of unity, all the primitive cube roots of unity, and all the primitive quartic (fourth) roots of unity.

(b) Sketch *all* the square roots of unity on a complex plane. (Not just the primitive ones, but all.) Repeat for the cube and quartic roots of unity, each on a separate plane.

(c) Are any cube roots of unity *not* primitive? what about quartic roots of unity?

Exercise 2.79.

(a) Suppose that a and b are both positive powers of ω. Adapt Lemma 2.73 to show that ab is also a power of ω.

(b) Explain why this shows that Ω_n is closed under multiplication.

Exercise 2.80.

(a) Let ω be a 14th root of unity; let $\alpha = \omega^5$, and $\beta = \omega^{14-5} = \omega^9$. Show that $\alpha\beta = 1$.

(b) More generally, let ω be a primitive n-th root of unity, Let $\alpha = \omega^a$, where $a \in \mathbb{N}$ and $a < n$. Show that $\beta = \omega^{n-a}$ satisfies $\alpha\beta = 1$.

(c) Explain why this shows that every element of Ω_n has an inverse.

Exercise 2.81. Suppose β is a root of $x^n - b$.

(a) Show that $\omega\beta$ is also a root of $x^n - b$, where ω is *any* nth root of unity.

(b) Use (a) and the idea of unique factorization that we described right before the proof of Theorem 2.74 to explain how we can use β and Ω_n to list all n roots of $x^n - b$.

Exercise 2.82.

(a) For each $\omega \in \Omega_6$, find $x, y \in \mathbb{R}$ such that $\omega = x + yi$. Plot all the points (x, y) on a graph.

(b) Do you notice any pattern to the points? If not, repeat part (a) for Ω_7, Ω_8, etc., until you see the pattern.

Exercise 2.83.

(a) Show that $\mathbb{C}$ satisfies the requirements of a vector space for scalar multiplication.

(b) Show that $\mathbb{C}$ and $\mathbb{R}^2$ are isomorphic as monoids under addition.

Exercise 2.84. Recall from Exercise 0.90 the set of quaternions $\{\pm\mathbf{1}, \pm\mathbf{i}, \pm\mathbf{j}, \pm\mathbf{k}\}$, where

$$\mathbf{1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mathbf{i} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix},$$
$$\mathbf{j} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \mathbf{k} = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}.$$

(a) Use the properties of these matrices that you proved earlier to build the Cayley table of Q_8. (In this case, the Cayley table is the multiplication table.)
(c) Show that Q_8 is a group under matrix multiplication.
(d) Explain why Q_8 is not an abelian group.

Exercise 2.85. In Exercise 2.84 you showed that the quaternions form a group under matrix multiplication. Verify that $H = \{\mathbf{1}, -\mathbf{1}, \mathbf{i}, -\mathbf{i}\}$ is a cyclic group. What elements generate H?

Exercise 2.86. Show that Q_8 is not cyclic.

Chapter 3:
Subgroups

A subset of a group is not necessarily a group; for example, $\{2,4\} \subset \mathbb{Z}$, but $\{2,4\}$ doesn't satisfy any properties of an additive group unless we change the definition of addition. Some subsets of groups are groups, and one of the keys to algebra consists in understanding the relationship between subgroups and groups.

We start this chapter by describing the properties that guarantee that a subset is a "subgroup" of a group (Section 3.1). We then explore how subgroups create *cosets*, equivalence classes within the group that perform a role similar to division of integers (Section 3.2). It turns out that in finite groups, we can count the number of these equivalence classes quite easily (Section 3.3).

Cosets open the door to a special class of groups called *quotient groups*, (Sections 3.4), one of which is a very natural, very useful tool (Section 3.5) that will eventually allow us to devise some "easy" solutions for problems in Number Theory (Chapter 6).

3.1: Subgroups

Definition 3.1. Let G be a group and $H \subseteq G$ be nonempty. If H is also a group under the same operation as G, then H is a **subgroup** of G. If $\{e\} \subsetneq H \subsetneq G$, then H is a **proper subgroup** of G.

Notation 3.2. If H is a subgroup of G, then we write $H < G$.

Example 3.3. Check that the following statements are true by verifying that the properties of a group are satisfied.

(a) $\mathbb{Z}$ is a subgroup of $\mathbb{Q}$.

(b) Let $4\mathbb{Z} := \{4m : m \in \mathbb{Z}\} = \{\ldots, -4, 0, 4, 8, \ldots\}$. Then $4\mathbb{Z}$ is a subgroup of $\mathbb{Z}$.

(c) Let $d \in \mathbb{Z}$ and $d\mathbb{Z} := \{dm : m \in \mathbb{Z}\}$. Then $d\mathbb{Z}$ is a subgroup of $\mathbb{Z}$.

(d) $\langle \mathbf{i} \rangle$ is a subgroup of Q_8.

Checking all four properties of a group is cumbersome. It would be convenient to verify that a set is a subgroup by checking fewer properties. It also makes sense that if a group is abelian, then its subgroups would be abelian, so we shouldn't have to check the abelian property. In that case, which properties *must* we check to decide whether a subset is a subgroup?

We can eliminate the associative and abelian properties from consideration. In fact, the operation remains associative and commutative for any sub*set*.

Lemma 3.4. Let G be a group and $H \subseteq G$. Then H satisfies the associative property of a group. In addition, if G is abelian, then H satisfies the commutative property of an abelian group. So, we only need to check the closure, identity, and inverse properties to ensure that G is a group.

Be careful: Lemma 3.4 neither assumes nor concludes that H is a subgroup. The other three properties may not be satisfied: H may not be closed; it may lack an identity; or some element may

lack an inverse. The lemma merely states that any subset automatically satisfies two important properties of a group.

Proof. If $H = \emptyset$, then the lemma is true trivially.

Otherwise, $H \neq \emptyset$. Let $a, b, c \in H$. Since $H \subseteq G$, we have $a, b, c \in G$. Since the operation is associative in G, $a(bc) = (ab)c$; that is, the operation remains associative for H. Likewise, if G is abelian, then $ab = ba$; that is, the operation also remains commutative for H. □

Lemma 3.4 has reduced the number of requirements for a subgroup from four to three. Amazingly, we can simplify this further, to *only one criterion*.

Theorem 3.5 (The Subgroup Theorem)**.** Let $H \subseteq G$ be nonempty. The following are equivalent:
(A) $H < G$;
(B) for every $x, y \in H$, we have $xy^{-1} \in H$.

Notation 3.6. If G were an additive group, we would write $x - y$ instead of xy^{-1}.

Proof. By Exercise 2.33 on page 65, (A) implies (B).

Conversely, assume (B). By Lemma 3.4, we need to show only that H satisfies the closure, identity, and inverse properties. We do this slightly out of order:

identity: Let $x \in H$. By (B), $e = x \cdot x^{-1} \in H$.[9]
inverse: Let $x \in H$. Since H satisfies the identity property, $e \in H$. By (B), $x^{-1} = e \cdot x^{-1} \in H$.
closure: Let $x, y \in H$. Since H satisfies the inverse property, $y^{-1} \in H$. By (B), $xy = x \cdot \left(y^{-1}\right)^{-1} \in H$.

Since H satisfies the closure, identity, and inverse properties, $H < G$. □

Let's take a look at the Subgroup Theorem in action.

Example 3.7. Let $d \in \mathbb{Z}$. We claim that $d\mathbb{Z} < \mathbb{Z}$. (Here $d\mathbb{Z}$ is the set defined in Example 3.3.) *Why?* Let's use the Subgroup Theorem.

Let $x, y \in d\mathbb{Z}$. If we can show that $x - y \in d\mathbb{Z}$, we will satisfy part (B) of the Subgroup Theorem. The theorem states that (B) is equivalent to (A); that is, $d\mathbb{Z}$ is a group. That's what we want, so let's try to show that $x - y \in d\mathbb{Z}$; that is, $x - y$ is an integer multiple of d.

Since x and y are by definition integer multiples of d, we can write $x = dm$ and $y = dn$ for some $m, n \in \mathbb{Z}$. Note that $-y = -(dn) = d(-n)$. Then

$$\begin{aligned} x - y = x + (-y) &= dm + d(-n) \\ &= d(m + (-n)) = d(m - n). \end{aligned}$$

Now, $m - n \in \mathbb{Z}$, so $x - y = d(m - n) \in d\mathbb{Z}$.

We did it! We took two integer multiples of d, and showed that their difference is also an integer multiple of d. By the Subgroup Theorem, $d\mathbb{Z} < \mathbb{Z}$.

The following geometric example gives a visual image of what a subgroup "looks" like.

[9] Notice that here we are replacing the y in (B) with x. This is fine, since nothing in (B) requires x and y to be distinct.

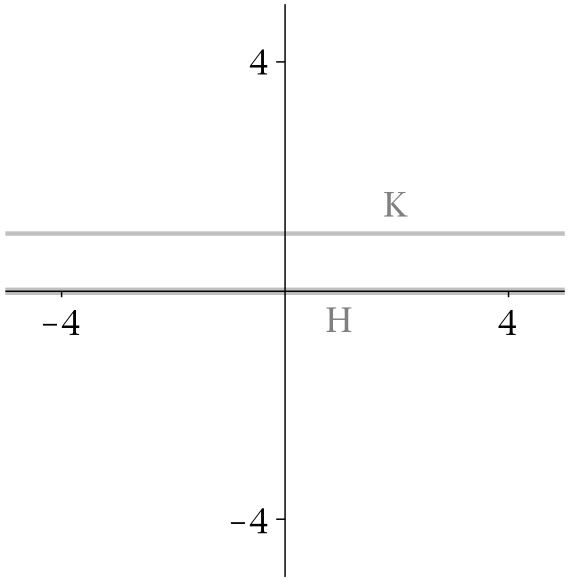


Figure 3.1. H and K from Example 3.8

Example 3.8. Recall that $\mathbb{R}$is a group under addition, and let G be the direct product $\mathbb{R}\times\mathbb{R}$. Geometrically, this is the set of points in the x-y plane. As is usual with a direct product, we define an addition for elements of G in the natural way: for $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$, define

$$P_1 + P_2 = (x_1 + x_2, y_1 + y_2)\,.$$

Let H be the x-axis; a set definition would be, $H = \{x \in G : x = (a, 0)\ \exists a \in \mathbb{R}\}$. We claim that $H < G$. *Why?* Use the Subgroup Theorem! Let $P, Q \in H$. By the definition of H, we can write $P = (p, 0)$ and $Q = (q, 0)$ where $p, q \in \mathbb{R}$. Then

$$P - Q = P + (-Q) = (p, 0) + (-q, 0) = (p - q, 0)\,.$$

Membership in H requires the first ordinate to be real, and the second to be zero. As $P - Q$ satisfies these requirements, $P - Q \in H$. The Subgroup Theorem implies that $H < G$.

Let K be the line $y = 1$; a set definition would be, $K = \{x \in G : x = (a, 1)\ \exists a \in \mathbb{R}\}$. We claim that $K \not< G$. *Why not?* Again, use the Subgroup Theorem! Let $P, Q \in K$. By the definition of K, we can write $P = (p, 1)$ and $Q = (q, 1)$ where $p, q \in \mathbb{R}$. Then

$$P - Q = P + (-Q) = (p, 1) + (-q, -1) = (p - q, 0)\,.$$

Membership in K requires the second ordinate to be one, but the second ordinate of $P - Q$ is zero, not one. Since $P - Q \notin K$, the Subgroup Theorem tells us that K is not a subgroup of G.

There's a more intuitive explanation as to why K is not a subgroup; it doesn't contain the origin. In a direct product of groups, the identity is formed using the identities of the component groups. In this case, the identity is $(0, 0)$, which is *not* in K.

Figure 3.1 gives a visualization of H and K. You will diagram another subgroup of G in Exercise 3.16.

Examples 3.7 and 3.8 give us examples of how the Subgroup Theorem verifies subgroups of *abelian* groups. Two interesting examples of nonabelian subgroups appear in D_3.

Example 3.9. Recall D_3 from Section 2.2. Both $H = \{\iota, \varphi\}$ and $K = \{\iota, \rho, \rho^2\}$ are subgroups of D_3. *Why?* Certainly $H, K \subsetneq G$, and Theorem 2.55 on page 78 tells us that H and K are groups, since $H = \langle\varphi\rangle$, and $K = \langle\rho\rangle$.

If a group satisfies a given property, a natural question to ask is whether its subgroups also satisfy this property. Cyclic groups are a good example: is every subgroup of a cyclic group also cyclic? The answer relies on the Division Theorem (Theorem 0.34 on page 13).

Theorem 3.10. Subgroups of cyclic groups are also cyclic.

Proof. Let G be a cyclic group, and $H < G$. From the fact that G is cyclic, choose $g \in G$ such that $G = \langle g\rangle$.

First we must find a candidate generator of H. If $H = \{e\}$, then $H = \langle e\rangle = \langle g^0\rangle$, and we are done. So assume there exists $x \in H$ such that $x \neq e$. By inclusion, every element $x \in H$ is also an element of G, which is generated by g, so $x = g^n$ for some $n \in \mathbb{Z}$. Without loss of generality, we may assume that $n \in \mathbb{N}^+$; after all, we just showed that we can choose $x \neq e$, so $n \neq 0$, and if $n \notin \mathbb{N}$, then closure of H implies that $x^{-1} = g^{-n} \in H$, so choose x^{-1} instead.

Now, if you were to take all the positive powers of g that appear in H, which would you expect to generate H? Certainly not the larger ones! The ideal candidate for the generator would be the smallest positive power of g in H, if it exists. Let S be the set of positive natural numbers i such that $g^i \in H$; in other words, $S = \left\{i \in \mathbb{N}^+ : g^i \in H\right\}$. From the well-ordering of $\mathbb{N}$, there exists a smallest element of S; call it d, and assign $h = g^d$.

We claim that $H = \langle h\rangle$. Let $x \in H$; then $x \in G$. By hypothesis, G is cyclic, so $x = g^a$ for some $a \in \mathbb{Z}$. By the Division Theorem, we know that there exist unique $q, r \in \mathbb{Z}$ such that

- $a = qd + r$, and
- $0 \leq r < d$.

Let $y = g^r$; by Exercise 2.61, we can rewrite this as

$$y = g^r = g^{a-qd} = g^a g^{-(qd)} = x \cdot \left(g^d\right)^{-q} = x \cdot h^{-q}.$$

Now, $x \in H$ by definition, and $h^{-q} \in H$ by closure and the existence of inverses, so by closure $y = x \cdot h^{-q} \in H$ as well. We chose d as the smallest positive power of g in H, and we just showed that $g^r \in H$. Recall that $0 \leq r < d$. If $0 < r$; then $g^r \in H$, so $r \in S$. But $r < d$, which contradicts the choice of d as the smallest element of S. Hence r cannot be positive; instead, $r = 0$ and $x = g^a = g^{qd} = h^q \in \langle h\rangle$.

Since x was arbitrary in H, every element of H is in $\langle h\rangle$; that is, $H \subseteq \langle h\rangle$. Since $h \in H$ and H is a group, closure implies that $H \supseteq \langle h\rangle$, so $H = \langle h\rangle$. In other words, H is cyclic. □

We again look to $\mathbb{Z}$ for an example.

Example 3.11. Recall from Example 2.53 on page 77 that $\mathbb{Z}$ is cyclic; in fact $\mathbb{Z} = \langle 1\rangle$. By Theorem 3.10, $d\mathbb{Z}$ is cyclic. In fact, $d\mathbb{Z} = \langle d\rangle$. Can you find another generator of $d\mathbb{Z}$?

Exercises.

Let G be any group and $g \in G$.
Claim: $\langle g \rangle < G$.
Proof:

1. Let $x, y \in$_____.
2. By definition of _____, there exist $m, n \in \mathbb{Z}$ such that $x = g^m$ and $y = g^n$.
3. By _____, $y^{-1} = g^{-n}$.
4. By _____, $xy^{-1} = g^{m+(-n)} = g^{m-n}$.
5. By _____, $xy^{-1} \in \langle g \rangle$.
6. By _____, $\langle g \rangle < G$.

Figure 3.2. Material for Exercise 3.14

Exercise 3.12. Recall that Ω_n, the nth roots of unity, is the cyclic group $\langle \omega \rangle$.
(a) Compute Ω_2 and Ω_4, and explain why $\Omega_2 < \Omega_4$.
(b) Compute Ω_8, and explain why both $\Omega_2 < \Omega_8$ and $\Omega_4 < \Omega_8$.
(b) Explain why, if $d \mid n$, then $\Omega_d < \Omega_n$.

Exercise 3.13. Show that even though the Klein 4-group is not cyclic, each of its proper subgroups is cyclic (see Exercises 2.32 on page 65 and 2.64 on page 83).

Exercise 3.14.
(a) Fill in each blank of Figure 3.2 with the appropriate justification or expression.
(b) Why would someone take this approach, rather than using the definition of a subgroup?

Exercise 3.15.

(a) Let $D_n(\mathbb{R}) = \{aI_n : a \in \mathbb{R}\} \subseteq \mathbb{R}^{n \times n}$; that is, $D_n(\mathbb{R})$ is the set of all diagonal matrices whose values along the diagonal is constant. Show that $D_n(\mathbb{R}) < \mathbb{R}^{n \times n}$. (In case you've forgotten Exercise 2.27, the operation here is addition.)
(b) Let $D_n^*(\mathbb{R}) = \{aI_n : a \in \mathbb{R} \setminus \{0\}\} \subseteq \mathrm{GL}_n(\mathbb{R})$; that is, $D_n^*(\mathbb{R})$ is the set of all non-zero diagonal matrices whose values along the diagonal is constant. Show that $D_n^*(\mathbb{R}) < \mathrm{GL}_n(\mathbb{R})$. (In case you've forgotten Definition 2.5, the operation here is multiplication.)

Exercise 3.16. Let $G = \mathbb{R}^2 := \mathbb{R} \times \mathbb{R}$, with addition defined as in Exercise 2.25 and Example 3.8. Let

$$L = \{x \in G : \ x = (a, a) \ \exists a \in \mathbb{R}\}.$$

(a) Describe L geometrically.
(b) Show that $L < G$.
(c) Suppose $\ell \subseteq G$ is any line. Identify the simplest criterion possible that decides whether $\ell < G$. Justify your answer.

Exercise 3.17. Let G be an abelian group. Let H, K be subgroups of G. Let

$$H + K = \{x + y : \ x \in H, y \in K\}.$$

Show that $H + K < G$.

Exercise 3.18. Let $H = \{\iota, \varphi\} < D_3$.

Let G be a group and $A_1, A_2, \ldots, A_m$ subgroups of G. Let

$$B = A_1 \cap A_2 \cap \cdots \cap A_m.$$

Claim: $B < G$.
Proof:

1. Let $x, y \in$_____.
2. By _____, $x, y \in A_i$ for all $i = 1, \ldots, m$.
3. By _____, $xy^{-1} \in A_i$ for all $i = 1, \ldots, m$.
4. By _____, $xy^{-1} \in B$.
5. By _____, $B < G$.

Figure 3.3. Material for Exercise 3.20

(a) Find a different subgroup K of D_3 with only two elements.
(b) Let $HK = \{xy : x \in H, y \in K\}$. Show that $HK \not< D_3$.
(c) Why does the result of (b) not contradict the result of Exercise 3.17?

Exercise 3.19. Explain why $\mathbb{R}$ cannot be cyclic.

Exercise 3.20. Fill each blank of Figure 3.3 with the appropriate justification or expression.

Exercise 3.21. Let G be a group and H, K two subgroups of G. Let $A = H \cup K$. Show that A *need not* be a subgroup of G.

Exercise 3.22. Recall the set of orthogonal matrices from Exercise 0.91.
(a) Show that $\mathrm{O}(n) < \mathrm{GL}(n)$. We call $O(n)$ the **orthogonal group**.

Let $\mathrm{SO}(n)$ be the set of all orthogonal $n \times n$ matrices whose determinant is 1. We call $\mathrm{SO}(n)$ the **special orthogonal group**.
(b) Show that $\mathrm{SO}(n) < \mathrm{O}(n)$.

3.2: Cosets

One of the most powerful tools in group theory is that of cosets. Students often have a hard time wrapping their minds around cosets, so we'll start with an introductory example that should give you an idea of how cosets "look" in a group. Then we'll define cosets, and finally look at some of their properties.

The idea

Recall the illustration of how the Division Theorem partitions the integers according to their remainder (Section 0.2). Two aspects of division were critical for this:

- *existence of a remainder*, which implies that every integer belongs to at least one class, which in turn implies that the union of the classes covers $\mathbb{Z}$; and
- *uniqueness of the remainder*, which implies that every integer ends up in only one set, so that the classes are disjoint.

Using the vocabulary of groups, recall that $A = 4\mathbb{Z} < \mathbb{Z}$ (page 94). All the elements of B have the form $1 + a$ for some $a \in A$. For example, $-3 = 1 + (-4)$. Likewise, all the elements of C have

the form $2+a$ for some $a \in A$, and all the elements of D have the form $3+a$ for some $a \in A$. So if we define

$$1+A := \{1+a : a \in A\},$$

then

$$\begin{aligned} 1+A &= \{\ldots, 1+(-4), 1+0, 1+4, 1+8, \ldots\} \\ &= \{\ldots, -3, 1, 5, 9, \ldots\} \\ &= B. \end{aligned}$$

Likewise, we can write $A = 0+A$ and $C = 2+A$, $D = 3+A$.

Pursuing this further, you can check that

$$\cdots = -3+A = 1+A = 5+A = 9+A = \cdots$$

and so forth. Interestingly, all the sets in the previous line are the same as B! In addition, $1+A = 5+A$, and $1-5 = -4 \in A$. The same holds for C: $2+A = 10+A$, and $2-10 = -8 \in A$. This relationship will prove important at the end of the section.

So the partition by remainders of division by four is related to the subgroup A of multiples of 4. This will become very important in Chapter 6. How can we generalize this phenomen to other groups, even nonabelian ones?

Definition 3.23. Let G be a group and $A < G$. Let $g \in G$. We define the **left coset of A with g** as

$$gA = \{ga : a \in A\}$$

and the **right coset of A with g** as

$$Ag = \{ag : a \in A\}.$$

As usual, if A is an additive subgroup, we write the left and right cosets of A with g as $g+A$ and $A+g$.

In general, left cosets and right cosets are not equal, partly because the operation might not commute. If we speak of "cosets" without specifying "left" or "right", we means "left cosets".

Example 3.24. Recall the group D_3 from Section 2.2 and the subgroup $H = \langle\varphi\rangle = \{\iota, \varphi\}$ from Example 3.9. In this case,

$$\rho H = \{\rho, \rho\varphi\} \text{ and } H\rho = \{\rho, \varphi\rho\}.$$

Since $\varphi\rho = \rho^2\varphi \neq \rho\varphi$, we see that $\rho H \neq H\rho$.

Sometimes, the left coset and the right coset *are* equal. This is always true in abelian groups, as illustrated by Example 3.25.

Example 3.25. Consider the subgroup $H = \{(a, 0) : a \in \mathbb{R}\}$ of $\mathbb{R}^2$ from Exercise 3.16. Let $p =$

$(3,-1) \in \mathbb{R}^2$. The coset of H with p is

$$\begin{aligned} p+H &= \{(3,-1)+q : q \in H\} \\ &= \{(3,-1)+(a,0) : a \in \mathbb{R}\} \\ &= \{(3+a,-1) : a \in \mathbb{R}\}. \end{aligned}$$

Sketch some of the points in $p+H$, and compare them to your sketch of H in Exercise 3.16. How does the coset compare to the subgroup?

Generalizing this further, every coset of H has the form $p+H$ where $p \in \mathbb{R}^2$. Elements of $\mathbb{R}^2$ are points, so $p=(x,y)$ for some $x,y \in \mathbb{R}$. The coset of H with p is

$$p+H = \{(x+a,y) : a \in \mathbb{R}\}.$$

Sketch several more cosets. How would you describe the set of *all* cosets of H in $\mathbb{R}^2$?

The group does not *have* to be abelian in order to have the left and right cosets equal. When deciding if $gA = Ag$, we are not deciding *whether elements of G commute*, but *whether subsets of G are equal*. Returning to D_3, we can find a subgroup whose left and right cosets are equal even though the group is not abelian and the operation is not commutative.

Example 3.26. Let $K = \{\iota, \rho, \rho^2\}$; certainly $K < D_3$, after all, $K = \langle \rho \rangle$. In this case, $\alpha K = K\alpha$ for all $\alpha \in D_3$:

α	αK	$K\alpha$
ι	K	K
φ	$\{\varphi, \varphi\rho, \varphi\rho^2\}$	$\{\varphi, \rho\varphi, \rho^2\varphi\}$
ρ	K	K
ρ^2	K	K
$\rho\varphi$	$\{\rho\varphi, (\rho\varphi)\rho, (\rho\varphi)\rho^2\}$	$\{\rho\varphi, \varphi, \rho^2\varphi\}$
$\rho^2\varphi$	$\{\rho^2\varphi, (\rho^2\varphi)\rho, (\rho^2\varphi)\rho^2\}$	$\{\rho^2\varphi, \rho\varphi, \varphi\}$

In each case, the sets φK and $K\varphi$ are equal, even though φ does not commute with ρ. (You should verify these computations by hand.)

Properties of Cosets

We could forgive you for concluding from this that cosets are useful for little more than a generalization of division; after all, you don't realize how powerful division is. The rest of this chapter should correct any such misapprehension; for now, we present some properties of cosets that illustrate further their similarities to division.

Theorem 3.27. The cosets of a subgroup partition the group.

Putting this together with Theorem 0.42 implies another nice result.

Corollary 3.28. Let $A < G$. Define a relation $\sim$ on $x, y \in G$ by

$$x \sim y \iff x \text{ is in the same coset of } A \text{ as } y.$$

This relation is an equivalence relation.

We will make use of this result, in due course.

Proof of Theorem 3.27. Let G be a group, and $A < G$. We have to show two things:

(CP1) the cosets of A cover G, and
(CP2) distinct cosets of A are disjoint.

We show (CP1) first. Let $g \in G$. The definition of a group tells us that $g = ge$. Since $e \in A$ by definition of subgroup, $g = ge \in gA$. Since g was arbitrary, every element of G is in some coset of A. Hence the union of all the cosets is G.

For (CP2), let X and Y be arbitrary cosets of A. Assume that X and Y are distinct; that is, $X \neq Y$. We need to show that they are disjoint; that is, $X \cap Y = \emptyset$. By way of contradiction, assume that $X \neq Y$ but $X \cap Y \neq \emptyset$. Since $X \neq Y$, one of the two cosets contains an element that does not appear in the other; without loss of generality, assume that $z \in X$ but $z \notin X$. By definition, there exist $x, y \in G$ such that $X = xA$ and $Y = yA$; we can write $z = xa$ for some $a \in A$. Since $X \cap Y \neq \emptyset$, there exists some $w \in X \cap Y$; by definition, we can find $b, c \in A$ such that $w = xb = yc$. Solve this last equation for x, and we have $x = (yc)\, b^{-1}$. Substitute this into the equation for z, and we have

$$z = xa = \left[(yc)\, b^{-1}\right] a \underset{\text{ass.}}{=} y\left(cb^{-1}a\right).$$

Since A is a subgroup, hence a group, it is closed under inverses and multiplication, so $cb^{-1}a \in A$. But then $z = y\left(cb^{-1}a\right) \in yA$, which contradicts the choice of z! The assumption that we could find distinct cosets that are not disjoint must have been false, and since X and Y were arbitrary, this holds for all cosets of A.

Having shown (CP2) and (CP1), we have shown that the cosets of A partition G. □

We conclude this section with three facts that allow us to decide when cosets are equal.

Lemma 3.29 (Equality of cosets)**.** Let G be a group and $H < G$. All of the following hold:
(CE1) $eH = H$.
(CE2) For all $a \in G$, $a \in H$ iff $aH = H$.
(CE3) For all $a, b \in G$, $aH = bH$ if and only if $a^{-1}b \in H$.

As usual, you should keep in mind that in additive groups these conditions translate to

(CE1) $0 + H = H$.
(CE2) For all $a \in G$, if $a \in H$ then $a + H = H$.
(CE3) For all $a, b \in G$, $a + H = b + H$ if and only if $a - b \in H$.

Proof. We only sketch the proof here. You will fill in the details in Exercise 3.36. Remember that part of this problem involves proving that two sets are equal, and to prove that, you should prove that each is a subset of the other.

(CE1) is "obvious" (but fill in the details anyway).

We'll skip (CE2) for the moment, and move to (CE3). Since (CE3) is also an equivalence, we have to prove two directions. Let $a, b \in G$. First, assume that $aH = bH$. By the identity property, $e \in H$, so $b = be \in bH$. Hence, $b \in aH$; that is, we can find $h \in H$ such that $b = ah$. By substitution and the properties of a group, $a^{-1}b = a^{-1}(ah) = h$, so $a^{-1}b \in H$.

Conversely, assume that $a^{-1}b \in H$. We must show that $aH = bH$, which requires us to show that $aH \subseteq bH$ and $aH \supseteq bH$. Since $a^{-1}b \in H$, we have

$$b = a\left(a^{-1}b\right) \in aH.$$

We can thus write $b = ah$ for some $h \in H$. Let $y \in bH$; then $y = b\widehat{h}$ for some $\widehat{h} \in H$, and we have $y = (ah)\,\widehat{h} \in H$. Since y was arbitrary in bH, we now have $aH \supseteq bH$.

Although we could build a similar argument to show that $aH \subseteq bH$, instead we point out that $aH \supseteq bH$ implies that $aH \cap bH \neq \emptyset$. The cosets are not disjoint, so by Theorem 3.27, they are not distinct: $aH = bH$.

Now we turn to (CE2). Let $a \in G$, and assume $a \in H$. By the inverse property, $a^{-1} \in H$. We know that $e \in H$, so by closure, $a^{-1}e \in H$. We can now use (CE3) and (CE1) to determine that $aH = eH = H$. □

Exercises.

Exercise 3.30. Show explicitly why left and right cosets are equal in abelian groups.

Exercise 3.31. In Exercise 3.12, you showed that $\Omega_2 < \Omega_8$. Compute the left and right cosets of Ω_2 in Ω_8.

Exercise 3.32. Let $\{e, a, b, a+b\}$ be the Klein 4-group. (See Exercises 2.32 on page 65, 2.64 on page 83, and 3.13 on page 98.) Compute the cosets of $\langle a \rangle$.

Exercise 3.33. In Exercise 3.18 on page 98, you found another subgroup K of order 2 in D_3. Does K satisfy the property $\alpha K = K\alpha$ for all $\alpha \in D_3$?

Exercise 3.34. Recall the subgroup L of $\mathbb{R}^2$ from Exercise 3.16 on page 98.

(a) Give a geometric interpretation of the coset $(3,-1) + L$.
(b) Give an algebraic expression that describes $p + L$, for arbitrary $p \in \mathbb{R}^2$.
(c) Give a geometric interpretation of the cosets of L in $\mathbb{R}^2$.
(d) Use your geometric interpretation of the cosets of L in $\mathbb{R}^2$ to explain why the cosets of L partition $\mathbb{R}^2$.

Exercise 3.35. Recall $D_n(\mathbb{R})$ from Exercise 3.15 on page 98. Give a description in set notation for

$$\begin{pmatrix} 0 & 3 \\ 0 & 0 \end{pmatrix} + D_2(\mathbb{R}).$$

List some elements of the coset.

Exercise 3.36.

(a) Fill in each blank of Figure 3.4 with the appropriate justification or statement.

3.3: Lagrange's Theorem

Let G be a group and $H < G$.
Claim: $eH = H$.

1. First we show that _____. Let $x \in eH$.
   (a) By definition, _____.
   (b) By the identity property, _____.
   (c) By definition, _____.
   (d) We had chosen an arbitrary element of eH , so by inclusion, _____.
2. Now we show the converse. Let _____.
   (a) By the identity property, _____.
   (b) By definition, _____$\in eH$.
   (c) We had chosen an arbitrary element, so by inclusion, _____.

Figure 3.4. Material for Exercise 3.36

This section introduces an important result describing the number of cosets a subgroup can have. This leads to some properties regarding the order of a group and any of its elements.

Notation 3.37. Let G be a group, and $A < G$. We write G/A for the set of all left cosets of A. That is,

$$G/A = \{gA : g \in G\}.$$

We also write $A\backslash G$ for the set of all right cosets of A:

$$A\backslash G = \{Ag : g \in G\}.$$

Example 3.38. Let $G = \mathbb{Z}$ and $A = 4\mathbb{Z}$. We saw in Example 0.40 that

$$G/A = \mathbb{Z}/4\mathbb{Z} = \{A, 1+A, 2+A, 3+A\}.$$

We actually "waved our hands" in Example 0.40. That means that we did not provide a very detailed argument, so let's show the details here. Recall that $4\mathbb{Z}$ is the set of multiples of $\mathbb{Z}$, so $x \in A$ iff x is a multiple of 4. What about the remaining elements of $\mathbb{Z}$?

Let $x \in \mathbb{Z}$; then

$$x + A = \{x + z : z \in A\} = \{x + 4n : n \in \mathbb{Z}\}.$$

Use the Division Theorem to write

$$x = 4q + r$$

for unique $q, r \in \mathbb{Z}$, where $0 \leq r < 4$. Then

$$x + A = \{(4q + r) + 4n : n \in \mathbb{Z}\} = \{r + 4(q + n) : n \in \mathbb{Z}\}.$$

By closure, $q + n \in \mathbb{Z}$. If we write m in place of $4(q + n)$, then $m \in 4\mathbb{Z}$. So

$$x + A = \{r + m : m \in 4\mathbb{Z}\} = r + 4\mathbb{Z}.$$

The distinct cosets of A are thus determined by the distinct remainders from division by 4. Since

the remainders from division by 4 are 0, 1, 2, and 3, we conclude that

$$\mathbb{Z}/A = \{A, 1+A, 2+A, 3+A\}$$

as claimed above.

Example 3.39. Let $G = D_3$ and $K = \{\iota, \rho, \rho^2\}$ as in Example 3.26, then

$$G/K = D_3/\langle\rho\rangle = \{K, \varphi K\}.$$

Example 3.40. Let $H < \mathbb{R}^2$ be as in Example 3.8 on page 95; that is,

$$H = \{(a, 0) \in \mathbb{R}^2 : a \in \mathbb{R}\}.$$

Then

$$\mathbb{R}^2/H = \{r + H : r \in \mathbb{R}^2\}.$$

It is not possible to list all the elements of G/H, but some examples would be

$$(1,1) + H,\ (4,-2) + H.$$

Here's a question for you to think about. Speaking *geometrically*, what do the elements of G/H look like? This question is similar to Exercise 3.34.

It is important to keep in mind that G/A is a set whose elements are also sets. As a result, showing equality of two elements of G/A requires one to show that two sets are equal.

When G is finite, a simple formula gives us the size of G/A.

Theorem 3.41 (Lagrange's Theorem)**.** Let G be a group of finite order, and $A < G$. Then

$$|G/A| = \frac{|G|}{|A|}.$$

Lagrange's Theorem states that the number of elements in G/A is the same as the quotient of the order of G by the order of A. The notation of cosets is somewhat suggestive of the relationship we illustrated at the begining of Section 3.2 between cosets and division of the integers. Nevertheless, Lagrange's Theorem is *not* as obvious as the notation might imply: we can't "divide" the sets G and A. We are not moving the absolute value bars "inside" the fraction; nor can we, as G/A is not a number. Rather, we are dividing, or partitioning, if you will, the group G by by the cosets of its subgroup A, obtaining the set of cosets G/A.

Proof. From Theorem 3.27 we know that the cosets of A partition G. How many such cosets are there? $|G/A|$, by definition! Each coset has the same size, $|A|$. A basic principle of counting tells us that the number of elements of G is thus the product of the number of elements in each coset and the number of cosets. That is, $|G/A| \cdot |A| = |G|$. This implies the theorem. □

The next-to-last sentence of the proof contains the statement $|G/A| \cdot |A| = |G|$. Since $|A|$ is the order of the group A, and $|G/A|$ is an integer, we conclude that:

Claim: The order of an element of a group divides the order of a group.
Proof:

1. Let G _____.
2. Let x _____.
3. Let $H = \langle$_____$\rangle$.
4. By _____, every integer power of x is in G.
5. By _____, H is the set of integer powers of x.
6. By _____, $H < G$.
7. By _____, $|H|$ divides $|G|$.
8. By _____, $\text{ord}(x)$ divides $|H|$.
9. By definition, there exist $m, n \in$_____ such that $|H| = m\,\text{ord}(x)$ and $|G| = n|H|$.
10. By substitution, $|G| =$_____.
11. _____.
    (This last statement must include a justification.)

Figure 3.5. Material for Exercise 3.46

Corollary 3.42. The order of a subgroup divides the order of a group.

Example 3.43. Let G be the Klein 4-group (see Exercises 2.32 on page 65, 2.64 on page 83, and 3.13 on page 98). Every subgroup of the Klein 4-group has order 1, 2, or 4. As predicted by Corollary 3.42, the orders of the subgroups divide the order of the group.

Likewise, the order of $\{\iota, \varphi\}$ divides the order of D_3.

By contrast, the subset HK of D_3 that you computed in Exercise 3.18 on page 98 has four elements. Since $4 \nmid 6$, the contrapositive of Lagrange's Theorem implies that HK *cannot* be a subgroup of D_3.

From the fact that every element g generates a cyclic subgroup $\langle g \rangle < G$, Lagrange's Theorem also implies an important consequence about the order of any element of any finite group.

Corollary 3.44. In a finite group G, the order of any element divides the order of a group.

Proof. You do it! See Exercise 3.46. □

Exercises.

Exercise 3.45. Recall from Exercise 3.12 that if $d \mid n$, then $\Omega_d < \Omega_n$. How many cosets of Ω_d are there in Ω_n?

Exercise 3.46. Fill in each blank of Figure 3.5 with the appropriate justification or expression.

Exercise 3.47. Suppose that a group G has order 8, but is not cyclic. Show that $g^4 = e$ for all $g \in G$.

Exercise 3.48. Let G be a group, and $g \in G$. Show that $g^{|G|} = e$.

Exercise 3.49. Suppose that a group has five elements. Why *must* it be abelian?

Exercise 3.50. Find a sufficient (but not necessary) condition on the order of a group of order at least two that guarantees that the group is cyclic.

3.4: Quotient Groups

Let $A < G$. Is there a natural generalization of the operation of G that makes G/A a group? By a "natural" generalization, we mean something like

$$(gA)(hA) = (gh)A.$$

"Normal" subgroups

The first order of business it to make sure that the operation even makes sense. The technical word for this is that the operation is **well-defined**. *What does that mean?* A coset can have different representations. An operation must be a function: for every pair of elements, it must produce *exactly one* result. The relation above would not be an operation if different representations of a coset gave us different answers. Example 3.51 shows how it can go wrong.

Example 3.51. Recall $H = \langle\varphi\rangle < D_3$ from Example 3.24. Let $X = \rho H$ and $Y = \rho^2 H$. Notice that $(\rho\varphi)H = \{\rho\varphi, \iota\} = \rho H$, so X has two representations, ρH and $(\rho\varphi)H$.

Were the operation well-defined, XY would have the same value, *regardless of the representation of* X. That is not the case! When we use the the first representation,

$$XY = (\rho H)\left(\rho^2 H\right) = \left(\rho \circ \rho^2\right)H = \rho^3 H = \iota H = H.$$

When we use the second representation,

$$\begin{aligned} XY = ((\rho\varphi)H)\left(\rho^2 H\right) &= \left((\rho\varphi)\rho^2\right)H = \left(\rho\left(\varphi\rho^2\right)\right)H \\ &= (\rho(\rho\varphi))H = \left(\rho^2\varphi\right)H \neq H. \end{aligned}$$

On the other hand, sometimes the operation *is* well-defined.

Example 3.52. Recall the subgroup $A = 4\mathbb{Z}$ of $\mathbb{Z}$. Let $B, C, D \in \mathbb{Z}/A$, so $B = b + 4\mathbb{Z}$, $C = c + 4\mathbb{Z}$, and $D = d + 4\mathbb{Z}$ for some $b, c, d \in \mathbb{Z}$.

We have to make sure that we cannot have $B = D$ and $B + C \neq D + C$. For example, if $B = 1 + 4\mathbb{Z}$ and $D = 5 + 4\mathbb{Z}$, $B = D$. Does it follow that $B + C = D + C$?

From Lemma 3.29, we know that $B = D$ iff $b - d \in A = 4\mathbb{Z}$. That is, $b - d = 4m$ for some $m \in \mathbb{Z}$. Let $x \in B + C$; then $x = (b + c) + 4n$ for some $n \in \mathbb{Z}$. By substitution,

$$x = ((d + 4m) + c) + 4n = (d + c) + 4(m + n) \in D + C.$$

Since x was arbitrary in $B + C$, we have $B + C \subseteq D + C$. A similar argument shows that $B + C \supseteq D + C$, so the two are, in fact, equal.

The operation was well-defined in the second example, but not the first. What made for the difference? In the second example, we rewrote

$$((d+4m)+c)+4n=(d+c)+4(m+n),$$

but that relies on the fact that *addition commutes in an abelian group*. Without that fact, we could not have swapped c and $4m$.

Does that mean we cannot make a group out of cosets of nonabelian groups? Not quite. The key in Example 3.52 was not that $\mathbb{Z}$ is abelian, but that we could rewrite $(4m+c)+4n$ as $c+(4m+4n)$, then simplify $4m+4n$ to $4(m+n)$. The abelian property makes it easy to do that, but we don't need the *group* G to be abelian; we need the *subgroup* A to satisfy it. If A were not abelian, we could still make it work if, after we move c left, we get *some* element of A to its right, so that it can be combined with the other one. That is, we have to be able to rewrite any ac as ca', where a' is also in A. We *need not have* $a=a'$*!* Let's emphasize that, changing c to g for an arbitrary group G:

The operation defined above is well-defined
iff
for every $g\in G$ *and for every* $a\in A$
there exists $a'\in A$ *such that* $ga=a'g$.

Think about this in terms of sets: for every $g\in G$ and for *every* $a\in A$, there exists $a'\in A$ such that $ga=a'g$. Here $ga\in gA$ is arbitrary, so $gA\subseteq Ag$. The other direction must also be true, so $gA\supseteq Ag$. In other words,

The operation defined above is well-defined
iff $gA=Ag$ *for all* $g\in G$.

This property merits a definition.

Definition 3.53. Let $A<G$. If

$$gA=Ag$$

for every $g\in G$, then A is a **normal subgroup** of G.

Notation 3.54. We write $A\triangleleft G$ to indicate that A is a normal subgroup of G.

Although we have outlined the argument above, we should show explicitly that if A is a normal subgroup, then the operation proposed for G/A is indeed well-defined.

Lemma 3.55. Let $A<G$. Then (CO1) implies (CO2).

(CO1) $A\triangleleft G$.

(CO2) Let $X,Y\in G/A$ and $x,y\in G$ such that $X=xA$ and $Y=yA$. The operation $\cdot$ on G/A defined by

$$XY=(xy)A$$

is well-defined for all $x,y\in G$.

Proof. Let $W,X,Y,Z\in G/A$ and choose $w,x,y,z\in G$ such that $W=wA$, $X=xA$, $Y=yA$, and $Z=zA$. To show that the operation is well-defined, we must show that if $W=X$ and

$Y = Z$, then $WY = XZ$ regardless of the values of w, x, y, or z. Assume therefore that $W = X$ and $Y = Z$. By substitution, $wA = xA$ and $yA = zA$. By Lemma 3.29(CE3), $w^{-1}x \in A$ and $y^{-1}z \in A$.

Since WY and XZ are sets, showing that they are equal requires us to show that each is a subset of the other. First we show that $WY \subseteq XZ$. To do this, let $t \in WY = (wy)A$. By definition of a coset, $t = (wy)a$ for some $a \in A$. What we will do now is rewrite t by

- using the fact that A is normal to move some element of a left, then right, through the representation of t; and
- using the fact that $W = X$ and $Y = Z$ to rewrite products of the form $w\check{\alpha}$ as $x\hat{\alpha}$ and $y\dot{\alpha}$ as $z\ddot{\alpha}$, where $\check{\alpha}, \hat{\alpha}, \dot{\alpha}, \ddot{\alpha} \in A$.

How, precisely? By the associative property, $t = w(ya)$. By definition of a coset, $ya \in yA$. By hypothesis, A is normal, so $yA = Ay$; thus, $ya \in Ay$. By definition of a coset, there exists $\check{a} \in A$ such that $ya = \check{a}y$. By substitution, $t = w(\check{a}y)$. By the associative property, $t = (w\check{a})y$. By definition of a coset, $w\check{a} \in wA$. By hypothesis, A is normal, so $wA = Aw$. Thus $w\check{a} \in Aw$. By hypothesis, $W = X$; that is, $wA = xA$. Thus $w\check{a} \in xA$, and by definition of a coset, $w\check{a} = x\hat{a}$ for some $\hat{a} \in A$. By substitution, $t = (x\hat{a})y$. The associative property again gives us $t = x(\hat{a}y)$; since A is normal we can write $\hat{a}y = y\dot{a}$ for some $\dot{a} \in A$. Hence $t = x(y\dot{a})$. Now,

$$y\dot{a} \in yA = Y = Z = zA,$$

so we can write $y\dot{a} = z\ddot{a}$ for some $\ddot{a} \in A$. By substitution and the definition of coset arithmetic,

$$t = x(z\ddot{a}) = (xz)\ddot{a} \in (xz)A = (xA)(zA) = XZ.$$

Since t was arbitrary in WY, we have shown that $WY \subseteq XZ$. A similar argument shows that $WY \supseteq XZ$; thus $WY = XZ$ and the operation is well-defined. □

An easy generalization of the argument of Example 3.52 shows the following Theorem.

Theorem 3.56. Let G be an abelian group, and $H < G$. Then $H \triangleleft G$.

Proof. You do it! See Exercise 3.65. □

We said before that we don't need an abelian group to have a normal subgroup. Here's a *great* example.

Example 3.57. Let

$$A_3 = \left\{\iota, \rho, \rho^2\right\} < D_3.$$

We call A_3 the **alternating group** on three elements. We claim that $A_3 \triangleleft D_3$. Indeed,

σ	σA_3	$A_3\sigma$
ι	A_3	A_3
ρ	A_3	A_3
ρ^2	A_3	A_3
φ	$\varphi A_3 = \{\varphi, \varphi\rho, \varphi\rho^2\}$ $= \{\varphi, \rho^2\varphi, \rho\varphi\} = A_3\varphi$	$A_3\varphi = \varphi A_3$
$\rho\varphi$	$\{\rho\varphi, (\rho\varphi)\rho, (\rho\varphi)\rho^2\}$ $= \{\rho\varphi, \varphi, \rho^2\varphi\} = \varphi A_3$	φA_3
$\rho^2\varphi$	$\{\rho^2\varphi, (\rho^2\varphi)\rho, (\rho^2\varphi)\rho^2\}$ $= \{\rho^2\varphi, \rho\varphi, \varphi\} = \varphi A_3$	φA_3

We have left out some details, though we also computed this table in Example 3.26, where we called the subgroup K instead of A_3. You should check the computation carefully, using extensively the fact that $\varphi\rho = \rho^2\varphi$.

Quotient groups

The set of cosets of a normal subgroup is, as desired, a group.

Theorem 3.58. Let G be a group. If $A \triangleleft G$, then G/A is a group.

Proof. Assume $A \triangleleft G$. By Lemma 3.55, the operation is well-defined, so it remains to show that G/A satisfies the properties of a group.

(closure) Closure follows from the fact that multiplication of cosets is well-defined when $A \triangleleft G$, as shown in Lemma 3.55: Let $X, Y \in G/A$, and choose $g_1, g_2 \in G$ such that $X = g_1A$ and $Y = g_2A$. By definition of coset multiplication, $XY = (g_1A)(g_2A) = (g_1g_2)A \in G/A$. Since X, Y were arbitrary in G/A, coset multiplication is closed.

(associativity) The associative property of G/A follows from the associative property of G. Let $X, Y, Z \in G/A$; choose $g_1, g_2, g_3 \in G$ such that $X = g_1A$, $Y = g_2A$, and $Z = g_3A$. Then

$$(XY)Z = [(g_1A)(g_2A)](g_3A).$$

By definition of coset multiplication,

$$(XY)Z = ((g_1g_2)A)(g_3A).$$

By the definition of coset multiplication,

$$(XY)Z = ((g_1g_2)g_3)A.$$

(Note the parentheses grouping g_1g_2.) Now apply the associative property of G

and reverse the previous steps to obtain

$$\begin{aligned}(XY)Z &= (g_1(g_2g_3))A \\ &= (g_1A)((g_2g_3)A) \\ &= (g_1A)[(g_2A)(g_3A)] \\ &= X(YZ).\end{aligned}$$

Since $(XY)Z = X(YZ)$ and X, Y, Z were arbitrary in G/A, coset multiplication is associative.

(identity) We claim that the identity of G/A is A itself. Let $X \in G/A$, and choose $g \in G$ such that $X = gA$. Since $e \in A$, Lemma 3.29 on page 102 implies that $A = eA$, so

$$XA = (gA)(eA) = (ge)A = gA = X.$$

Since X was arbitrary in G/A and $XA = X$, A is the identity of G/A.

(inverse) Let $X \in G/A$. Choose $g \in G$ such that $X = gA$, and let $Y = g^{-1}A$. We claim that $Y = X^{-1}$. By applying substitution and the operation on cosets,

$$XY = (gA)\left(g^{-1}A\right) = \left(gg^{-1}\right)A = eA = A.$$

Hence X has an inverse in G/A. Since X was arbitrary in G/A, every element of G/A has an inverse.

We have shown that G/A satisfies the properties of a group. □

Definition 3.59. Let G be a group, and $A \triangleleft G$. Then G/A is **the quotient group of G with respect to A**, also called G **mod** A.

Normally we simply say "the quotient group" rather than "the quotient group of G with respect to A."

Example 3.60. Since A_3 is a normal subgroup of D_3, D_3/A_3 is a group. By Lagrange's Theorem, it has $6/3 = 2$ elements. The composition table is

$\circ$	A_3	φA_3
A_3	A_3	φA_3
φA_3	φA_3	A_3

We meet an important quotient group in Section 3.5.

Exercises.

Exercise 3.61. Show that for any group G, $\{e\} \triangleleft G$ and $G \triangleleft G$.

Exercise 3.62. Recall from Exercise 3.12 that if $d \mid n$, then $\Omega_d < \Omega_n$.

(a) Explain how we know that, in fact, $\Omega_d \triangleleft \Omega_n$.

(b) Compute the Cayley table of the quotient group Ω_8/Ω_2. Does it have the same structure as the Klein 4-group, or as the Cyclic group of order 4?

Exercise 3.63. Let $H = \langle \mathbf{i} \rangle < Q_8$.

(a) Show that $H \triangleleft Q_8$ by computing all the cosets of H.

(b) Compute the multiplication table of Q_8/H.

Exercise 3.64. Let $H = \langle -1 \rangle < Q_8$.

(a) Show that $H \triangleleft Q_8$ by computing all the cosets of H.

(b) Compute the multiplication table of Q_8/H.

(c) With which well-known group does Q_8/H have the same structure?

Exercise 3.65. Let G be an abelian group. Explain why for any $H < G$ we know that $H \triangleleft G$.

Definition 3.66. Let G be a group, $g \in G$, and $H < G$. Define the **conjugation** of H by g as

$$gHg^{-1} = \{h^g : h \in H\}.$$

(The notation h^g is the definition of conjugation from Exercise 2.37 on page 66; that is, $h^g = ghg^{-1}$.)

Let G be a group, and $H < G$.
Claim: $H \triangleleft G$ if and only if $H = gHg^{-1}$ for all $g \in G$.
Proof:

1. First, we show that if $H \triangleleft G$, then _____.
   (a) Assume _____.
   (b) By definition of normal, _____.
   (c) Let g_____.
   (d) We first show that $H \subseteq gHg^{-1}$.
      i. Let h_____.
      ii. By 1b, $hg \in$_____.
      iii. By definition, there exists $h' \in H$ such that $hg =$_____.
      iv. Multiply both sides on the right by g^{-1} to see that $h =$_____.
      v. By _____, $h \in gHg^{-1}$.
      vi. Since h was arbitrary, _____.
   (e) Now we show that $H \supseteq gHg^{-1}$.
      i. Let $x \in$_____.
      ii. By _____, $x = ghg^{-1}$ for some $h \in H$.
      iii. By _____, $gh \in Hg$.
      iv. By _____, there exists $h' \in H$ such that $gh = h'g$.
      v. By _____, $x = (h'g)\,g^{-1}$.
      vi. By _____, $x = h'$.
      vii. By _____, $x \in H$.
      viii. Since x was arbitrary, _____.
   (f) We have shown that $H \subseteq gHg^{-1}$ and $H \supseteq gHg^{-1}$. Thus, _____.
2. Now, we show _____: that is, if $H = gHg^{-1}$ for all $g \in G$, then $H \triangleleft G$.
   (a) Assume _____.
   (b) First, we show that $gH \subseteq Hg$.
      i. Let $x \in$_____.
      ii. By _____, there exists $h \in H$ such that $x = gh$.
      iii. By _____, $g^{-1}x = h$.
      iv. By _____, there exists $h' \in H$ such that $h = g^{-1}h'g$.
      (A key point here is that this is true for *all* $g \in G$.)
      v. By _____, $g^{-1}x = g^{-1}h'g$.
      vi. By _____, $x = g\,(g^{-1}h'g)$.
      vii. By _____, $x = h'g$.
      viii. By _____, $x \in Hg$.
      ix. Since x was arbitrary, _____.
   (c) The proof that _____ is similar.
   (d) We have show that _____. Thus, $gH = Hg$.

Figure 3.6. Material for Exercise 3.67

Exercise 3.67. Fill in each blank of Figure 3.6 with the appropriate justification or statement.[10]

[10]Certain texts define a normal subgroup this way; that is, a subgroup H is normal if every conjugate of H is precisely H. They then prove that in this case, any left coset equals the corresponding right coset.

Let G be a group. The **centralizer** of G is

$$Z(G) = \{g \in G : xg = gx \ \forall x \in G\}.$$

Claim: $Z(G) \triangleleft G$.
Proof:

1. First, we must show that $Z(G) < G$.
   (a) Let g, h, x_____.
   (b) By _____, $xg = gx$ and $xh = hx$.
   (c) By _____, $xh^{-1} = h^{-1}x$.
   (d) By _____, $h^{-1} \in Z(G)$.
   (e) By the associative property and the definition of $Z(G)$, $(gh^{-1})x =$_____=_____=...$= x(gh^{-1})$.
   (Fill in more blanks as needed.)
   (f) By _____, $gh^{-1} \in Z(G)$.
   (g) By _____, $Z(G) < G$.
2. Now, we show that $Z(G)$ is normal.
   (a) Let x_____.
   (b) First we show that $xZ(G) \subseteq Z(G)x$.
      i. Let y_____.
      ii. By definition of cosets, there exists $g \in Z(G)$ such that $y =$_____.
      iii. By definition of $z(G)$, _____.
      iv. By definition of _____, $y \in Z(G)x$.
      v. By _____, $xZ(G) \subseteq Z(G)x$.
   (c) A similar argument shows that _____.
   (d) By definition, _____. That is, $Z(G)$ is normal.

Figure 3.7. Material for Exercise 3.71

Exercise 3.68. Recall the subgroup L of $\mathbb{R}^2$ from Exercises 3.16 on page 98 and 3.34 on page 103.
(a) Explain how we know that $L \triangleleft \mathbb{R}^2$ *without* checking that $p + L = L + p$ for any $p \in \mathbb{R}^2$.
(b) Sketch two elements of $\mathbb{R}^2/L$ and show their sum.

Exercise 3.69. Explain why every subgroup of $D_m(\mathbb{R})$ is normal.

Exercise 3.70. Show that Q_8 is not a normal subgroup of $GL_m(\mathbb{C})$.

Exercise 3.71. Fill in every blank of Figure 3.7 with the appropriate justification or statement.

Exercise 3.72. Let G be a group, and $H < G$. Define the **normalizer** of H as

$$N_G(H) = \{g \in G : gH = Hg\}.$$

Show that $H \triangleleft N_G(H)$.

Exercise 3.73. Let G be a group, and $A < G$. Suppose that $|G/A| = 2$; that is, the subgroup A partitions G into precisely two left cosets. Show that:

- $A \triangleleft G$; and
- G/A is abelian.

Exercise 3.74. Recall from Exercise 2.37 on page 66 the commutator of two elements of a group. Let $[G,G]$ denote the intersection of all subgroups of G that contain $[x,y]$ for all $x,y \in G$.

(a) Compute $[D_3, D_3]$.
(b) Compute $[Q_8, Q_8]$.
(c) Show that $[G,G] < G$.
(d) Fill in each blank of Figure 3.8 with the appropriate justification or statement.

Definition 3.75. We call $[G,G]$ the **commutator subgroup** of G, and make use of it in Section 3.6.

Claim: For any group G, $[G,G]$ is a normal subgroup of G.
Proof:

1. Let _____.
2. We will use Exercise 3.67 to show that $[G,G]$ is normal. Let $g \in$_____.
3. First we show that $[G,G] \subseteq g\,[G,G]\,g^{-1}$. Let $h \in [G,G]$.
   (a) We need to show that $h \in g\,[G,G]\,g^{-1}$. It will suffice to show that this is true if h has the simpler form $h = [x,y]$, since _____. Thus, choose $x,y \in G$ such that $h = [x,y]$.
   (b) By _____, $h = x^{-1}y^{-1}xy$.
   (c) By _____, $h = ex^{-1}ey^{-1}exeye$.
   (d) By _____, $h = \left(gg^{-1}\right)x^{-1}\left(gg^{-1}\right)y^{-1}\left(gg^{-1}\right)x\left(gg^{-1}\right)y\left(gg^{-1}\right)$.
   (e) By _____, $h = g\left(g^{-1}x^{-1}g\right)\left(g^{-1}y^{-1}g\right)\left(g^{-1}xg\right)\left(g^{-1}yg\right)g^{-1}$.
   (f) By _____, $h = g\left(x^{-1}\right)^{g^{-1}}\left(y^{-1}\right)^{g^{-1}}\left(x^{g^{-1}}\right)\left(y^{g^{-1}}\right)g^{-1}$.
   (g) By Exercise 2.37 on page 66(c), $h =$_____.
   (h) By definition of the commutator, $h =$_____.
   (i) By _____, $h \in g\,[G,G]\,g^{-1}$.
   (j) Since _____, $[G,G] \subseteq g\,[G,G]\,g^{-1}$.
4. Conversely, we show that $[G,G] \supseteq g\,[G,G]\,g^{-1}$. Let $h \in g\,[G,G]\,g^{-1}$.
   (a) We need to show that $h \in [G,G]$. It will suffice to show this is true if h has the simpler form $h = g\,[x,y]\,g^{-1}$, since _____. Thus, choose $x,y \in G$ such that $h = g\,[x,y]\,g^{-1}$.
   (b) By _____, $h = [x,y]^g$.
   (c) By _____, $h = [x^g, y^g]$.
   (d) By _____, $h \in G$.
   (e) Since _____, $[G,G] \supseteq g\,[G,G]\,g^{-1}$.
5. We have shown that $[G,G] \subseteq g\,[G,G]\,g^{-1}$ and $[G,G] \supseteq g\,[G,G]\,g^{-1}$. By _____, $[G,G] = g\,[G,G]\,g^{-1}$.

Figure 3.8. Material for Exercise 3.74

3.5: “Clockwork” groups

By Theorem 3.56, every subgroup H of $\mathbb{Z}$ is normal. Let $n \in \mathbb{Z}$; since $n\mathbb{Z} < \mathbb{Z}$, it follows that $n\mathbb{Z} \triangleleft \mathbb{Z}$. Thus $\mathbb{Z}/n\mathbb{Z}$ is a quotient group.

We used $n\mathbb{Z}$ in many examples of subgroups. One reason is that you are accustomed to working with $\mathbb{Z}$, so it should be conceptually easy. Another reason is that the quotient group $\mathbb{Z}/n\mathbb{Z}$ has a vast array of applications in number theory and computer science. You will see some of these in Chapter 6. Because this group is so important, we give it several special names.

Definition 3.76. Let $n \in \mathbb{Z}$. We call the quotient group $\mathbb{Z}/n\mathbb{Z}$
- $\mathbb{Z}$ **mod** n, or
- the **linear residues modulo** n.

Notation 3.77. It is common to write $\mathbb{Z}_n$ instead of $\mathbb{Z}/n\mathbb{Z}$.

Example 3.78. You already saw a bit of $\mathbb{Z}_4 = \mathbb{Z}/4\mathbb{Z}$ at the beginning of Section 3.2 and again in Example 3.52. Recall that $\mathbb{Z}_4 = \{4\mathbb{Z}, 1+4\mathbb{Z}, 2+4\mathbb{Z}, 3+4\mathbb{Z}\}$. Addition in this group will always give us one of those four representations of the cosets:

$$\begin{aligned}(2+4\mathbb{Z}) + (1+4\mathbb{Z}) &= 3+4\mathbb{Z};\\ (1+4\mathbb{Z}) + (3+4\mathbb{Z}) &= 4+4\mathbb{Z} = 4\mathbb{Z};\\ (2+4\mathbb{Z}) + (3+4\mathbb{Z}) &= 5+4\mathbb{Z} = 1+4\mathbb{Z};\end{aligned}$$

and so forth.

Reasoning similar to that used at the beginning of Section 3.2 would show that

$$\mathbb{Z}_{31} = \mathbb{Z}/31\mathbb{Z} = \{31\mathbb{Z}, 1+31\mathbb{Z}, \ldots, 30+31\mathbb{Z}\}.$$

We show this explicitly in Theorem 3.82.

Before looking at some properties of $\mathbb{Z}_n$, let's look for an easier way to talk about its elements. It is burdensome to write $a+n\mathbb{Z}$ whenever we want to discuss an element of $\mathbb{Z}_n$, so we adopt the following convention.

Notation 3.79. Let $A \in \mathbb{Z}_n$ and choose $r \in \mathbb{Z}$ such that $A = r + n\mathbb{Z}$.
- If it is clear from context that A is an element of $\mathbb{Z}_n$, then we simply write r instead of $r+n\mathbb{Z}$.
- If we want to emphasize that A is an element of $\mathbb{Z}_n$ (perhaps there are a lot of integers hanging about) then we write $[r]_n$ instead of $r+n\mathbb{Z}$.
- If the value of n is obvious from context, we simply write $[r]$.

To help you grow accustomed to the notation $[r]_n$, we use it for the rest of this chapter, even when n is mind-bogglingly obvious.

The first property is that, for most values of n, $\mathbb{Z}_n$ has finitely many elements. To show that there are finitely many elements of $\mathbb{Z}_n$, we rely on the following fact, which is important enough to highlight as a separate result.

Lemma 3.80. Let $n \in \mathbb{Z}\setminus\{0\}$ and $[a]_n \in \mathbb{Z}_n$. Use the Division Theorem to choose $q, r \in \mathbb{Z}$ such that $a = qn + r$ and $0 \leq r < |n|$. Then $[a]_n = [r]_n$.

The proof of Lemma 3.80 on the preceding page is similar to the discussion in Example 3.38 on page 104, so you might want to reread that.

Proof. We give two different proofs. Both are based on the fact that $[a]_n$ and $[r]_n$ are *cosets;* so showing that they are equal is tantamount to showing that a and r are different elements of the same set.

(1) By definition and substitution,

$$\begin{aligned}[a]_n &= a + n\mathbb{Z} \\ &= (qn + r) + n\mathbb{Z} \\ &= \{(qn + r) + nd : d \in \mathbb{Z}\} \\ &= \{r + n(q + d) : d \in \mathbb{Z}\} \\ &= \{r + nm : m \in \mathbb{Z}\} \\ &= r + n\mathbb{Z} \\ &= [r]_n .\end{aligned}$$

(2) Rewrite $a = qn + r$ as $a - r = qn$. By definition, $a - r \in n\mathbb{Z}$. The immensely useful Lemma 3.29 shows that $a + n\mathbb{Z} = r + n\mathbb{Z}$, and the notation implies that $[a]_n = [r]_n$. □

Definition 3.81. On account of Lemma 3.80, we can designate the remainder of division of a by n, whose value is between 0 and $|n| - 1$, inclusive, as the **canonical representation** of $[a]_n$ in $\mathbb{Z}_n$.

Theorem 3.82. $\mathbb{Z}_n$ is finite for every nonzero $n \in \mathbb{Z}$. In fact, if $n \neq 0$ then $\mathbb{Z}_n$ has $|n|$ elements corresponding to the remainders from division by n: $0, 1, 2, \ldots, n - 1$.

Proof. Lemma 3.80 on the preceding page states that every element of such $\mathbb{Z}_n$ can be represented by $[r]_n$ for some $r \in \mathbb{Z}$ where $0 \leq r < |n|$. But there are only $|n|$ possible choices for such a remainder. □

Let’s look at how we can perform arithmetic in $\mathbb{Z}_n$.

Lemma 3.83. Let $d, n \in \mathbb{Z}$ and $[a]_n, [b]_n \in \mathbb{Z}_n$. Then

$$[a]_n + [b]_n = [a + b]_n$$

and

$$d\,[a]_n = [da]_n .$$

For example, $[3]_7 + [9]_7 = [3 + 9]_7 = [12]_7 = [5]_7$ and $-4\,[3]_5 = [-4 \cdot 3]_5 = [-12]_5 = [3]_5$.

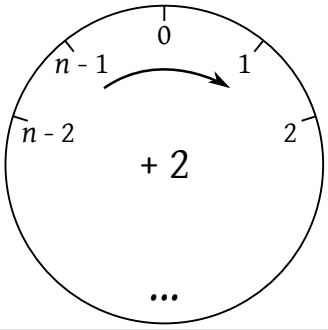


Figure 3.9. Addition in $\mathbb{Z}_n$ is “clockwork”: $[n-1]_n + [2]_n = [1]_n$.

Proof. The proof really amounts to little more than manipulating the notation. By the definitions of coset addition and of $\mathbb{Z}_n$,

$$\begin{aligned}[a]_n + [b]_n &= (a + n\mathbb{Z}) + (b + n\mathbb{Z}) \\ &= (a+b) + n\mathbb{Z} \\ &= [a+b]_n .\end{aligned}$$

For $d\,[a]_n$, we consider three cases.

If $d = 0$, then $d\,[a]_n = [0]_n$ by Notation 2.51 on page 76, and $[0]_n = [0 \cdot a]_n = [da]_n$. By substitution, then, $d\,[a]_n = [da]_n$.

If d is positive, then the expression $d\,[a]_n$ is the sum of d copies of $[a]_n$, which the Lemma’s first claim (now proved) implies to be

$$\begin{aligned}\underbrace{[a]_n + [a]_n + \cdots + [a]_n}_{d \text{ times}} &= [2a]_n + \underbrace{[a]_n + \cdots + [a]_n}_{d-2 \text{ times}} \\ &\vdots \\ &= [da]_n .\end{aligned}$$

If d is negative, then Notation 2.51 again tells us that $d\,[a]_n$ is the sum of $|d|$ copies of $-[a]_n$. So, what is the additive inverse of $[a]_n$? Using the first claim, $[a]_n + [-a]_n = [a + (-a)]_n = [0]_n$, so $-[a]_n = [-a]_n$. By substitution,

$$\begin{aligned}d\,[a]_n &= |d|\,(-[a_n]) = |d|\,[-a]_n \\ &= [|d| \cdot (-a)]_n = [-d \cdot (-a)]_n = [da]_n .\end{aligned}$$

□

Lemmas 3.80 and 3.83 imply that each $\mathbb{Z}_n$ acts as a **“clockwork” group**. Why?

- To add $[a]_n$ and $[b]_n$, let $c = a + b$.
- If $0 \le c < |n|$, then you are done. After all, division of c by n gives $q = 0$ and $r = c$.
- Otherwise, $c < 0$ or $c \ge |n|$, so we divide c by n, obtaining q and r where $0 \le r < |n|$. The sum is $[r]_n$.

We call this “clockwork” because it counts like a clock: if you sit down at 5 o’clock and wait two hours, you rise at not at 13 o’clock, but at $13 - 12 = 1$ o’clock. See Figure 3.9.

It should be clear from Example 2.9 on page 60 as well as Exercise 2.31 on page 65 that $\mathbb{Z}_2$ and $\mathbb{Z}_3$ have precisely the same structure as the groups of order 2 and 3. On the other hand, we saw in Exercise 2.32 on page 65 that there are two possible structures for a group of order 4: the Klein 4-group, and a cyclic group. Which structure does $\mathbb{Z}_4$ have?

Example 3.84. Use Lemma 3.83 to observe that

$$\langle [1]_4 \rangle = \{[0]_4, [1]_4, [2]_4, [3]_4\}$$

since $[2]_4 = [1]_4 + [1]_4$, $[3]_4 = [2]_4 + [1]_4$, and $[0]_4 = 0 \cdot [1]_4$ (or $[0]_4 = [3]_4 + [1]_4$).

The fact that $\mathbb{Z}_4$ was cyclic makes one wonder: is $\mathbb{Z}_n$ always cyclic? Yes!

Theorem 3.85. $\mathbb{Z}_n$ is cyclic for every $n \in \mathbb{Z}$.

This theorem has a more general version, which you will prove in the homework.

Proof. Let $n \in \mathbb{Z}$ and $[a]_n \in \mathbb{Z}_n$. By Lemma 3.83,

$$[a]_n = [a \cdot 1]_n = a\,[1]_n \in \langle [1]_n \rangle .$$

Since $[a]_n$ was arbitrary in $\mathbb{Z}_n$, $\mathbb{Z}_n \subseteq \langle [1]_n \rangle$. Closure implies that $\mathbb{Z}_n \supseteq \langle [1]_n \rangle$, so in fact $\mathbb{Z}_n = \langle [1]_n \rangle$, and $\mathbb{Z}_n$ is therefore cyclic. □

Not *every* non-zero element necessarily generates $\mathbb{Z}_n$. We know that $[2]_4 + [2]_4 = [4]_4 = [0]_4$, so in $\mathbb{Z}_4$, we have

$$\langle [2]_4 \rangle = \{[0]_4, [2]_4\} \subsetneq \mathbb{Z}_4.$$

A natural and interesting followup question is, which non-zero elements *do* generate $\mathbb{Z}_n$? You need a bit more background in number theory before you can answer that question, but in the exercises you will build some more addition tables and use them to formulate a hypothesis.

The following important lemma gives an "easy" test for whether two integers are in the same coset of $\mathbb{Z}_n$.

Lemma 3.86. Let $a, b, n \in \mathbb{Z}$ and assume that $n \neq 0$. The following are equivalent.
(A) $a + n\mathbb{Z} = b + n\mathbb{Z}$.
(B) $[a]_n = [b]_n$.
(C) $n \mid (a - b)$.

Proof. You do it! See Exercise 3.93. □

Exercises.

Exercise 3.87. We showed that $\mathbb{Z}_n$ is finite for $n \neq 0$. What if $n = 0$? How many elements would it have? Illustrate a few additions and subtractions, and indicate whether you think that $\mathbb{Z}_0$ is an interesting or useful group.

Exercise 3.88. In the future, we won't actually talk about $\mathbb{Z}_n$ for $n < 0$. Show that this is because $\mathbb{Z}_n = \mathbb{Z}_{|n|}$.

Exercise 3.89. Write out the Cayley tables for $\mathbb{Z}_2$ and $\mathbb{Z}_3$. Remember that the operation is addition.

Exercise 3.90. Write down the Cayley table for $\mathbb{Z}_5$. Remember that the operation is addition. Which elements generate $\mathbb{Z}_5$?

Exercise 3.91. Write down the Cayley table for $\mathbb{Z}_6$.Remember that the operation is addition. Which elements generate $\mathbb{Z}_6$?

Exercise 3.92. Compare the results of Example 3.84 and Exercises 3.89, 3.90, and 3.91. Formulate a conjecture as to which elements generate $\mathbb{Z}_n$. Do not try to prove your example.

Exercise 3.93. Prove Lemma 3.86.

Exercise 3.94. Prove the following generalization of Theorem 3.85: If G is a cyclic group and $A \triangleleft G$, then G/A is cyclic.

3.6: "Solvable" groups

One of the major motivations of group theory was the question of whether a polynomial can be solved by radicals. For example, if we have a quadratic equation $ax^2 + bx + c = 0$, then[11]

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

Since the solution contains nothing more than addition, multiplication, and radicals, we say that a quadratic equation is *solvable by radicals*.

Similar formulas can be found for cubic and quartic equations. When mathematicians turned their attention to quintic equations, however, they hit a wall: they weren't able to use previous techniques to find a "quintic formula". Eventually, it was shown that this is because some quintic equations are not solvable by radicals. The method they used to show this is based on the following concept.

Definition 3.95. If a group G contains subgroups $G_0, G_1, \ldots, G_n$ such that

- $G_0 = \{e\}$;
- $G_n = G$;
- $G_{i-1} \triangleleft G_i$; and
- G_i / G_{i-1} is abelian,

then G is a **solvable group**. The chain of subgroups $G_0, \ldots, G_n$ is called a **normal series**.

[11] Well, as long as $a \neq 0$. But then you wouldn't consider it quadratic, would you?

Example 3.96. Any finite abelian group G is solvable: let $G_0 = \{e\}$ and $G_1 = G$. Subgroups of an abelian group are always normal, so $G_0 \triangleleft G_1$. In addition, $X, Y \in G_1/G_0$ implies that $X = x\{e\}$ and $Y = y\{e\}$ for some $x, y \in G_1 = G$. Since G is abelian,

$$XY = (xy)\{e\} = (yx)\{e\} = YX.$$

Example 3.97. The group D_3 is solvable. To see this, let $n = 2$ and $G_1 = \langle \rho \rangle$:

- By Exercise 3.61 on page 111, $\{e\} \triangleleft G_1$. To see that $G_1/\{e\}$ is abelian, note that for any $X, Y \in G_1/\{e\}$, we can write $X = x\{e\}$ and $Y = y\{e\}$ for some $x, y \in G_1$. By definition of G_1, we can write $x = \rho^a$ and $y = \rho^b$ for some $a, b \in \mathbb{Z}$. We can then fall back on the commutative property of addition in $\mathbb{Z}$ to show that

$$\begin{aligned} XY &= (xy)\{e\} = \rho^{a+b}\{e\} \\ &= \rho^{b+a}\{e\} = (yx)\{e\} = YX. \end{aligned}$$

- By Exercise 3.73 on page 114 and the fact that $|G_1| = 3$ and $|G_2| = 6$, we know that $G_1 \triangleleft G_2$. The same exercise tells us that G_2/G_1 is abelian.

The following properties of solvable subgroups are very useful in a branch of algebra called *Galois Theory*.

Theorem 3.98. Every quotient group of a solvable group is solvable.

Proof. Let G be a group and $A \triangleleft G$. We need to show that G/A is solvable. Since G is solvable, choose a normal series $G_0, \ldots, G_n$. For each $i = 0, \ldots, n$, put

$$A_i = \{gA : g \in G_i\}.$$

We claim that the chain $A_0, A_1, \ldots, A_n$ likewise satisfies the definition of a solvable group.

First, we show that $A_{i-1} \triangleleft A_i$ for each $i = 1, \ldots, n$. Let $X \in A_i$; by definition, $X = xA$ for some $x \in G_i$. We have to show that $XA_{i-1} = A_{i-1}X$. Let $Y \in A_{i-1}$; by definition, $Y = yA$ for some $y \in G_{i-1}$. Recall that $G_{i-1} \triangleleft G_i$, so there exists $\widehat{y} \in G_{i-1}$ such that $xy = \widehat{y}x$. Let $\widehat{Y} = \widehat{y}A$; since $\widehat{y} \in G_{i-1}$, $\widehat{Y} \in A_{i-1}$. Using substitution and the definition of coset arithmetic, we have

$$XY = (xy)A = (\widehat{y}x)A = \widehat{Y}X \in A_{i-1}X.$$

Since Y was arbitrary in A_{i-1}, $XA_{i-1} \subseteq A_{i-1}X$. A similar argument shows that $XA_{i-1} \supseteq A_{i-1}X$, so the two are equal. Since X is an arbitrary coset of A_{i-1} in A_i, we conclude that $A_{i-1} \triangleleft A_i$.

Second, we show that A_i/A_{i-1} is abelian. Let $X, Y \in A_i/A_{i-1}$. By definition, we can write $X = SA_{i-1}$ and $Y = TA_{i-1}$ for some $S, T \in A_i$. Again by definition, there exist $s, t \in G_i$ such that $S = sA$ and $T = tA$. Let $U \in A_{i-1}$; we can likewise write $U = uA$ for some $u \in G_{i-1}$. Since G_i/G_{i-1} is abelian, $(st)G_{i-1} = (ts)G_{i-1}$; thus, $(st)u = (ts)v$ for some $v \in G_{i-1}$. By

definition, $vA \in A_{i-1}$. By substitution and the definition of coset arithmetic, we have

$$\begin{aligned} XY &= (ST)A_{i-1} = ((st)A)A_{i-1} \\ &= [(st)A](uA) = ((st)u)A \\ &= ((ts)v)A = [(ts)A](vA) \\ &= ((ts)A)A_{i-1} = (TS)A_{i-1} \\ &= YX. \end{aligned}$$

Since X and Y were arbitrary in the quotient group A_i/A_{i-1}, we conclude that it is abelian.

We have constructed a normal series in G/A; it follows that G/A is solvable. □

The following result is also true:

Theorem 3.99. Every subgroup of a solvable group is solvable.

Proving it, however, is a little more difficult. We need the definition of the commutator from Exercises 2.37 on page 66 and 3.74 on page 115.

Definition 3.100. Let G be a group. The **commutator subgroup** G' of G is the intersection of all subgroups of G that contain $[x,y]$ for all $x,y \in G$.

Notice that $G' < G$ by Exercise 3.20.

Notation 3.101. We wrote G' as $[G,G]$ in Exercise 3.74.

Lemma 3.102. For any group G, $G' \triangleleft G$. In addition, G/G' is abelian.

Proof. You showed that $G' \triangleleft G$ in Exercise 3.74 on page 115. To show that G/G' is abelian, let $X, Y \in G/G'$. Write $X = xG'$ and $Y = yG'$ for appropriate $x, y \in G$. By definition, $XY = (xy)G'$. Let $g' \in G'$; by definition, $g' = [a,b]$ for some $a, b \in G$. Since G' is a group, it is closed under the operation, so $[x,y][a,b] \in G'$. Let $z \in G'$ such that $[x,y][a,b] = z$. Rewrite this expression as

$$\left(x^{-1}y^{-1}xy\right)[a,b] = z \quad \Longrightarrow \quad (xy)[a,b] = (yx)z.$$

(Multiply both sides of the equation on the left by yx.) Hence

$$(xy)g' = (xy)[a,b] = (yx)z \in (yx)G'.$$

Since g' was arbitrary, $(xy)G' \subseteq (yx)G'$. A similar argument shows that $(xy)G' \supseteq (yx)G'$. Thus

$$XY = (xy)G' = (yx)G' = YX,$$

and G/G' is abelian. □

Lemma 3.103. If $H \subseteq G$, then $H' \subseteq G'$.

Proof. You do it! See Exercise 3.107. □

Notation 3.104. Define $G^{(0)} = G$ and $G^{(i)} = \left(G^{(i-1)}\right)'$; that is, $G^{(i)}$ is the commutator subgroup of $G^{(i-1)}$.

Lemma 3.105. A group is solvable if and only if $G^{(n)} = \{e\}$ for some $n \in \mathbb{N}$.

Proof. ($\Longrightarrow$) Suppose that G is solvable. Let $G_0, \ldots, G_n$ be a normal series for G. We claim that $G^{(n-i)} \subseteq G_i$. If this claim were true, then $G^{(n-0)} \subseteq G_0 = \{e\}$, and we would be done. We proceed by induction on $n - i \in \mathbb{N}$.

Inductive base: If $n - i = 0$, then $G^{(n-i)} = G = G_n$. Also, $i = n$, so $G^{(n-i)} = G_n = G_i$, as claimed.

Inductive hypothesis: Assume that the assertion holds for $n - i$.

Inductive step: By definition, $G^{(n-i+1)} = \left(G^{(n-i)}\right)'$. By the inductive hypothesis, $G^{(n-i)} \subseteq G_i$; by Lemma 3.103, $\left(G^{(n-i)}\right)' \subseteq G_i'$. Hence

$$G^{(n-i+1)} \subseteq G_i'. \tag{12}$$

Recall from the properties of a normal series that G_i / G_{i-1} is abelian; for any $x, y \in G_i$, we have

$$\begin{aligned}(xy)\, G_{i-1} &= \left(xG_{i-1}\right)\left(yG_{i-1}\right) \\ &= \left(yG_{i-1}\right)\left(xG_{i-1}\right) = (yx)\, G_{i-1}.\end{aligned}$$

By Lemma 3.29 on page 102, $(yx)^{-1}(xy) \in G_{i-1}$; in other words, $[x,y] = x^{-1}y^{-1}xy \in G_{i-1}$. Since x and y were arbitrary in G_i, we have $G_i' \subseteq G_{i-1}$. Along with (12), this implies that $G^{(n-(i-1))} = G^{(n-i+1)} \subseteq G_{i-1}$.

We have shown the claim; thus, $G^{(n)} = \{e\}$ for some $n \in \mathbb{N}$.

($\Longleftarrow$) Suppose that $G^{(n)} = \{e\}$ for some $n \in \mathbb{N}$. We have

$$\{e\} = G^{(n)} < G^{(n-1)} < \cdots < G^{(0)} = G.$$

By Lemma 3.102, the subgroups form a normal series; that is,

$$\{e\} = G^{(n)} \triangleleft G^{(n-1)} \triangleleft \cdots \triangleleft G^{(0)} = G$$

and $G^{(n-i)} / G^{(n-(i-1))}$ is abelian for each $i = 0, \ldots, n-1$. As this is a normal series, we have shown that G is solvable. □

We can now prove Theorem 3.99.

Proof of Theorem 3.99. Let $H < G$. Assume G is solvable; by Lemma 3.105, $G^{(n)} = \{e\}$. By Lemma 3.103, $H^{(i)} \subseteq G^{(i)}$ for all $n \in \mathbb{N}$, so $H^{(n)} \subseteq \{e\}$. By the definition of a group, $H^{(n)} \supseteq \{e\}$, so the two are equal. By the same lemma, H is solvable. □

Exercises.

Exercise 3.106. Explain why Ω_n is solvable for any $n \in \mathbb{N}^+$.

Exercise 3.107. Show that if $H \subseteq G$, then $H' \subseteq G'$.

Exercise 3.108. Show that Q_8 is solvable.

Exercise 3.109. In the textbook *God Created the Integers...* the theoretical physicist Stephen Hawking reprints some of the greatest mathematical results in history, adding some commentary. For an excerpt from Evariste Galois' *Memoirs,* Hawking sums up the main result this way.

> To be brief, Galois demonstrated that the general polynomial of degree n could be solved by radicals if and only if every subgroup N of the group of permutations S_n is a normal subgroup. Then he demonstrated that every subgroup of S_n is normal for all $n \leq 4$ but not for any $n > 5$.
>
> —p. 105

Unfortunately, Hawking's explanation is completely wrong, and this exercise leads you towards an explanation as to why.[12] You have not yet studied the groups of permutations S_n, but you will learn in Section 5.1 that the group S_3 is really the same as D_3. So we look at D_3, instead.

(a) Find all six subgroups of D_3.
(b) It is known that the general polynomial of degree 3 can be solved by radicals. According to the quote above, what must be true about all the subgroups of D_3?
(c) Why is Hawking's explanation of Galois' result "obviously" wrong?

(To be precise, S_3 is "isomorphic" to D_3. We discuss group isomorphisms in Chapter 4. Exercise 5.37 of Chapter 5 asks you to show that $S_3 \cong D_3$. We talk about solvability by radicals in Chapter 9.)

[12] Perhaps Hawking was trying to simplify what Galois actually showed, and went too far. (I've done much worse in my lifetime.) In fact, Galois showed that a polynomial of degree n could be solved by radicals if and only if a corresponding group, now called its **Galois group,** was a solvable group. He then showed that the Galois group of $x^5 + 2x + 5$ was not a solvable group.

Chapter 4:
Isomorphisms

We have on occasion observed that different groups have the same Cayley table. We have also talked about different groups having the same structure: regardless of whether a group of order two is additive or multiplicative, its elements behave in exactly the same fashion. The groups may consist of elements whose construction was quite different, and the definition of the operation may also be different, but the "group behavior" is nevertheless identical.

We saw in Chapter 1 that algebraists describe such a relationship between two monoids as *isomorphic*. Isomorphism for groups has the same intuitive meaning as isomorphism for monoids:

If two groups G and H have identical group structure,
we say that G and H are *isomorphic*.

We want to study isomorphism of groups in quite a bit of detail, so to define isomorphism precisely, we start by reconsidering another topic that you studied in the past, functions. There we will also introduce the related notion of *homomorphism*. Despite the same basic intuitive definition, the precise definition of group homorphism turns out simpler than for monoids. This is the focus of Section 4.1. Section 4.2 lists some results that should help convince you that the existence of an isomorphism does, in fact, show that two groups have an identical group structure. Section 4.3 describes how we can create new isomorphisms from a homomorphism's *kernel*, a special subgroup defined by a homomorphism. Section 4.4 introduces a class of isomorphism that is important for later applications, an *automorphism*.

4.1: Homomorphisms

Groups have more structure than monoids. Just as a monoid homomorphism would require that we preserve both identities and the operation (page 45), you might infer that the requirements for a group isomorphism are stricter than those for a monoid isomorphism. After all, you have to preserve not only identities and the operation, but inverses as well.

In fact, the additional structure of groups allows us to have *fewer* requirements for a group homomorphism.

Group isomorphisms

Definition 4.1. Let $(G,\times)$ and $(H,+)$ be groups. If there exists a function $f : G \to H$ that preserves the operation, which is to say that

$$f(xy) = f(x) + f(y) \quad \text{for every } x, y \in G,$$

then we call f a **group homomorphism**.

This definition requires the preservation of neither inverses nor identities! You might conclude from this that group homomorphism aren't even monoid homomorphisms; we will see in a moment that this is quite untrue!

Notation 4.2. As with monoids, you have to be careful with the fact that different groups have different operations. Depending on the context, the proper way to describe the homomorphism property may be

- $f(xy) = f(x) + f(y)$;
- $f(x+y) = f(x)f(y)$;
- $f(x \circ y) = f(x) \odot f(y)$;
- etc.

Example 4.3. A trivial example of a homomorphism, but an important one, is the identity function $\iota : G \to G$ by $\iota(g) = g$ for all $g \in G$. It should be clear that this is a homomorphism, since for all $g, h \in G$ we have

$$\iota(gh) = gh = \iota(g)\iota(h).$$

For a non-trivial homomorphism, let $f : \mathbb{Z} \to 2\mathbb{Z}$ by $f(x) = 4x$. Then f is a group homomorphism, since for any $x \in \mathbb{Z}$ we have

$$f(x) + f(y) = 4x + 4y = 4(x+y) = f(x+y).$$

Hopefully, the homomorphism property reminds you of certain special functions and operations that you studied in Linear Algebra or Calculus. Recall from Exercise 2.29 that $\mathbb{R}^+$, the set of all positive real numbers, is a multiplicative group.

Example 4.4. Let $f : (\mathrm{GL}_m(\mathbb{R}), \times) \to (\mathbb{R}\setminus\{0\}, \times)$ by $f(A) = \det A$. By Theorem 0.82, $\det A \cdot \det B = \det(AB)$. Thus

$$f(A) \cdot f(B) = \det A \cdot \det B = \det A \cdot \det B = \det(AB) = f(AB),$$

implying that f is a homomorphism of groups.

Let's look at a clockwork group.

Example 4.5. Let $n \in \mathbb{Z}$ such that $n > 1$, and let $f : (\mathbb{Z}, +) \to (\mathbb{Z}_n, +)$ by the assignment $f(x) = [x]_n$. We claim that f is a homomorphism. *Why?* From Lemma 3.83, we know that for any $x, y \in \mathbb{Z}_n$, $f(x+y) = [x+y]_n = [x]_n + [y]_n = f(x) + f(y)$.

By preserving the operation, we preserve an enormous amount of information about a group. If there is a homomorphism f from G to H, then elements of the **image of** G,

$$f(G) = \{h \in H : \exists g \in G \text{ such that } f(g) = h\}$$

act the same way as their **preimages** in G.

This does *not* imply that the *group structure* is the same. In Example 4.5, for example, f is a homomorphism from an infinite group to a finite group; even if the group operations behave in a similar way, the groups themselves are inherently different. If we can show that the groups have the same "size" in addition to a similar operation, then the groups are, for all intents and purposes, identical.

How do we decide that two groups have the same size? For finite groups, this is "easy": count the elements. We can't do that for infinite groups, so we need something a little more general.[13]

[13]The standard method in set theory of showing that two sets are the same "size" is to show that there exists a one-

Definition 4.6. Let $f : G \to H$ be a homomorphism of groups. If f is also a bijection, then we say that G is **isomorphic** to H, write $G \cong H$, and call f an **isomorphism**.

Example 4.7. Recall the homomorphisms of Example 4.3,

$$\iota : G \to G \quad \text{by} \quad \iota(g) = g \qquad \text{and} \qquad f : \mathbb{Z} \to 2\mathbb{Z} \quad \text{by} \quad f(x) = 4x.$$

First we show that ι is an isomorphism. We already know it's a homomorphism, so we need only show that it's a bijection.

one-to-one: Let $g, h \in G$. Assume that $\iota(g) = \iota(h)$. By definition of ι, $g = h$. Since g and h were arbitrary in G, ι is one-to-one.

onto: Let $g \in G$. We need to find $x \in G$ such that $\iota(x) = g$. Using the definition of ι, $x = g$ does the job. Since g was arbitrary in G, ι is onto.

Now we show that f is not a bijection, and hence not an isomorphism.

***not** onto:* There is no element $a \in \mathbb{Z}$ such that $f(a) = 2$. If there were, $4a = 2$. The only possible solution to this equation is $a = 1/2 \notin \mathbb{Z}$.

This is despite the fact that f is one-to-one:

one-to-one: Let $a, b \in \mathbb{Z}$. Assume that $f(a) = f(b)$. By definition of f, $4a = 4b$. Then $4(a-b) = 0$; by the zero product property of the integers, $4 = 0$ or $a - b = 0$. Since $4 \neq 0$, we must have $a - b = 0$, or $a = b$. We assumed $f(a) = f(b)$ and showed that $a = b$. Since a and b were arbitrary, f is one-to-one.

Example 4.8. Recall the homomorphism of Example 4.4,

$$f : \mathrm{GL}_m(\mathbb{R}) \to \mathbb{R}^+ \quad \text{by} \quad f(A) = |\det A|.$$

We claim that f is onto, but not one-to-one.

That f is not one-to-one: Observe that f maps both of the following two diagonal matrices to 2, even though the matrices are unequal:

$$A = \begin{pmatrix} 2 & & & \\ & 1 & & \\ & & 1 & \\ & & & \ddots \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 1 & & & & \\ & 2 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & \ddots \end{pmatrix}.$$

(Unmarked entries are zeroes.)

That f is onto: Let $x \in \mathbb{R}^+$; then $f(A) = x$ where A is the diagonal matrix

$$A = \begin{pmatrix} x & & & \\ & 1 & & \\ & & 1 & \\ & & & \ddots \end{pmatrix}.$$

to-one, onto function between the sets. For example, one can use this definition to show that $\mathbb{Z}$ and $\mathbb{Q}$ are the same size, but $\mathbb{Z}$ and $\mathbb{R}$ are not. So an isomorphism is a homomorphism that also shows that two sets are the same size.

(Again, unmarked entries are zeroes.)

We *cannot* conclude from these examples that $\mathbb{Z} \ncong 2\mathbb{Z}$ and that $\mathbb{R}^+ \ncong \mathbb{R}^{m\times n}$. *Why not?* In each case, we were considering only one of the (possibly many) homomorphisms. It is quite possible that we could find different homomorphisms that *would* be bijections, showing that $\mathbb{Z} \cong 2\mathbb{Z}$ and that $\mathbb{R}^+ \cong \mathbb{R}^{m\times n}$. The first assertion is in fact true, while the second is not; you will explain why in the exercises.

Properties of group homorphism

We turn now to three important properties of group *homomorphism*. For the rest of this section, we assume that $(G,\times)$ and $(H,\circ)$ are groups. Notice that the operations are both "multiplicative".

We still haven't explored the relationship between group homomorphisms and monoid homomorphisms. If a group homomorphism has fewer criteria, can it actually guarantee more structure? Theorem 4.9 answers in the affirmative.

Theorem 4.9. Let $f : G \to H$ be a homomorphism of groups. Denote the identity of G by e_G, and the identity of H by e_H. Then f
preserves identities: $f(e_G) = e_H$; and
preserves inverses: for every $x \in G$, $f\left(x^{-1}\right) = f(x)^{-1}$.

Read the proof below carefully, and identify precisely why this theorem holds for groups, but not for monoids.

Proof. *That f preserves identities:* Let $x \in G$, and $y = f(x)$. By the property of homomorphisms,

$$e_H y = y = f(x) = f(e_G x) = f(e_G) f(x) = f(e_G) y.$$

By the transitive property of equality,

$$e_H y = f(e_G) y.$$

Multiply both sides of the equation *on the right* by y^{-1} to obtain

$$e_H = f(e_G).$$

This shows that f, an arbitrary homomorphism of arbitrary groups, maps the identity of the domain to the identity of the range.

That f preserves inverses: Let $x \in G$. By the property of homomorphisms and by the fact that f preserves identity,

$$e_H = f(e_G) = f\left(x \cdot x^{-1}\right) = f(x) \cdot f\left(x^{-1}\right).$$

Thus

$$e_H = f(x) \cdot f\left(x^{-1}\right).$$

Pay careful attention to what this equation says! "The product of $f(x)$ and $f\left(x^{-1}\right)$ is the identity," which means that those two elements must be inverses! Hence, $f\left(x^{-1}\right)$ is the inverse of $f(x)$, which we write as

$$f\left(x^{-1}\right) = f(x)^{-1}.$$

□

The trick, then, is that the property of inverses guaranteed to groups allows us to do more than we can do in a monoid. In this case, *more* structure in the group led to *fewer* conditions for equivalence. This is not true in general; we we discuss rings, we will see that more structure can lead to more conditions.

If homomorphisms preserve the inverse after all, it makes sense that "the inverse of the image is the image of the inverse." Corollary 4.10 affirms this.

Corollary 4.10. Let $f : G \to H$ be a homomorphism of groups. Then $f\left(x^{-1}\right)^{-1} = f(x)$ for every $x \in G$.

Proof. You do it! See Exercise 4.23. □

It will probably not surprise you that homomorphisms preserve powers of an element.

Theorem 4.11. Let $f : G \to H$ be a homomorphism of groups. Then f preserves powers of elements of G. That is, if $f(g) = h$, then $f(g^n) = f(g)^n = h^n$.

Proof. You do it! See Exercise 4.28. □

Naturally, if homomorphisms preserve powers of an element, they must also preserve cyclic groups.

Corollary 4.12. Let $f : G \to H$ be a homomorphism of groups. If $G = \langle g \rangle$ is a cyclic group, then $f(g)$ determines f completely. In other words, the image $f(G)$ is a cyclic group, and $f(G) = \langle f(g) \rangle$.

Proof. Assume that $G = \langle g \rangle$; that is, G is cyclic. We have to show that two sets are equal. By definition, for any $x \in G$ we can find $n \in \mathbb{Z}$ such that $x = g^n$.

First we show that $f(G) \subseteq \langle f(g) \rangle$. Let $y \in f(G)$ and choose $x \in G$ such that $y = f(x)$. Since G is a cyclic group generated by g, we can choose $n \in \mathbb{Z}$ such that $x = g^n$. By substitution and Theorem 4.11, $y = f(x) = f(g^n) = f(g)^n$. By definition, $y \in \langle f(g) \rangle$. Since y was arbitrary in $f(G)$, $f(G) \subseteq \langle f(g) \rangle$.

Now we show that $f(G) \supseteq \langle f(g) \rangle$. Let $y \in \langle f(g) \rangle$, and choose $n \in \mathbb{Z}$ such that $y = f(g)^n$. By Theorem 4.11, $y = f(g^n)$. Since $g^n \in G$, $f(g^n) \in f(G)$, so $y \in f(G)$. Since y was arbitrary in $\langle f(g) \rangle$, $f(G) \supseteq \langle f(g) \rangle$.

We have shown that $f(G) \subseteq \langle f(g) \rangle$ and $f(G) \supseteq \langle f(g) \rangle$. By equality of sets, $f(G) = \langle f(g) \rangle$. □

The final property of homomorphism that we check here is an important algebraic property of functions; it should remind you of a topic in Section 0.3. It will prove important in subsequent sections and chapters.

Definition 4.13. Let G and H be groups, and $f : G \to H$ a homomorphism. Let

$$Z = \{g \in G : f(g) = e_H\};$$

that is, Z is the set of all elements of G that f maps to the identity of H. We call Z the **kernel** of f, written $\ker f$.

Theorem 4.14. Let $f : G \to H$ be a homomorphism of groups. Then $\ker f \triangleleft G$.

Proof. You do it! See Exercise 4.25. □

Exercises.

Exercise 4.15.
(a) Show that $f : \mathbb{Z} \to 2\mathbb{Z}$ by $f(x) = 2x$ is an isomorphism. Hence $\mathbb{Z} \cong 2\mathbb{Z}$.
(b) Show that $\mathbb{Z} \cong n\mathbb{Z}$ for every nonzero integer n.

Exercise 4.16. Let $n \geq 1$ and $f : \mathbb{Z} \longrightarrow \mathbb{Z}_n$ by $f(a) = [a]_n$.
(a) Show that f is a homomorphism.
(b) Explain why f cannot possibly be an isomorphism.
(c) Determine $\ker f$. (It might help to use a specific value of n first.)
(d) Indicate how we know that $\mathbb{Z}/\ker f \cong \mathbb{Z}_n$. (Eventually, we will show that $G/\ker f \cong H$ for *any* homomorphism $f : G \longrightarrow H$ that is onto.)

Exercise 4.17. Show that $\mathbb{Z}_2$ is isomorphic to the group of order two from Example 2.9 on page 60. *Caution!* Remember to denote the operations properly: $\mathbb{Z}_2$ is additive, but we used $\circ$ for the operation of the group of order two.

Exercise 4.18. Show that $\mathbb{Z}_2$ is isomorphic to the Boolean xor group of Exercise 2.21 on page 64. *Caution!* Remember to denote the operation in the Boolean xor group correctly.

Exercise 4.19. Show that $\mathbb{Z}_n \cong \Omega_n$ for $n \in \mathbb{N}^+$.

Exercise 4.20. Suppose we try to define $f : Q_8 \longrightarrow \Omega_4$ by $f(\mathbf{i}) = f(\mathbf{j}) = f(\mathbf{k}) = i$, and $f(\mathbf{xy}) = f(\mathbf{x})\,f(\mathbf{y})$ for all other $\mathbf{x}, \mathbf{y} \in Q_8$. Show that f is *not* a homomorphism.

Exercise 4.21. Show that $\mathbb{Z}$ is isomorphic to $\mathbb{Z}_0$. (Because of this, people generally don't pay attention to $\mathbb{Z}_0$. See also Exercise 3.87 on page 119.)

Exercise 4.22. Recall the subgroup L of $\mathbb{R}^2$ from Exercises 3.16 on page 98, 3.34 on page 103, and 3.68 on page 114. Show that $L \cong \mathbb{R}$.

Exercise 4.23. Prove Corollary 4.10.

Exercise 4.24. Suppose f is an isomorphism. How many elements does $\ker f$ contain?

Claim: $\ker\varphi \triangleleft G$.
Proof:

1. By _____, it suffices to show that for any $g \in G$, $\ker\varphi = g\,(\ker\varphi)\,g^{-1}$. So, let $g \in$_____.
2. First we show that $(\ker\varphi) \supseteq g\,(\ker\varphi)\,g^{-1}$. Let $x \in g\,(\ker\varphi)\,g^{-1}$.
   (a) By _____, there exists $k \in \ker\varphi$ such that $x = gkg^{-1}$.
   (b) By _____, $\varphi(x) = \varphi\left(gkg^{-1}\right)$.
   (c) By _____, $\varphi(x) = \varphi(g)\,\varphi(k)\,\varphi(g)^{-1}$.
   (d) By _____, $\varphi(x) = \varphi(g)\,e_H\,\varphi(g)^{-1}$.
   (e) By _____, $\varphi(x) = e_H$.
   (f) By definition of the kernel, _____.
   (g) Since _____, $g\,(\ker\varphi)\,g^{-1} \subseteq \ker\varphi$.
3. Now we show the converse; that is, _____. Let $k \in \ker\varphi$.
   (a) Let $x = g^{-1}kg$. Notice that if $x \in \ker\varphi$, then we would have what we want, since in this case _____.
   (b) In fact, $x \in \ker\varphi$. After all, _____.
   (c) Since _____, $\ker\varphi \subseteq g\,(\ker\varphi)\,g^{-1}$.
4. By _____, $\ker\varphi = g\,(\ker\varphi)\,g^{-1}$.

Figure 4.1. Material for Exercise 4.25

Exercise 4.25. Let G and H be groups, and $\varphi : G \to H$ a homomorphism.

(a) Show that $\ker\varphi < G$.
(b) Fill in each blank of Figure 4.1 with the appropriate justification or statement.

Exercise 4.26. Let φ be a homomorphism from a finite group G to a group H. Recall from Exercise 4.25 that $\ker\varphi \triangleleft G$. Explain why $|\ker\varphi| \cdot |\varphi(G)| = |G|$. (This is sometimes called **the Homomorphism Theorem**.)

Exercise 4.27. Let $f : G \to H$ be an isomorphism. Isomorphisms are by definition one-to-one functions, so f has an inverse function f^{-1}. Show that $f^{-1} : H \to G$ is also an isomorphism.

Exercise 4.28. Prove Theorem 4.11.

Exercise 4.29. Let $f : G \to H$ be a homomorphism of groups. Assume that G is abelian.
(a) Show that $f(G)$ is abelian.
(b) Is H abelian? Explain why or why not.

Exercise 4.30. Let $f : G \to H$ be a homomorphism of groups. Let $A < G$. Show that $f(A) < H$.

Exercise 4.31. Let $f : G \to H$ be a homomorphism of groups. Let $A \triangleleft G$.
(a) Show that $f(A) \triangleleft f(G)$.
(b) Do you think that $f(A) \triangleleft H$? Justify your answer.

Exercise 4.32. Show that if G is a group, then $G/\{e\} \cong G$ and $G/G \cong \{e\}$.

Exercise 4.33. Recall the orthogonal group and the special orthogonal group from Exercise 3.22. Let $\varphi : O(n) \to \Omega_2$ by $\varphi(A) = \det A$.

(a) Show that φ is a homomorphism, but not an isomorphism.
(b) Explain why $\ker\varphi = \mathrm{SO}(n)$.

Exercise 4.34. In Chapter 1, the definition of an isomorphism for *monoids* required that the function map the identity to the identity (Definition 1.26 on page 45). By contrast, Theorem 4.9 shows that the preservation of the operation guarantees that a group homomorphism maps the identity to the identity, so we don't need to require this in the definition of an isomorphism for *groups* (Definition 4.6).

The difference between a group and a monoid is the existence of an inverse. Use this to show that, in a monoid, you *can* have a function that preserves the operation, but not the identity. In other words, show that Theorem 4.9 is false for monoids.

4.2: Consequences of isomorphism

Throughout this section, $(G,\times)$ and $(H,\circ)$ are groups.

The purpose of this section is to show why we use the name *isomorphism:* if two groups are isomorphic, then they are indistinguishable *as groups*. The elements of the sets are different, and the operation may be defined differently, but as groups the two are identical. Suppose that two groups G and H are isomorphic. We will show that

- isomorphism is an equivalence relation;
- G is abelian iff H is abelian;
- G is cyclic iff H is cyclic;
- every subgroup A of G corresponds to a subgroup A' of H (in particular, if A is of order n, so is A');
- every normal subgroup N of G corresponds to a normal subgroup N' of H;
- the quotient group G/N corresponds to a quotient group H/N'.

All of these depend on the existence of an isomorphism $f : G \to H$. In particular, uniqueness is guaranteed only for any one isomorphism; if two different isomorphisms f, f' exist between G and H, then a subgroup A of G may well correspond to two distinct subgroups B and B' of H.

Isomorphism is an equivalence relation

The fact that isomorphism is an equivalence relation will prove helpful with the equivalence properties; for example, "G is cyclic iff H is cyclic." So, we start with that one first.

Theorem 4.35. Isomorphism is an equivalence relation. That is, $\cong$ satisfies the reflexive, symmetric, and transitive properties.

Proof. First we show that $\cong$ is reflexive. Let G be any group, and let ι be the identity homomorphism from Example 4.3. We showed in Example 4.7 that ι is an isomorphism. Since $\iota : G \to G$, $G \cong G$. Since G was an arbitrary group, $\cong$ is reflexive.

Next, we show that $\cong$ is symmetric. Let G, H be groups and assume that $G \cong H$. By definition, there exists an isomorphism $f : G \to H$. By Exercise 4.27, f^{-1} is also a isomorphism. Hence $H \cong G$.

Finally, we show that $\cong$ is transitive. Let G, H, K be groups and assume that $G \cong H$ and $H \cong K$. By definition, there exist isomorphisms $f : G \to H$ and $g : H \to K$. Define $h : G \to K$ by

$$h(x) = g(f(x)).$$

We claim that h is an isomorphism. We show each requirement in turn:

That h is a homomorphism, let $x, y \in G$. By definition of h, $h(x \cdot y) = g(f(x \cdot y))$. Applying the fact that g and f are both homomorphisms,

$$h(x \cdot y) = g(f(x \cdot y)) = g(f(x) \cdot f(y)) = g(f(x)) \cdot g(f(y)) = h(x) \cdot h(y).$$

Thus h is a homomorphism.

That h is one-to-one, let $x, y \in G$ and assume that $h(x) = h(y)$. By definition of h,

$$g(f(x)) = g(f(y)).$$

By hypothesis, g is an isomorphism, so by definition it is one-to-one, so if its outputs are equal, so are its inputs. In other words,

$$f(x) = f(y).$$

Similarly, f is an isomorphism, so $x = y$. Since x and y were arbitrary in G, h is one-to-one.

That h is onto, let $z \in K$. We claim that there exists $x \in G$ such that $h(x) = z$. Since g is an isomorphism, it is by definition onto, so there exists $y \in H$ such that $g(y) = z$. Since f is an isomorphism, there exists $x \in G$ such that $f(x) = y$. Putting this together with the definition of h, we see that

$$z = g(y) = g(f(x)) = h(x).$$

Since z was arbitrary in K, h is onto.

We have shown that h is a one-to-one, onto homorphism. Thus h is an isomorphism, and $G \cong K$. □

Isomorphism preserves basic properties of groups

We now show that isomorphism preserves two basic properties of groups that we introduced in Chapter 2: abelian and commutative. Both proofs make use of the fact that isomorphism is an equivalence relation; in particular, that the relation is symmetric.

Theorem 4.36. Suppose that $G \cong H$. Then G is abelian iff H is abelian.

Proof. Let $f : G \to H$ be an isomorphism. Assume that G is abelian. We must show that H is abelian. By Exercise 4.29, $f(G)$ is abelian. Since f is an isomorphism, and therefore onto, $f(G) = H$. Hence H is abelian.

We turn to the converse. Assume that H is abelian. Since isomorphism is symmetric, $H \cong G$. Along with the above argument, this implies that if H is abelian, then G is, too.

Hence, G is abelian iff H is abelian. □

Theorem 4.37. Suppose $G \cong H$. Then G is cyclic iff H is cyclic.

Proof. Let $f : G \to H$ be an isomorphism. Assume that G is cyclic. We must show that H is cyclic; that is, we must show that every element of H is generated by a fixed element of H.

Since G is cyclic, by definition $G = \langle g \rangle$ for some $g \in G$. Let $h = f(g)$; then $h \in H$. We claim that $H = \langle h \rangle$.

Let $x \in H$. Since f is an isomorphism, it is onto, so there exists $a \in G$ such that $f(a) = x$. Since G is cyclic, there exists $n \in \mathbb{Z}$ such that $a = g^n$. By Theorem 4.11,

$$x = f(a) = f(g^n) = f(g)^n = h^n.$$

Since x was an arbitrary element of H and x is generated by h, all elements of H are generated by h. Hence $H = \langle h \rangle$ is cyclic.

Since isomorphism is symmetric, $H \cong G$. Along with the above argument, this implies that if H is cyclic, then G is, too.

Hence, G is cyclic iff H is cyclic. □

Isomorphism preserves the structure of subgroups

Theorem 4.38. Suppose $G \cong H$. Every subgroup A of G is isomorphic to a subgroup B of H. Moreover, each of the following holds.
(A) $|A|$ iff $|B|$.
(B) A is normal iff B is normal.

Proof. Let $f : G \to H$ be an isomorphism. Let A be a subgroup of G. By Exercise 4.30, $f(A) < H$.

We claim that f is one-to-one and onto from A to $f(A)$. Onto is immediate from the definition of $f(A)$. The one-to-one property holds because f is one-to-one in G and $A \subseteq G$. We have shown that $f(A) < H$ and that f is one-to-one and onto from A to $f(A)$. Hence $A \cong f(A)$.

Claim (A) follows from the fact that f is a bijection: this is the definition of when two sets have equal size.

For claim (B), assume $A \triangleleft G$. We want to show that $B \triangleleft H$; that is, $xB = Bx$ for every $x \in H$. Let $x \in H$ and $y \in B$; since f is an isomorphism, it is onto, so $f(g) = x$ and $f(a) = y$ for some $g \in G$ and some $a \in A$. By substitution and the homomorphism property,

$$xy = f(g) f(a) = f(ga).$$

Since $A \triangleleft G$, $gA = Ag$, so there exists $a' \in A$ such that $ga = a'g$. Let $y' = f(a')$. By substitution and the homomorphism property,

$$xy = f(a'g) = f(a') f(g) = y'x.$$

By definition and substitution, we have $y' = f(a') \in f(A) = B$. We conclude that, $xy = y'x \in Bx$.

We have shown that for arbitrary $x \in H$ and arbitrary $y \in B$, there exists $y' \in B$ such that $xy = y'x$. Hence $xB \subseteq Bx$. A similar argument shows that $xB \supseteq Bx$, so $xB = Bx$. This is the definition of a normal subgroup, so $B \triangleleft H$.

Since isomorphism is symmetric, $B \cong A$. Along with the above argument, this implies that if $B \triangleleft H$, then $A \triangleleft G$, as well.

Hence, A is normal iff B is normal. □

Theorem 4.39. Suppose $G \cong H$ as groups. Every quotient group of G is isomorphic to a quotient group of H.

We use Lemma 3.29(CE3) on page 102 on coset equality heavily in this proof; you may want to go back and review it.

Proof. Let $f : G \to H$ be an isomorphism. Consider an arbitrary quotient group of G defined as G/A, where $A \triangleleft G$. Let $B = f(A)$; by Theorem 4.38 $B \triangleleft H$, so H/B is a quotient group. We want to show that $G/A \cong H/B$.

To that end, define a new function $f_A : G/A \to H/B$ by

$$f_A(X) = f(g)B \qquad \text{where} \qquad X = gA \in G/A.$$

Keep in mind that f_A maps cosets to cosets, using the relation f from group elements to group elements.

We claim that f_A is an isomorphism. You probably expect that we "only" have to show that f_A is a bijection and a homomorphism, but this is not true. We have to show first that f_A is well-defined. Do you remember what this means? If not, reread page 107. Once you understand the definition, ask yourself, why do we have to show f_A is well-defined?

Just we must define the operation for cosets to give the same result regardless of two cosets' representation, a function on cosets must give the same result regardless of that coset's representation. Let X be any coset in G/A. It is usually the case that X can have more than one representation; that is, we can find $g \neq \widehat{g}$ where $X = gA = \widehat{g}A$. For example, suppose you want to build a function from $\mathbb{Z}_5$ to another set. Suppose that we want $f([2]) = x$. Recall that in $\mathbb{Z}_5$, $\cdots = [-3] = [2] = [7] = [12] = \cdots$. If f is defined in such a way that we would think $f([-3]) \neq x$, we would have a problem, since we need to ensure that $f([-3]) = f([2])$! For another example, consider D_3. We know that $\varphi A_3 = (\rho\varphi)A_3$, even though $\varphi \neq \rho\varphi$; see Example 3.57 on page 109. If $f(g) \neq f(\widehat{g})$, then $f_A(X)$ would have more than one possible value, since

$$f_A(X) = f_A(gA) = f(g) \neq f(\widehat{g}) = f_A(\widehat{g}A) = f(X).$$

In other words, f_A would not be a function, since at least one element of the domain (X) would correspond to at least two elements of the range ($f(g)$ and $f(\widehat{g})$). See Figure 4.2. A homomorphism must first be a function, so if f_A is not even a function, then it is not well-defined.

That f_A is well-defined: Let $X \in G/A$ and consider two representations g_1A and g_2A of X. Let $Y_1 = f_A(g_1A)$ and $Y_2 = f_A(g_2A)$. By definition of f_A,

$$Y_1 = f(g_1)B \quad \text{and} \quad Y_2 = f(g_2)B.$$

To show that f_A is well-defined, we must show that $Y_1 = Y_2$. By hypothesis, $g_1A = g_2A$. Lemma 3.29(CE3) implies that $g_2^{-1}g_1 \in A$. Recall that $f(A) = B$; by definitino of the image,

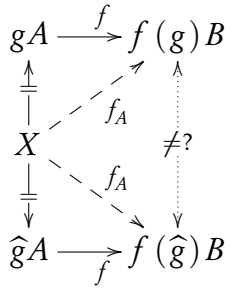


Figure 4.2. When defining a mapping whose domain is a quotient group, we must be careful to ensure that a coset with different representations has the same value. In the diagram above, X has the two representations gA and $\widehat{g}A$, and f_A is defined using f. Inb this case, is $f(g) = f(\widehat{g})$? If not, then $f_A(X)$ would have two different values, and f_A would not be a function.

$f\left(g_2^{-1}g_1\right) \in B$. The homomorphism property implies that

$$f(g_2)^{-1} f(g_1) = f\left(g_2^{-1}\right) f(g_1) = f\left(g_2^{-1}g_1\right) \in B.$$

Lemma 3.29(CE3) again implies that $f(g_1)B = f(g_2)B$, or $Y_1 = Y_2$, so there is no ambiguity in the definition of f_A as the image of X in H/B; the function is well-defined.

That f_A is a homomorphism: Let $X, Y \in G/A$ and write $X = g_1A$ and $Y = g_2A$ for appropriate $g_1, g_2 \in G$. Now

$$\begin{aligned}
f_A(XY) &= f_A((g_1A)\cdot(g_2A)) && \text{(substitution)}\\
&= f_A(g_1g_2\cdot A) && \text{(coset multiplication in } G/A)\\
&= f(g_1g_2)B && \text{(definition of } f_A)\\
&= (f(g_1)f(g_2))\cdot B && \text{(homomorphism property)}\\
&= f(g_1)A'\cdot f(g_2)B && \text{(coset multiplication in } H/B)\\
&= f_A(g_1A)\cdot f_A(g_2A) && \text{(definition of } f_A)\\
&= f_A(X)\cdot f_A(Y) && \text{(substitution).}
\end{aligned}$$

By definition, f_A is a homomorphism.

That f_A is one-to-one: Let $X, Y \in G/A$ and assume that $f_A(X) = f_A(Y)$. Let $g_1, g_2 \in G$ such that $X = g_1A$ and $Y = g_2A$. The definition of f_A implies that

$$f(g_1)B = f_A(X) = f_A(Y) = f(g_2)B,$$

so by Lemma 3.29(CE3) $f(g_2)^{-1} f(g_1) \in B$. Recall that $B = f(A)$, so there exists $a \in A$ such that $f(a) = f(g_2)^{-1} f(g_1)$. The homomorphism property implies that

$$f(a) = f\left(g_2^{-1}\right) f(g_1) = f\left(g_2^{-1}g_1\right).$$

Recall that f is an isomorphism, hence one-to-one. The definition of one-to-one implies that

$$g_2^{-1} g_1 = a \in A.$$

Applying Lemma 3.29(CE3) again gives us $g_1 A = g_2 A$, and

$$X = g_1 A = g_2 A = Y.$$

We took arbitrary $X, Y \in G/A$ and showed that if $f_A(X) = f_A(Y)$, then $X = Y$. It follows that f_A is one-to-one.

That f_A is onto: You do it! See Exercise 4.40. □

Exercises.

Exercise 4.40. Show that the function f_A defined in the proof of Theorem 4.39 is onto.

Exercise 4.41. Recall from Exercise 2.85 on page 93 that $\langle \mathbf{i} \rangle$ is a cyclic group of Q_8.

(a) Show that $\langle \mathbf{i} \rangle \cong \mathbb{Z}_4$ by giving an explicit isomorphism.
(b) Let A be a proper subgroup of $\langle \mathbf{i} \rangle$. Find the corresponding subgroup of $\mathbb{Z}_4$.
(c) Use the proof of Theorem 4.39 to determine the quotient group of $\mathbb{Z}_4$ to which $\langle \mathbf{i} \rangle / A$ is isomorphic.

Exercise 4.42. Recall from Exercise 4.22 on page 130 that the set

$$L = \left\{ x \in \mathbb{R}^2 : \; x = (a, a) \; \exists a \in \mathbb{R} \right\}$$

defined in Exercise 3.16 on page 98 is isomorphic to $\mathbb{R}$.

(a) Show that $\mathbb{Z} \triangleleft \mathbb{R}$.
(b) Give the precise definition of $\mathbb{R}/\mathbb{Z}$.
(c) Explain why we can think of $\mathbb{R}/\mathbb{Z}$ as the set of classes $[a]$ such that $a \in [0, 1)$. Choose one such $[a]$ and describe the elements of this class.
(d) Find the subgroup H of L that corresponds to $\mathbb{Z} < \mathbb{R}$. What do this section's theorems imply that you can conclude about H and L/H?
(e) Use the homomorphism f_A defined in the proof of Theorem 4.39 to find the images $f_{\mathbb{Z}}(\mathbb{Z})$ and $f_{\mathbb{Z}}(\pi + \mathbb{Z})$.
(f) Use the answer to (c) to describe L/H intuitively. Choose an element of L/H and describe the elements of this class.

4.3: The Isomorphism Theorem

In this section, we identify an important relationship between a subgroup $A < G$ that has a special relationship to a homomorphism, and the image of the quotient group $f(G/A)$. First, an example.

Motivating example

Example 4.43. Recall $A_3 = \{\iota, \rho, \rho^2\} \triangleleft D_3$ from Example 3.57. We saw that D_3/A_3 has only two elements, so it must be isomorphic to any group of two elements. First we show this explicitly: Let $\mu : D_3/A_3 \to \mathbb{Z}_2$ by

$$\mu(X) = \begin{cases} 0, & X = A_3; \\ 1, & \text{otherwise.} \end{cases}$$

Is μ a homomorphism? Recall that A_3 is the identity element of D_3/A_3, so for any $X \in D_3/A_3$

$$\mu(X \cdot A_3) = \mu(X) = \mu(X) + 0 = \mu(X) + \mu(A_3).$$

This verifies the homomorphism property for all products in the Cayley table of D_3/A_3 except $(\varphi A_3) \cdot (\varphi A_3)$, which is easy to check:

$$\mu((\varphi A_3) \cdot (\varphi A_3)) = \mu(A_3) = 0 = 1 + 1 = \mu(\varphi A_3) + \mu(\varphi A_3).$$

Hence μ is a homomorphism. The property of isomorphism follows from the facts that

- $\mu(A_3) \neq \mu(\varphi A_3)$, so μ is one-to-one, and
- both 0 and 1 have preimages, so μ is onto.

Notice further that $\ker \mu = A_3$.

Something subtle is at work here. Let $f : D_3 \to \mathbb{Z}_2$ by

$$f(x) = \begin{cases} 0, & x \in A_3; \\ 1, & \text{otherwise.} \end{cases}$$

Is f a homomorphism? The elements of A_3 are ι, ρ, and ρ^2; f maps these elements to zero, and the other three elements of D_3 to 1. Let $x, y \in D_3$ and consider the various cases:

Case 1. Suppose first that $x, y \in A_3$. Since A_3 is a group, closure implies that $xy \in A_3$. Thus

$$f(xy) = 0 = 0 + 0 = f(x) + f(y).$$

Case 2. Next, suppose that $x \in A_3$ and $y \notin A_3$. Since A_3 is a group, closure implies that $xy \notin A_3$. (Otherwise $xy = z$ for some $z \in A_3$, and multiplication by the inverse implies that $y = x^{-1}z \in A_3$, a contradiction.) Thus

$$f(xy) = 1 = 0 + 1 = f(x) + f(y).$$

Case 3. If $x \notin A_3$ and $y \in A_3$, then a similar argument shows that $f(xy) = f(x) + f(y)$.

Case 4. Finally, suppose $x, y \notin A_3$. Inspection of the Cayley table of D_3 (Exercise 2.45 on page 74) shows that $xy \in A_3$. Hence

$$f(xy) = 0 = 1 + 1 = f(x) + f(y).$$

We have shown that f is a homomorphism from D_3 to $\mathbb{Z}_2$. Again, $\ker f = A_3$.

In addition, consider the function $\eta : D_3 \to D_3/A_3$ by

$$\eta(x) = \begin{cases} A_3, & x \in A_3; \\ \varphi A_3, & \text{otherwise.} \end{cases}$$

It is easy to show that this is a homomorphism; we do so presently.

Now comes the important observation: Look at the composition function $\eta \circ \mu$ whose domain is D_3 and whose range is $\mathbb{Z}_2$:

$$\begin{aligned}
(\mu \circ \eta)(\iota) &= \mu(\eta(\iota)) = \mu(A_3) = 0;\\
(\mu \circ \eta)(\rho) &= \mu(\eta(\rho)) = \mu(A_3) = 0;\\
(\mu \circ \eta)\left(\rho^2\right) &= \mu\left(\eta\left(\rho^2\right)\right) = \mu(A_3) = 0;\\
(\mu \circ \eta)(\varphi) &= \mu(\eta(\varphi)) = \mu(\varphi A_3) = 1;\\
(\mu \circ \eta)(\rho\varphi) &= \mu(\eta(\rho\varphi)) = \mu(\varphi A_3) = 1;\\
(\mu \circ \eta)\left(\rho^2\varphi\right) &= \mu\left(\eta\left(\rho^2\varphi\right)\right) = \mu(\varphi A_3) = 1.
\end{aligned}$$

We have

$$(\mu \circ \eta)(x) = \begin{cases} 0, & x \in A_3;\\ 1, & \text{otherwise,} \end{cases}$$

or in other words

$$\mu \circ \eta = f.$$

In words, f is the composition of a "natural" mapping between D_3 and D_3/A_3, and the isomorphism from D_3/A_3 to $\mathbb{Z}_2$. But another way of looking at this is that the isomorphism μ is related to f and the "natural" homomorphism.

The Isomorphism Theorem

This remarkable correspondence can make it easier to study quotient groups G/A:

- find a group H that is "easy" to work with; and
- find a homomorphism $f : G \to H$ such that
  - $f(g) = e_H$ for all $g \in A$, and
  - $f(g) \neq e_H$ for all $g \notin A$.

If we can do this, then $H \cong G/A$, and as we saw in Section 4.2 studying G/A is equivalent to studying H.

The reverse is also true: suppose that a group G and its quotient groups are relatively easy to study, whereas another group H is difficult. The isomorphism theorem helps us identify a quotient group G/A that is isomorphic to H, making it easier to study.

Another advantage, which we realize later in the course, is that computation in G can be difficult or even impossible, while computation in G/A can be quite easy. This turns out to be the case with $\mathbb{Z}$ when the coefficients grow too large; we will work in $\mathbb{Z}_p$ for several values of p, and reconstruct the correct answers.

We need to formalize this observation in a theorem, but first we have to confirm something that we claimed earlier:

Lemma 4.44. Let G be a group and $A \triangleleft G$. The function $\eta : G \to G/A$ by

$$\eta(g) = gA$$

is a homomorphism.

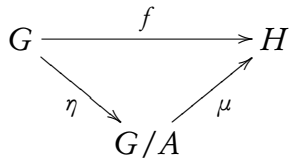
Proof. You do it! See Exercise 4.47. □

Definition 4.45. We call the homomorphism η of Lemma 4.44 the **natural homomorphism** from G to G/A.

What's special about A_3 in the example that began this section? Of course, A_3 is a normal subgroup of D_3, but something you might not have noticed is that it was the kernel of f. We use this to formalize the observation of Example 4.43.

Theorem 4.46 (The Isomorphism Theorem)**.** Let G and H be groups, $f : G \to H$ a homomorphism that is onto, and $\ker f = A$. Then $G/A \cong H$, and the isomorphism $\mu : G/A \to H$ satisfies $f = \mu \circ \eta$, where $\eta : G \to G/A$ is the natural homomorphism.

We can illustrate Theorem 4.46 by the following diagram:



The idea is that "the diagram commutes", or $f = \mu \circ \eta$.

Proof. We are given G, H, f and A. Define $\mu : G/A \to H$ in the following way:

$$\mu(X) = f(g), \text{ where } X = gA.$$

We claim that μ is an isomorphism from G/A to H, and moreover that $f = \mu \circ \eta$.

Since the domain of μ consists of cosets which may have different representations, we must show first that μ is well-defined. Suppose that $X \in G/A$ has two representations $X = gA = g'A$ where $g, g' \in G$ and $g \neq g'$. We need to show that $\mu(gA) = \mu(g'A)$. From Lemma 3.29(CE3), we know that $g^{-1}g' \in A$, so there exists $a \in A$ such that $g^{-1}g' = a$, so $g' = ga$. Applying the definition of μ and the homomorphism property,

$$\mu(g'A) = f(g') = f(ga) = f(g) f(a).$$

Recall that $a \in A = \ker f$, so $f(a) = e_H$. Substitution gives

$$\mu(g'A) = f(g) \cdot e_H = f(g) = \mu(gA).$$

Hence $\mu(g'A) = \mu(gA)$ and $\mu(X)$ is well-defined.

Is μ a homomorphism? Let $X, Y \in G/A$; we can represent $X = gA$ and $Y = g'A$ for some

$g, g' \in G$. We see that

$$\begin{aligned}
\mu(XY) &= \mu\left((gA)\left(g'A\right)\right) && \text{(substitution)}\\
&= \mu\left(\left(gg'\right)A\right) && \text{(coset multiplication)}\\
&= f\left(gg'\right) && \text{(definition of } \mu\text{)}\\
&= f(g) f\left(g'\right) && \text{(homomorphism)}\\
&= \mu(gA)\,\mu\left(g'A\right). && \text{(definiition of } \mu\text{)}
\end{aligned}$$

Thus μ is a homomorphism.

Is μ one-to-one? Let $X, Y \in G/A$ and assume that $\mu(X) = \mu(Y)$. Represent $X = gA$ and $Y = g'A$ for some $g, g' \in G$; we see that

$$\begin{aligned}
f\left(g^{-1}g'\right) &= f\left(g^{-1}\right) f\left(g'\right) && \text{(homomorphism)}\\
&= f(g)^{-1} f\left(g'\right) && \text{(homomorphism)}\\
&= \mu(gA)^{-1}\,\mu\left(g'A\right) && \text{(definition of } \mu\text{)}\\
&= \mu(X)^{-1}\,\mu(Y) && \text{(substitution)}\\
&= \mu(Y)^{-1}\,\mu(Y) && \text{(substitution)}\\
&= e_H, && \text{(inverses)}
\end{aligned}$$

so $g^{-1}g' \in \ker f$. By hypothesis, $\ker f = A$, so $g^{-1}g' \in A$. Lemma 3.29(CE3) now tells us that $gA = g'A$, so $X = Y$. Thus μ is one-to-one.

Is μ onto? Let $h \in H$; we need to find an element $X \in G/A$ such that $\mu(X) = h$. By hypotehesis, f is onto, so there exists $g \in G$ such that $f(g) = h$. By definition of μ and substitution,

$$\mu(gA) = f(g) = h,$$

so μ is onto.

We have shown that μ is an isomorphism; we still have to show that $f = \mu \circ \eta$, but the definition of μ makes this trivial: for any $g \in G$,

$$(\mu \circ \eta)(g) = \mu(\eta(g)) = \mu(gA) = f(g).$$

□

Exercises

Exercise 4.47. Prove Lemma 4.44.

Exercise 4.48. Use Exercise 4.33 to explain why $\Omega_2 \cong \mathrm{O}(n)/\mathrm{SO}(n)$.

Exercise 4.49. Recall the normal subgroup L of $\mathbb{R}^2$ from Exercises 3.16, 3.34, and 3.68 on pages 98, 103, and 114, respectively. In Exercise 4.22 on page 130 you found an explicit isomorphism $L \cong \mathbb{R}$.

(a) Use the Isomorphism Theorem to find an isomorphism $\mathbb{R}^2/L \cong \mathbb{R}$.

(b) Argue from this that $\mathbb{R}^2/\mathbb{R} \cong \mathbb{R}$.

Let G and H be groups, and $A \triangleleft G$.
Claim: If $G/A \cong H$, then there exists a homomorphism $\varphi : G \to H$ such that $\ker\varphi = A$.

1. Assume _____.
2. By hypothesis, there exists f _____.
3. Let $\eta : G \to G/A$ be the natural homomorphism. Define $\varphi : G \to H$ by $\varphi(g) =$_____.
4. By _____, φ is a homomorphism.
5. We claim that $A \subseteq \ker\varphi$. To see why,
   (a) By _____, the identity of G/A is A.
   (b) By _____, $f(A) = e_H$.
   (c) Let $a \in A$. By definition of the natural homomorphism, $\eta(a) =$_____.
   (d) By _____, $f(\eta(a)) = e_H$.
   (e) By _____, $\varphi(a) = e_H$.
   (f) Since _____, $A \subseteq \ker\varphi$.
6. We further claim that $A \supseteq \ker\varphi$. To see why,
   (a) Let $g \in G \backslash A$. By definition of the natural homomorphism, $\varphi(g) \neq$_____.
   (b) By _____, $f(\eta(g)) \neq e_H$.
   (c) By _____, $\varphi(g) \neq e_H$.
   (d) By _____, $g \notin \ker\varphi$.
   (e) Since g was arbitrary in $G \backslash A$, _____.
7. We have shown that $A \subseteq \ker\varphi$ and $A \supseteq \ker\varphi$. By _____, $A = \ker\varphi$.

Figure 4.3. Material for Exercise 4.52

(c) Describe geometrically how the cosets of $\mathbb{R}^2/L$ are mapped to elements of $\mathbb{R}$.

Exercise 4.50. Recall the normal subgroup $\langle -\mathbf{1} \rangle$ of Q_8 from Exercises 2.84 on page 92 and 3.64 on page 112.

(a) Use Lagrange's Theorem to explain why $Q_8/\langle -\mathbf{1} \rangle$ has order 4.
(b) We know from Exercise 2.32 on page 65 that there are only two groups of order 4, the Klein 4-group and the cyclic group of order 4, which we can represent by $\mathbb{Z}_4$. Use the Isomorphism Theorem to determine which of these groups is isomorphic to $Q_8/\langle -\mathbf{1} \rangle$.

Exercise 4.51. Recall the kernel of a monoid homomorphism from Exercise 1.43 on page 50, and that group homomorphisms are also monoid homomorphisms. These two definitions do not look the same, but in fact, one generalizes the other.

(a) Show that if $x \in G$ is in the kernel of a group homomorphism $f : G \to H$ if and only $(x, e) \in \ker f$ when we view f as a monoid homomorphism.
(b) Show that $x \in G$ is in the kernel of a group homomorphism $f : G \to H$ if and only if we can find $y, z \in G$ such that $f(y) = f(z)$ and $y^{-1}z = x$.
(c) Explain how this shows that Exercise 1.43 "lays the groundwork" for a "monoid generalization" of the Isomorphism Theorem.
(d) Formulate and prove a "Monoid Isomorphism Theorem."

Exercise 4.52. Fill in each blank of Figure 4.3 with the appropriate justification or statement.

4.4: Automorphisms and groups of automorphisms

In this section, we use isomorphisms to build a new kind of group, useful for analyzing roots of polynomial equations. We will discuss the applications of these groups in Chapter 9, but they are of independent interest, as well.

Definition 4.53. Let G be a group. If $f : G \to G$ is an isomorphism, then we call f an **automorphism**.

An automorphism[14] is an isomorphism whose domain and range are the same set. Thus, to show that some function f is an automorphism, you must show first that the domain and the range of f are the same set. Afterwards, you show that f satisfies the homomorphism property, and then that it is both one-to-one and onto.

Example 4.54.

(a) An easy automorphism for any group G is the identity isomorphism $\iota(g) = g$:
- its range is by definition G;
- it is a *homomorphism* because $\iota(g \cdot g') = g \cdot g' = \iota(g) \cdot \iota(g')$;
- it is *one-to-one* because $\iota(g) = \iota(g')$ implies (by evaluation of the function) that $g = g'$; and
- it is *onto* because for any $g \in G$ we have $\iota(g) = g$.

(b) An automorphism in $(\mathbb{Z}, +)$ is $f(x) = -x$:
- its range is $\mathbb{Z}$ because of closure;
- it is a *homomorphism* because $f(x+y) = -(x+y) = -x - y = f(x) + f(y)$;
- it is *one-to-one* because $f(x) = f(y)$ implies that $-x = -y$, so $x = y$; and
- it is *onto* because for any $x \in \mathbb{Z}$ we have $f(-x) = x$.

(c) An automorphism in D_3 is $f(x) = \rho^2 x \rho$:
- its range is D_3 because of closure;
- it is a *homomorphism* because $f(xy) = \rho^2(xy)\rho = \rho^2(x \cdot \iota \cdot y)\rho = \rho^2(x \cdot \rho^3 \cdot y)\rho = (\rho^2 x \rho) \cdot (\rho^2 y \rho) = f(x) \cdot f(y)$;
- it is *one-to-one* because $f(x) = f(y)$ implies that $\rho^2 x \rho = \rho^2 y \rho$, and multiplication on the left by ρ and on the right by ρ^2 gives us $x = y$; and
- it is *onto* because for any $y \in D_3$, choose $x = \rho y \rho^2$ and then $f(x) = \rho^2(\rho y \rho^2)\rho = (\rho^2\rho) \cdot y \cdot (\rho^2\rho) = \iota \cdot y \cdot \iota = y$.

The automorphism of Example 4.54(c) generalizes to an important way. Recall the conjugation of one element of a group by another, introduced in Exercise 2.37 on page 66. By fixing the second element, we can turn this into a function on a group.

Definition 4.55. Let G be a group and $a \in G$. Define the function of **conjugation by** a to be $\text{conj}_a(x) = a^{-1}xa$.

In Example 4.54(c), we had $a = \rho$ and $\text{conj}_a(x) = a^{-1}xa = \rho^2 x \rho$.

You have already worked with conjugation in previous exercises, such as showing that it can provide an alternate definition of a normal subgroup (Exercises 2.37 on page 66 and 3.67 on page 113). Beyond that, conjugating a subgroup *always* produces another subgroup:

[14] The word comes Greek words that mean *self* and *shape*.

Lemma 4.56. Let G be a group, and $a \in G$. Then conj_a is an automorphism. Moreover, for any $H < G$,

$$\{\text{conj}_a(h) : h \in H\} < G.$$

Proof. You do it! See Exercise 4.64. □

The subgroup $\{\text{conj}_a(h) : h \in H\}$ is important enough to identify by a special name.

Definition 4.57. Suppose $H < G$, and $a \in G$. We say that $\{\text{conj}_a(h) : h \in H\}$ is the **group of conjugations of H by a**, and denote it by $\text{Conj}_a(H)$.

Conjugation of a subgroup H by an arbitrary $a \in G$ is *not* necessarily an automorphism; there can exist $H < G$ and $a \in G \backslash H$ such that $H \neq \{\text{conj}_a(h) : h \in H\}$. On the other hand, if H is a *normal* subgroup of G, then we *do* have $H = \{\text{conj}_a(h) : h \in H\}$; this property can act as an alternate definition of a normal subgroup. You will explore this in the exercises.

Now it is time to identify the new group that we promised at the beginning of the section.

The automorphism group

Notation 4.58. Write $\text{Aut}(G)$ for the set of all automorphisms of G. We typically denote elements of $\text{Aut}(G)$ by Greek letters ($\alpha, \beta, \ldots$), rather than Latin letters ($f, g, \ldots$).

Example 4.59. We compute $\text{Aut}(\mathbb{Z}_4)$. Let $\alpha \in \text{Aut}(\mathbb{Z}_4)$ be arbitrary; what do we know about α? By definition, its range is $\mathbb{Z}_4$, and by Theorem 4.9 on page 128 we know that $\alpha(0) = 0$. Aside from that, we consider all the possibilities that preserve the isomorphism properties.

Recall from Theorem 3.85 on page 119 that $\mathbb{Z}_4$ is a cyclic group; in fact $\mathbb{Z}_4 = \langle 1 \rangle$. Corollary 4.12 on page 129 tells us that $\alpha(1)$ will tell us everything we want to know about α. So, what can $\alpha(1)$ be?

Case 1. Can we have $\alpha(1) = 0$? If so, then $\alpha(1) = \alpha(0)$. This is not one-to-one, so we cannot have $\alpha(1) = 0$.

Case 2. Can we have $\alpha(1) = 1$? Certainly $\alpha(1) = 1$ if α is the identity homomorphism ι, so we can have $\alpha(1) = 1$.

Case 3. Can we have $\alpha(1) = 2$? If so, then the homomorphism property implies that

$$\alpha(2) = \alpha(1+1) = \alpha(1) + \alpha(1) = 4 = 0 = \alpha(0).$$

This is not one-to-one, so we cannot have $\alpha(1) = 2$.

Case 4. Can we have $\alpha(1) = 3$? If so, then the homomorphism property implies that

$$\alpha(2) = \alpha(1+1) = \alpha(1) + \alpha(1) = 3+3 = 6 = 2; \text{ and}$$
$$\alpha(3) = \alpha(2+1) = \alpha(2) + \alpha(1) = 2+3 = 5 = 1.$$

In this case, α is both one-to-one *and* onto. We were careful to observe the homomorphism property when determining α, so we know that α is a homomorphism. So we *can* have $\alpha(1) = 2$.

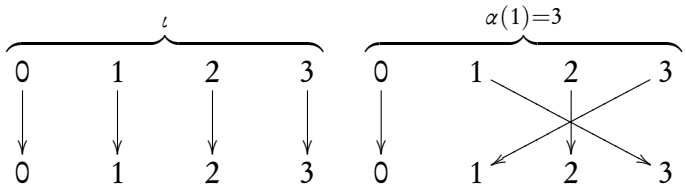


Figure 4.4. The elements of Aut $(\mathbb{Z}_4)$.

We found only two possible elements of Aut $(\mathbb{Z}_4)$: the identity automorphism and the automorphism determined by $\alpha(1) = 3$. Figure 4.4 illustrates the two mappings.

If Aut $(\mathbb{Z}_4)$ were a group, then the fact that it contains only two elements would imply that Aut $(\mathbb{Z}_4) \cong \mathbb{Z}_2$. But *is* it a group?

Lemma 4.60. For any group G, Aut (G) is a group under the operation of composition of functions.

On account of this lemma, we can justifiably refer to Aut (G) as **the automorphism group**.

Proof. Let G be any group. We show that Aut (G) satisfies each of the group properties from Definition 2.1.

(closed) Let $\alpha, \theta \in \text{Aut}(G)$. We must show that $\alpha \circ \theta \in \text{Aut}(G)$ as well:

- the domain and range of $\alpha \circ \theta$ are both G because the domain and range of both α and θ are both G;
- $\alpha \circ \theta$ is a *homomorphism* because for any $g, g' \in G$ we have,

$$\begin{aligned}
(\alpha \circ \theta)(g \cdot g') &= \alpha(\theta(g \cdot g')) && \text{(def. of comp.)} \\
&= \alpha(\theta(g) \cdot \theta(g')) && (\theta \text{ a homom.}) \\
&= \alpha(\theta(g)) \cdot \alpha(\theta(g')) && (\alpha \text{ a homom.}) \\
&= (\alpha \circ \theta)(g) \cdot (\alpha \circ \theta)(g'); && \text{(def. of comp.)}
\end{aligned}$$

- $\alpha \circ \theta$ is *one-to-one* because
  - if $(\alpha \circ \theta)(g) = (\alpha \circ \theta)(g')$, then by the definition of composition, $\alpha(\theta(g)) = \alpha(\theta(g'))$;
  - since α is one-to-one, $\theta(g) = \theta(g')$;
  - since θ is one-to-one, $g = g'$; and
- $\alpha \circ \theta$ is *onto* because for any $z \in G$,
  - α is onto, so there exists $y \in G$ such that $\alpha(y) = z$, and
  - θ is onto, so there exists $x \in G$ such that $\theta(x) = y$, so
  - $(\alpha \circ \theta)(x) = \alpha(\theta(x)) = \alpha(y) = z$.

We have shown that $\alpha \circ \theta$ satisfies the properties of an automorphism; hence, $\alpha \circ \theta \in \text{Aut}(G)$, and Aut (G) is closed under the composition of functions.

(associative) The associative property is sastisfied because the operation is composition of functions, which is associative.

(identity) Denote by ι the identity homomorphism; that is, $\iota(g) = g$ for all $g \in G$. We showed in Example 4.54(a) that ι is an automorphism, so $\iota \in \mathrm{Aut}(G)$. Let $\alpha \in \mathrm{Aut}(G)$; we claim that $\iota \circ \alpha = \alpha \circ \iota = \alpha$. Let $x \in G$ and write $y = \alpha(x)$. We have

$$(\iota \circ \alpha)(x) = \iota(\alpha(x)) = \iota(y) = y = \alpha(x),$$

and likewise $(\alpha \circ \iota)(x) = \alpha(x)$. Since x was arbitrary in G, we have $\iota \circ \alpha = \alpha \circ \iota = \alpha$.

(inverse) Let $\alpha \in \mathrm{Aut}(G)$. Since α is an automorphism, it is an isomorphism. You showed in Exercise 4.27 that α^{-1} is also an isomorphism. The domain and range of α are both G, so the domain and range of α^{-1} are also both G. Hence $\alpha^{-1} \in \mathrm{Aut}(G)$.

□

Since $\mathrm{Aut}(G)$ is a group, we can compute $\mathrm{Aut}(\mathrm{Aut}(G))$, and the same theory holds, so we can compute $\mathrm{Aut}(\mathrm{Aut}(\mathrm{Aut}(G)))$, and so forth. In the exercises, you will compute $\mathrm{Aut}(G)$ for some other groups.

Exercises.

Exercise 4.61. Show that $f(x) = x^2$ is an automorphism on the group $(\mathbb{R}^+, \times)$, but not on the group $(\mathbb{R}, \times)$.

Exercise 4.62. Recall the subgroup $A_3 = \{\iota, \rho, \rho^2\}$ of D_3.

(a) List the elements of $\mathrm{Conj}_\rho(A_3)$.

(b) List the elements of $\mathrm{Conj}_\varphi(A_3)$.

(c) In both (a) and (b), we saw that $\mathrm{Conj}_a(A_3) = A_3$ for $a = \rho, \varphi$. This makes sense, since $A_3 \triangleleft D_3$. Find a subgroup K of D_3 and an element $a \in D_3$ where $\mathrm{Conj}_a(K) \neq K$.

Exercise 4.63. Let $H = \langle \mathbf{i} \rangle < Q_8$. List the elements of $\mathrm{Conj}_{\mathbf{j}}(H)$.

Exercise 4.64. Prove Lemma 4.56 on page 144 in two steps:

(a) Show first that conj_a is an automorphism.

(b) Show that $\{\mathrm{conj}_a(h) : h \in H\}$ is a group.

Exercise 4.65. Determine the automorphism group of $\mathbb{Z}_5$.

Exercise 4.66. Determine the automorphism group of D_3.

Chapter 5:
Groups of permutations

This chapter introduces groups of permutations. Now, what is a permutation, and why are they so important?

Certain applications of mathematics involve the rearrangement of a list of n elements. It is common to refer to such rearrangements as *permutations*.

Definition 5.1. A **list** is a sequence. Let V be any finite list. A **permutation** is a one-to-one function whose domain and range are both V.

We require V to be a list rather than a set because for a permutation, the order of the elements matters: the lists $(a,d,k,r) \neq (a,k,d,r)$ even though $\{a,d,k,r\} = \{a,k,d,r\}$. For the sake of convenience, we usually write V as a list of natural numbers between 1 and $|V|$, but it can be any finite list.

Let's take a concrete example. Suppose you have a list of numbers, $(1,3,2,7)$, and you rearrange them by switching the first two entries in the list, $(3,1,2,7)$. The action of switching those first two numbers is a permutation. There is no doubt as to the outcome of the action, so this action is a function. Thus, permutations are special kinds of functions.

The importance of permutations is twofold. First, group theory is a pretty neat and useful thing in itself, and we will see in this chapter that all finite groups can be modeled by groups of permutations. Anything that can model every possible group is by that very fact important.

The second reason permutations are important has to do with the factorization of polynomials. The polynomial $x^4 - 1$ can be factored as

$$(x+1)(x-1)(x+i)(x-i),$$

but it can also be factored as

$$(x-1)(x+1)(x-i)(x+i).$$

On account of the commutative property, it doesn't matter what order we list the factors; this corresponds to a permutation, and is related to another idea that we will study, called field extensions. Field extensions can be used to solve polynomials equations, and since the order of the extensions doesn't really matter, permutations are important to determining the structure of the extension that solves a polynomial.

Section 5.1 introduces you to groups of permutations, while Section 5.2 describes a convenient way to write permutations. Sections 5.3 and 5.5 introduce you to two special classes of groups of permutation. The main goal of this chapter is to show that groups of permutations are, in some sense, "all there is" to group theory, which we accomplish in Section 5.4. We conclude with a great example of an application of symmetry groups in Section 5.6.

5.1: Permutations

In this first section, we consider some basic properties of permutations.

Permutations as functions

Example 5.2. Let $S = (a,d,k,r)$. Define a permutation on the elements of S by

$$f(x) = \begin{cases} r, & x = a; \\ a, & x = d; \\ k, & x = k; \\ d, & x = r. \end{cases}$$

Notice that f is one-to-one, and $f(S) = (r,a,k,d)$.

We can represent the same permutation on $V = (1,2,3,4)$, a generic list of four elements. Define a permutation on the elements of V by

$$\pi(i) = \begin{cases} 2, & i = 1; \\ 4, & i = 2; \\ 3, & i = 3; \\ 1, & i = 4. \end{cases}$$

Here π is one-to-one, and $\pi(i) = j$ is interpreted as "the jth element of the permuted list is the ith element of the original list." You could visualize this as

position i in original list		position j in permuted list
1	$\to$	2
2	$\to$	4
3	$\to$	3
4	$\to$	1

Thus $\pi(V) = (4,1,3,2)$. If you look back at $f(S)$, you will see that in fact the first element of the permuted list, $f(S)$, is the fourth element of the original list, S.

It should not surprise you that the identity function is a "do-nothing" permutation, just as it was a "do-nothing" symmetry of the triangle in Section 2.2.

Proposition 5.3. Let V be a set of n elements. The function $\iota : V \to V$ by $\iota(x) = x$ is a permutation on V. In addition, for any $\alpha \in S_n$, $\iota \circ \alpha = \alpha$ and $\alpha \circ \iota = \alpha$.

Proof. You do it! See Exercise 5.13. □

Permutations have a convenient property.

Lemma 5.4. The composition of two permutations is a permutation.

Proof. Let V be a set of n elements, and α, β permutations of V. Let $\gamma = \alpha \circ \beta$. We claim that γ is a permutation. To show this, we must show that γ is a one-to-one function whose domain and range are both V. The definition of α and β imply that the domain and range of γ are both V; it remains to show that γ is one-to-one. Let $x, y \in V$ and assume that $\gamma(x) = \gamma(y)$; substituting the definition of γ,

$$\alpha(\beta(x)) = \alpha(\beta(y)).$$

Because they are permutations, α and β are one-to-one functions. Since α is one-to-one, we can simplify the above equation to

$$\beta(x) = \beta(y);$$

and since β is one-to-one, we can simplify the above equation to

$$x = y.$$

We assumed that $\gamma(x) = \gamma(y)$, and found that this forced $x = y$. By definition, γ is a one-to-one function. We already explained why its domain and range are both V, so γ is a permutation. □

In Example 5.2, we wrote a permutation as a piecewise function. This is burdensome; we would like a more efficient way to denote permutations.

Notation 5.5. The **tabular notation** for a permutation on a list of n elements is a $2 \times n$ matrix

$$\alpha = \begin{pmatrix} 1 & 2 & \cdots & n \\ \alpha_1 & \alpha_2 & \cdots & \alpha_n \end{pmatrix}$$

indicating that $\alpha(1) = \alpha_1$, $\alpha(2) = \alpha_2$, …, $\alpha(n) = \alpha_n$. Again, $\alpha(i) = j$ indicates that the jth element of the permuted list is the ith element of the original list.

Example 5.6. Recall V and π from Example 5.2. In tabular notation,

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

because π moves

- the element in the first position to the second;
- the element in the second position to the fourth;
- the element in the third position nowhere; and
- the element in the fourth position to the first.

Then

$$\pi(1,2,3,4) = (4,1,3,2).$$

Notice that the tabular notation for π looks similar to the table in Example 5.2.

We can also use π to permute different lists, so long as the new lists have four elements:

$$\pi(3,2,1,4) = (4,3,1,2);$$
$$\pi(2,4,3,1) = (1,2,3,4);$$
$$\pi(a,b,c,d) = (d,a,c,b).$$

Groups of permutations

It comes as a pleasant revelation that sets of permutations form groups in a very natural way. In particular, consider the following set.

Definition 5.7. For $n \geq 2$, denote by S_n the set of all permutations of a list of n elements.

Example 5.8. For $n = 1,2,3$ we have

$$S_1 = \left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}$$
$$S_2 = \left\{ \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \right\}$$
$$S_3 = \left\{ \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} \right\}.$$

Is there some structure to S_n? By definition, a permutation is a one-to-one function. In Example 1.9 on page 40, we found that for any set, the set of functions on that set was a monoid under the operation of composition of functions. The identity function is one-to-one, and the composition of one-to-one functions is also one-to-one, so S_n has an identity and is closed under composition. In addition, S_n inherits the associative property from the larger set of functions. Already, then, we can conclude that S_n is a monoid. However, one-to-one functions have inverses, which leads us to ask whether S_n is also a group.

Theorem 5.9. For all $n \geq 2$ $(S_n, \circ)$ is a group.

Notation 5.10. Normally we just write S_n, understanding from context that the operation is composition of functions. It is common to refer to S_n as the **symmetric group** of n elements.

Proof. Let $n \geq 2$. We have to show that S_n satisfies the properties of a group under the operation of composition of functions. Proposition 5.3 tells us that the identity function acts as an identity in S_n, and Lemma 5.4 tells us that S_n is closed under composition.

We still have to show that S_n satisfies the inverse and associative properties. Let V be a finite list with n elements. The fact that $S_n \subseteq F_V$ implies that S_n satisfies the associative property. Let $\alpha \in S_n$. By definition of a permutation, α is one-to-one; since V is finite, α is onto. By Exercise 0.33 on page 12, α has an inverse function α^{-1}, which satisfies the relationship that, for every $v \in V$,

$$\alpha^{-1}(\alpha(v)) = v \quad \text{and} \quad \alpha\left(\alpha^{-1}(v)\right) = v.$$

Since $\iota(v) = v$ for every $v \in V$, we have shown that $\alpha^{-1} \circ \alpha = \alpha \circ \alpha^{-1} = \iota$. Again, Exercise 0.33 indicates that α^{-1} is a one-to-one, onto function on V, so $\alpha^{-1} \in S_n$! We chose α as an arbitrary permutation of n elements, so S_n satisfies the inverse property.

As claimed, S_n satisfies all four properties of a group. □

A final question: *how large is each S_n?* To answer this, we must count the number of permutations of n elements. A counting argument called *the multiplication principle* shows that there are

$$n! = n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1$$

such permutations. Why? Given any list of n elements,

- we have n positions to move the first element, including its current position;
- we have $n-1$ positions to move the second element, since the first element has already taken one spot;

- we have $n-2$ positions to move the third element, since the first and second elements have already take two spots;
- etc.

We have shown the following.

Lemma 5.11. For each $n \in \mathbb{N}^+, |S_n| = n!$

Exercises

Exercise 5.12. For the permutation

$$\alpha = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 5 & 2 & 4 & 6 & 3 \end{pmatrix},$$

(a) Evaluate $\alpha(1,2,3,4,5,6)$.
(b) Evaluate $\alpha(1,5,2,4,6,3)$.
(c) Evaluate $\alpha(6,3,5,2,1,4)$.

Exercise 5.13. Prove Proposition 5.3.

Exercise 5.14. How many elements are there of S_4?

Exercise 5.15. Identify at least one normal subgroup of S_3, and at least one subgroup that is not normal.

Exercise 5.16. Find an explicit isomorphism from S_2 to $\mathbb{Z}_2$.

Exercise 5.17. Do you think $S_3 \cong \mathbb{Z}_6$, $S_3 \cong D_3$, or neither? Why or why not? (Do not provide a full proof; a short justification will do.)

5.2: Cycle notation

Tabular notation of permutations is rather burdensome; a simpler notation is possible.

Cycles

Definition 5.18. A **cycle** is a vector

$$\alpha = (\alpha_1\ \alpha_2\ \cdots\ \alpha_n)$$

that corresponds to the permutation where the entry in position α_1 is moved to position α_2; the entry in position α_2 is moved to position α_3, ... and the element in position α_n is moved to position α_1. If a position is not listed in α, then the entry in that position is not moved. We call such positions **stationary**. For the identity permutation where no entry is moved, we write

$$\iota = (1).$$

The fact that the permutation α moves the entry in position α_n to position α_1 is the reason this is called a *cycle;* applying it repeatedly cycles the list of elements around, and on the nth application the list returns to its original order.

Example 5.19. Recall π from Example 5.6. In tabular notation,

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

To write it as a cycle, we can start with any position we like. However, the convention is to start with the smallest position that changes. Since π moves elements out of position 1, we start with

$$\pi = (1\,?)\,.$$

The second entry in cycle notation tells us where π moves the element whose position is that of the first entry. The first entry indicates position 1. From the tabular notation, we see that π moves the element in position 1 to position 2, so

$$\pi = (1\,2\,?)\,.$$

The third entry of cycle notation tells us where π moves the element whose position is that of the second entry. The second entry indicates position 2. From the tabular notation, we see that π moves the element in position 2 to position 4, so

$$\pi = (1\,2\,4\,?)\,.$$

The fourth entry of cycle notation tells us where π moves the element whose position is that of the third entry. The third element indicates position 4. From the tabular notation, we see that π moves the element in position 4 to position 1, so you might feel the temptation to write

$$\pi = (1\,2\,4\,1\,?)\,,$$

but there is no need. Since we have now returned to the first element in the cycle, we close it:

$$\pi = (1\,2\,4)\,.$$

The cycle $(1\,2\,4)$, indicates that

- the element in position 1 of a list moves to position 2;
- the element in position 2 of a list moves to position 4;
- the element in position 4 of a list moves to position 1.

What about the element in position 3? Since it doesn't appear in the cycle notation, it must be stationary. This agrees with what we wrote in the piecewise and tabular notations for π.

Not all permutations can be written as one cycle.

Example 5.20. Consider the permutation in tabular notation

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

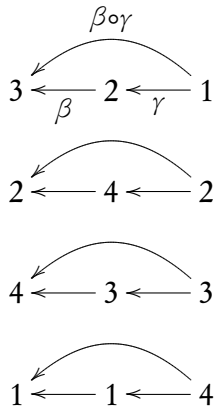


Figure 5.1. Diagram of how $\beta \circ \gamma$ modifies a list of four elements, for $\beta = (2\,3\,4)$ and $\gamma = (1\,2\,4)$.

We can easily start the cycle with $\alpha = (1\,2)$, and this captures the behavior on the elements in the first and second positions of a list, but what about the third and fourth positions? We cannot write $(1\,2\,3\,4)$; that would imply that the element in the second position is moved to the third, and the element in the fourth position is moved to the fourth.

To solve this difficulty, we develop a simple arithmetic of cycles.

Cycle arithmetic

What operation should we apply to cycles? Cycles represent permutations; permutations are functions; functions can be *composed*. Hence, the appropriate operation is *composition*.

Example 5.21. Consider the cycles

$$\beta = (2\,3\,4) \quad \text{and} \quad \gamma = (1\,2\,4)\,.$$

What is the cycle notation for

$$\beta \circ \gamma = (2\,3\,4) \circ (1\,2\,4)?$$

Let's think about this. Since cycles represent permutations, and permutations are closed under composition, $\beta \circ \gamma$ must be a permutation. With any luck, it will be a permutation that we can write as a cycle. What we need to do, then, is determine how the permutation $\beta \circ \gamma$ moves a list of four elements around. If that permutation can be represented as a cycle, then we've answered the question.

Since an element in the first position is moved, we should be able to write

$$\beta \circ \gamma = (1\,?)\,.$$

Where is this first element moved? Let's apply the definition of composition: $\beta \circ \gamma$ means, "first apply γ; then apply β." Figure 5.1 gives us the basic idea; we will refer to it throughout the example. Since γ moves an element in the first position to the second, and β moves an element in the second position to the third, it must be that $\beta \circ \gamma$ moves an element from the first position to the third. We see this in the top row of Figure 5.1. We now know that

$$\beta \circ \gamma = (1\,3\,?)\,.$$

The next entry should tell us where $\beta \circ \gamma$ moves an element that starts in the third position. Applying the definition of composition again, we know that γ moves an element from the third position to… well, nowhere, actually. So an element in the third position *doesn't* move under γ; if we then apply β, however, it moves to the fourth position. It must be that $\beta \circ \gamma$ moves an element from the third position to the fourth. We see this in the *third* row of Figure 5.1. We now know that

$$\beta \circ \gamma = (1\,3\,4\,?)\,.$$

Time to look at elements in the fourth position, then. Since γ moves elements in the fourth position to the first position (4 is at the end of the cycle, so it moves to the beginning), and β moves elements in the first position… well, nowhere, we conclude that $\beta \circ \gamma$ moves elements from the fourth position to the first position. This completes the cycle, so we now know that

$$\beta \circ \gamma = (1\,3\,4)\,.$$

Haven't we missed something? What about an element that starts in the second position? Since γ moves elements in the second position to the fourth, and β moves elements from the fourth position to the second, they undo each other, and the second position is stationary. It is, therefore, *absolutely correct* that 2 does not appear in the cycle notation of $\beta \circ \gamma$, and we see this in the *second* row of Figure 5.1.

Another phenomenon occurs when each permutation moves elements that the other does not.

Example 5.22. Consider the two cycles

$$\beta = (1\,3) \quad \text{and} \quad \gamma = (2\,4)\,.$$

There is no way to simplify $\beta \circ \gamma$ into a *single* cycle, because β operates only on the first and third elements of a list, and γ operates only on the second and fourth elements of a list. The only way to write them is as the composition of two cycles,

$$\beta \circ \gamma = (1\,3) \circ (2\,4)\,.$$

This motivates the following.

Definition 5.23. We say that two cycles are **disjoint** if none of their entries are common.

Disjoint cycles enjoy an important property: their permutations commute under composition.

Lemma 5.24. Let α, β be two disjoint cycles. Then $\alpha \circ \beta = \beta \circ \alpha$.

Proof. Let $n \in \mathbb{N}^+$ be the largest entry in α or β. Let $V = (1, 2, \ldots, n)$. Let $i \in V$. We consider the following cases:

Case 1. $\alpha(i) \neq i$.

Let $j = \alpha(i)$. The definition of cycle notation implies that j appears immediately after i in the cycle α. The definition of "disjoint" means that, since i and j are entries of α, they cannot

be entries of β. By definition of cycle notation, $\beta(i) = i$ and $\beta(j) = j$. Hence

$$(\alpha \circ \beta)(i) = \alpha(\beta(i)) = \alpha(i) = j = \beta(j) = \beta(\alpha(i)) = (\beta \circ \alpha)(i).$$

Case 2. $\alpha(i) = i$.

Subcase (a): $\beta(i) = i$.

We have $(\alpha \circ \beta)(i) = i = (\beta \circ \alpha)(i)$.

Subcase (b): $\beta(i) \neq i$.

Let $j = \beta(i)$. The definition of cycle notation implies that j appears immediately after i in the cycle β. The definition of "disjoint" means that, since i and j are entries of β, they cannot be entries of α. By definition of cycle notation, $\alpha(j) = j$. Hence

$$(\alpha \circ \beta)(i) = \alpha(\beta(i)) = \alpha(j) = j = \beta(i) = \beta(\alpha(i)) = (\beta \circ \alpha)(i).$$

In both cases, we had $(\alpha \circ \beta)(i) = (\beta \circ \alpha)(i)$. Since i was arbitrary, $\alpha \circ \beta = \beta \circ \alpha$. □

Notation 5.25. Since the composition of two disjoint cycles $\alpha \circ \beta$ cannot be simplified, we normally write it without the circle; for example,

$$(1\,2)(3\,4).$$

By Lemma 5.24, we can also write this as

$$(3\,4)(1\,2).$$

That said, the usual convention for cycles is to write the smallest entry of a cycle first, and to write cycles with smaller first entries before cycles with larger first entries. Thus, we prefer

$$(1\,4)(2\,3)$$

to either of

$$(1\,4)(3\,2) \quad \text{or} \quad (2\,3)(1\,4).$$

The convention for writing a permutation in cycle form is the following:

1. The first entry in each cycle is the cycle's smallest.
2. We simplify the composition of cycles that are not disjoint, discarding all cycles of length 1.
3. The remaining cycles will be disjoint. From Lemma 5.24, we know that they commute; write them so that the first cycle's first entry is smallest, the second cycle's first entry is second-smallest, and so forth.

Example 5.26. We return to Example 5.20, with

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

To write this permutation in cycle notation, we begin again with

$$\alpha = (1\,2)\ldots?$$

Since α also moves entries in positions 3 and 4, we need to add a second cycle. We start with the smallest position whose entry changes position, 3:

$$\alpha = (1\,2)\,(3\,?)\,.$$

Since α moves the element in position 3 to position 4, we write

$$\alpha = (1\,2)\,(3\,4\,?)\,.$$

Now α moves the element in position 4 to position 3, so we can close the second cycle:

$$\alpha = (1\,2)\,(3\,4)\,.$$

Now α moves no more entries, so the cycle notation is complete.

Permutations as cycles

We have come to the main result of this section.

Theorem 5.27. Every permutation can be written as a composition of cycles.

The proof is constructive; we build the cycle notation for the permutation.

Proof. Let π be a permutation; denote its domain by V. Without loss of generality, we write $V = (1,2,\ldots,n)$.

Let i_1 be the smallest element of V such that $\pi(i_1) \neq i_1$. Recall that the range of π has at most n elements, so the sequence $\pi(i_1)$, $\pi(\pi(i_1)) = \pi^2(i_1)$, $\ldots$ cannot continue indefinitely; eventually, we must have $\pi^{k+1}(i_1) = i_1$ for some $k \leq n$. Let

$$\alpha^{(1)} = \left(i_1\ \pi(i_1)\ \pi(\pi(i_1))\ \cdots\ \pi^k(i_1)\right).$$

Is there is some $i_2 \in V$ that is not stationary with respect to π and not an entry of $\alpha^{(1)}$? If so, then generate the cycle $\alpha^{(2)}$ by $\left(i_2\ \pi(i_2)\ \pi(\pi(i_2))\ \cdots\ \pi^\ell(i_2)\right)$, where, as before $\pi^{\ell+1}(i_2) = i_2$.

Repeat this process until every non-stationary element of V corresponds to a cycle, generating $\alpha^{(3)}$, $\ldots$, $\alpha^{(m)}$ for non-stationary $i_3 \notin \alpha^{(1)}, \alpha^{(2)}$, $i_4 \notin \alpha^{(1)}, \alpha^{(2)}, \alpha^{(3)}$, and so on until $i_m \notin \alpha^{(1)}, \ldots, \alpha^{(m-1)}$. Since the list is finite, this process will not continue indefinitely, and we have a finite list of cycles.

The remainder of the proof consists of two claims.

Claim 1: Each of the cycles we created is disjoint from any of the rest.

By way of contradiction, assume that two cycles $\alpha^{(i)}$ and $\alpha^{(j)}$ are not disjoint. By construction, the first elements of these cycles are different; let r be the first entry in $\alpha^{(j)}$ that also appears in $\alpha^{(i)}$. Let a be the entry that precedes r in $\alpha^{(i)}$, and b the entry that precedes r in $\alpha^{(j)}$. By construction, we have $\alpha(a) = r = \alpha(b)$. Since r is the *first* entry of each cycle that is the same, $a \neq b$. This contradicts the hypothesis that α is a permutation, as permutations are one-to-one. Hence, $\alpha^{(i)}$ and $\alpha^{(j)}$ are disjoint.

Claim 2: $\pi = \alpha^{(1)}\alpha^{(2)}\cdots\alpha^{(m)}$.

Let $i \in V$. We consider two cases.

If $\pi(i) = i$, then i could not have been used to begin construction of an $\alpha^{(j)}$. Since π is a one-to-one function, we cannot have $\pi(k) = i$ for any $k \neq i$, either. By construction, i appears in none of the $\alpha^{(j)}$.

Assume, then, that $\pi(i) \neq i$. By construction, i appears in $\alpha^{(j)}$ for some $j = 1, 2, \ldots, m$. By definition, $\alpha^{(j)}(i) = \pi(i)$, so $\alpha^{(k)}(i) = i$ for $k \neq j$. By Claim 1, both i and $\pi(i)$ appear in *only* one of the α. By substitution, the expression $\left(\alpha^{(1)}\alpha^{(2)}\cdots\alpha^{(m)}\right)(i)$ simplifies to

$$\begin{aligned}\left(\alpha^{(1)}\alpha^{(2)}\cdots\alpha^{(m)}\right)(i) &= \alpha^{(1)}\left(\alpha^{(2)}\left(\cdots\alpha^{(m-1)}\left(\alpha^{(m)}(i)\right)\right)\right)\\ &= \alpha^{(1)}\left(\alpha^{(2)}\left(\cdots\alpha^{(j-1)}\left(\alpha^{(j)}(i)\right)\right)\right)\\ &= \alpha^{(1)}\left(\alpha^{(2)}\left(\cdots\alpha^{(j-1)}(\pi(i))\right)\right)\\ &= \pi(i).\end{aligned}$$

We have shown that

$$\left(\alpha^{(1)}\alpha^{(2)}\cdots\alpha^{(m)}\right)(i) = \pi(i).$$

Since i is arbitrary, $\pi = \alpha^{(1)} \circ \alpha^{(2)} \circ \cdots \circ \alpha^{(m)}$. That is, π is a composition of cycles. Since π was arbitrary, every permutation is a composition of cycles. □

Example 5.28. Consider the following permutation written in tabular notation,

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 7 & 5 & 3 & 2 & 4 & 8 & 1 & 6 \end{pmatrix}.$$

The proof of Theorem 5.27 constructs the cycles

$$\begin{aligned}\alpha^{(1)} &= (1\,7)\\ \alpha^{(2)} &= (2\,5\,4)\\ \alpha^{(3)} &= (6\,8).\end{aligned}$$

Notice that $\alpha^{(1)}$, $\alpha^{(2)}$, and $\alpha^{(3)}$ are disjoint. In addition, the only element of $V = (1, 2, \ldots, 8)$ that does not appear in an α is 3, because $\pi(3) = 3$. Inspection verifies that

$$\pi = \alpha^{(1)}\alpha^{(2)}\alpha^{(3)}.$$

We conclude with some examples of simplifying the composition of permutations.

Example 5.29. Let $\alpha = (1\,3)(2\,4)$ and $\beta = (1\,3\,2\,4)$. Notice that $\alpha \neq \beta$; check this on $V = (1, 2, 3, 4)$ if this isn't clear. In addition, α and β are not disjoint.

1. We compute the cycle notation for $\gamma = \alpha \circ \beta$. We start with the smallest entry moved by either α or β:

   $$\gamma = (\ 1\ \ ?\ ).$$

   The notation $\alpha \circ \beta$ means to apply β first, *then* α. What does β do with the entry in position 1? It moves it to position 3. Subsequently, α moves the entry in position 3 back to the entry in position 1. The next entry in the first cycle of γ should thus be 1, but that's

also the first entry in the cycle, so we close the cycle. So far, we have

$$\gamma = (1)\ldots?$$

We aren't finished, since α and β also move other entries around. The next smallest entry moved by either α or β is 2, so

$$\gamma = (1)\,(2\,?)\,.$$

Now β moves the entry in position 2 to the entry in position 4, and α moves the entry in position 4 to the entry in position 2. The next entry in the second cycle of γ should thus be 2, but that's also the first entry in the second cycle, so we close the cycle. So far, we have

$$\gamma = (1)\,(2)\ldots?$$

Next, β moves the entry in position 3, so

$$\gamma = (1)\,(2)\,(3\,?)\,.$$

Where does β move the entry in position 3? To the entry in position 2. Subsequently, α moves the entry in position 2 to the entry in position 4. We now have

$$\gamma = (1)\,(2)\,(3\,4\,?)\,.$$

You can probably guess that 4, as the largest possible entry, will close the cycle, but to be safe we'll check: β moves the entry in position 4 to the entry in position 1, and α moves the entry in position 1 to the entry in position 3. The next entry of the third cycle will be 3, but this is also the first entry of the third cycle, so we close the third cycle and

$$\gamma = (1)\,(2)\,(3\,4)\,.$$

Finally, we simplify γ by not writing cycles of length 1, so

$$\gamma = (3\,4)\,.$$

Hence

$$((1\,3)\,(2\,4)) \circ (1\,3\,2\,4) = (3\,4)\,.$$

2. Now we compute the cycle notation for $\beta \circ \alpha$, but with less detail. Again we start with 1, which α moves to 3, and β then moves to 2. So we start with

$$\beta \circ \alpha = (1\,2\,?)\,.$$

Next, α moves 2 to 4, and β moves 4 to 1. This closes the first cycle:

$$\beta \circ \alpha = (1\,2)\ldots?$$

We start the next cycle with position 3: α moves it to position 1, which β moves back to position 3. This generates a length-one cycle, so there is no need to add anything. Likewise,

the element in position 4 is also stable under $\beta \circ \alpha$. Hence we need write no more cycles;

$$\beta \circ \alpha = (1\,2)\,.$$

3. Let's look also at $\beta \circ \gamma$ where $\gamma = (1\,4)$. We start with 1, which γ moves to 4, and then β moves to 1. Since $\beta \circ \gamma$ moves 1 to itself, we don't have to write 1 in the cycle. The next smallest number that appears is 2: γ doesn't move it, and β moves 2 to 4. We start with

$$\beta \circ \gamma = (2\,4\,?)\,.$$

Next, γ moves 4 to 1, and β moves 1 to 3. This adds another element to the cycle:

$$\beta \circ \gamma = (2\,4\,3\,?)\,.$$

We already know that 1 won't appear in the cycle, so you might guess that we should not close the cycle. To be certain, we consider what $\beta \circ \gamma$ does to 3: γ doesn't move it, and β moves 3 to 2. The cycle is now complete:

$$\beta \circ \gamma = (2\,4\,3)\,.$$

Exercises.

Exercise 5.30. For the permutation

$$\alpha = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 5 & 2 & 4 & 6 & 3 \end{pmatrix},$$

(a) Write α in cycle notation.
(b) Write α as a piecewise function.

Exercise 5.31. For the permutation

$$\alpha = (1\,3\,4\,2)\,,$$

(a) Evaluate $\alpha\,(1,2,3,4)$.
(b) Evaluate $\alpha\,(1,4,3,2)$.
(c) Evaluate $\alpha\,(3,1,4,2)$.
(d) Write α in tabular notation.
(e) Write α as a piecewise function.

Exercise 5.32. Let $\alpha = (1\,2\,3\,4)$, $\beta = (1\,4\,3\,2)$, and $\gamma = (1\,3)$. Compute $\alpha \circ \beta$, $\alpha \circ \gamma$, $\beta \circ \gamma$, $\beta \circ \alpha$, $\gamma \circ \alpha$, $\gamma \circ \beta$, α^2, β^2, and γ^2. (Here $\alpha^2 = \alpha \circ \alpha$.) What are the inverses of α, β, and γ?

Exercise 5.33. Compute the order of

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

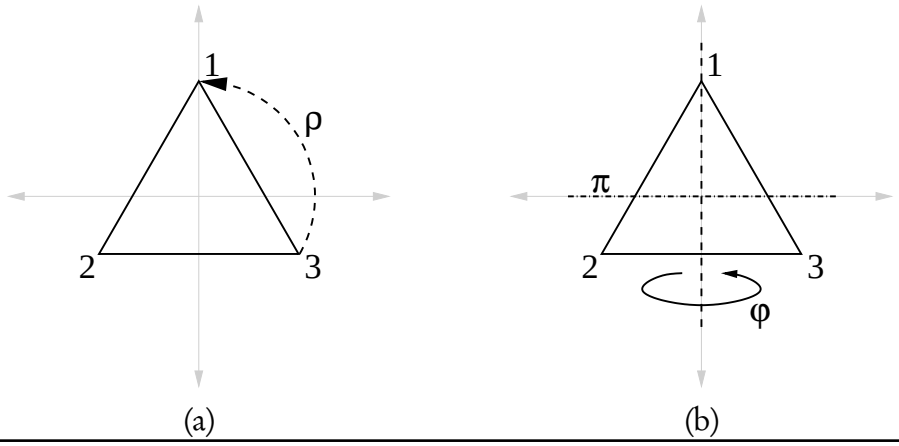


Figure 5.2. Rotation and reflection of an equilateral triangle centered at the origin

Exercise 5.34. Show that all the elements of S_3 can be written as compositions of the cycles $\alpha = (1\,2\,3)$ and $\beta = (2\,3)$.

Exercise 5.35. For α and β as defined in Exercise 5.34 on the previous page, show that $\beta \circ \alpha = \alpha^2 \circ \beta$. (Notice that $\alpha, \beta \in S_n$ for all $n > 2$, so as a consequence of this exercise S_n is not abelian for $n > 2$.)

Exercise 5.36. Write the Cayley table for S_3.

Exercise 5.37. Show that $D_3 \cong S_3$ by showing that the function $f : D_3 \to S_3$ by $f\left(\rho^a \varphi^b\right) = \alpha^a \beta^b$ is an isomorphism.

Exercise 5.38. List the elements of S_4 using cycle notation.

Exercise 5.39. Compute the cyclic subgroup of S_4 generated by $\alpha = (1\,3\,4\,2)$. Compare your answer to that of Exercise 5.33.

Exercise 5.40. Let $\alpha = (\alpha_1\, \alpha_2 \cdots \alpha_m) \in S_n$. (Note $m \leq n$.) Show that we can write α^{-1} as

$$\beta = \left(\alpha_1\, \alpha_m\, \alpha_{m-1} \cdots \alpha_2\right).$$

For example, if $\alpha = (2\,3\,5\,6)$, $\alpha^{-1} = (2\,6\,5\,3)$.

5.3: Dihedral groups

In Section 2.2 we studied the symmetries of a triangle; we presented the group as the products of matrices ρ and φ, derived from the symmetries of *rotation* and *reflection about the y-axis*. Figure 5.2, a copy of Figure 2.4 on page 68, shows how ρ and φ correspond to the symmetries of an equilateral triangle centered at the origin. In Exercises 5.34–5.37 you showed that D_3 and S_3 are isomorphic.

From symmetries to permutations

We now turn to the symmetries of a regular n-sided polygon.

Definition 5.41. The **dihedral set** D_n is the set of symmetries of a regular polygon with n sides.

We have two goals in introducing the dihedral group: first, to give you another concrete and interesting group; and second, to serve as a bridge to Section 5.4. The next example starts starts us in that directions.

Example 5.42. Another way to represent the elements of D_3 is to consider how they re-arrange the vertices of the triangle. We can represent the vertices of a triangle as the list $V = (1,2,3)$. Application of ρ to the triangle moves

- vertex 1 to vertex 2;
- vertex 2 to vertex 3; and
- vertex 3 to vertex 1.

This is equivalent to the permutation $(1\,2\,3)$. Application of φ to the triangle moves

- vertex 1 to itself—that is, vertex 1 does not move;
- vertex 2 to vertex 3; and
- vertex 3 to vertex 2.

This is equivalent to the permutation $(2\,3)$.

In the context of the symmetries of the triangle, it looks as if ρ and φ correspond to $(1\,2\,3)$ and $(2\,3)$, respectively. Recall that ρ and φ generate all the symmetries of a triangle; likewise, these two cycles generate all the permutations of a list of three elements! (See Example 5.8 and Exercise 2.45 on page 74.)

We can do this with D_4 and S_4 as well.

Example 5.43. Using the tabular notation for permutations, we identify some elements of D_4, the set of symmetries of a square. As with the triangle, we can represent the vertices of a square as the list $V = (1,2,3,4)$. The identity symmetry ι, which moves the vertices back onto themselves, is thus the cycle (1). We also have a 90° rotation which moves vertex 1 to vertex 2, vertex 2 to vertex 3, and so forth. As a permutation, we can write that as

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

The other rotations are clearly powers of ρ. We can visualize three kinds of flips: one across the y-axis,

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

one across the x-axis,

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

and one across a diagonal,

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

See Figure 5.3 on the next page. We can also imagine other diagonals; but they can be shown to be superfluous, just as we show shortly that ϑ and ψ are superflulous. There may be other symmetries of the square, but we'll stop here for the time being.

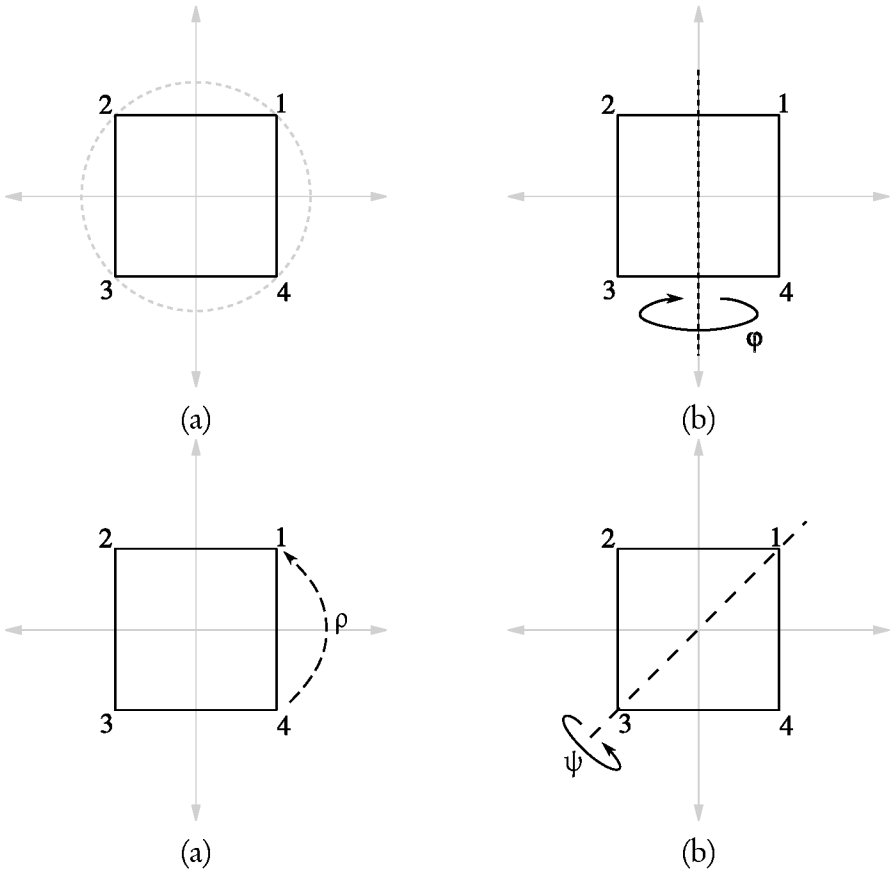


Figure 5.3. Rotation and reflection of a square centered at the origin

Is it possible to write ψ as a composition of φ and ρ? It turns out that $\psi = \varphi \circ \rho$. We can show this by observing that

$$\varphi \circ \rho = \begin{pmatrix} 1 & 2 \end{pmatrix} \begin{pmatrix} 3 & 4 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 \end{pmatrix} = \begin{pmatrix} 2 & 4 \end{pmatrix} = \psi.$$

We can see this geometrically; see Figure 5.4. First, ρ moves $(1,2,3,4)$ to $(4,1,2,3)$. Subsequently, φ moves $(4,1,2,3)$ to $(1,4,3,2)$. Likewise, ψ would permute $(1,2,3,4)$ directly to $(1,4,3,2)$. Either way, we see that $\psi = \varphi \circ \rho$. A similar argument shows that $\vartheta = \varphi \circ \rho^2$, so it looks as if we need only φ and ρ to generate D_4.

Similar arguments verify that the reflection and the rotation have a property similar to that in S_3:

$$\varphi \circ \rho = \rho^3 \circ \varphi,$$

so unless there is some symmetry of the square that cannot be described by rotation or reflection on the y-axis, we can list all the elements of D_4 using a composition of some power of ρ after some power of φ. There are four unique 90° rotations and two unique reflections on the y-axis, implying that D_4 has at least eight elements:

$$D_4 \supseteq \left\{ \iota, \rho, \rho^2, \rho^3, \varphi, \rho\varphi, \rho^2\varphi, \rho^3\varphi \right\}.$$

Can D_4 have other elements? There are in fact $|S_4| = 4! = 24$ possible permutations of the vertices, but are they all symmetries of a square? Consider the permutation from $(1,2,3,4)$ to

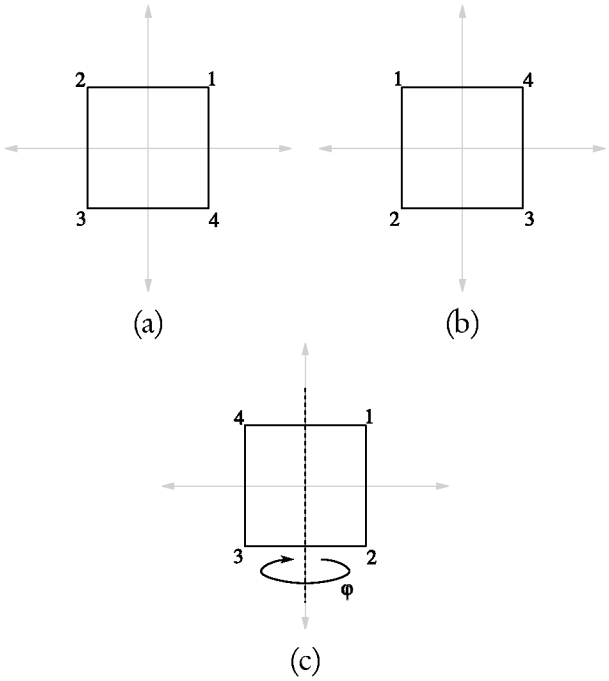


Figure 5.4. Rotation and reflection of a square centered at the origin

$(2,1,3,4)$: in the basic square, the distance between vertices 1 and 3 is $\sqrt{2}$, but in the configuration $(2,1,3,4)$ vertices 1 and 3 are adjacent on the square, so the distance between them has diminished to 1. Meanwhile, vertices 2 and 3 are no longer adjacent, so the distance between them has increased from 1 to $\sqrt{2}$. Since the distances between points on the square was not preserved, the permutation described, $\begin{pmatrix} 1 & 2 \end{pmatrix}$, is *not* an element of D_4. The same can be shown for the other fifteen permutations of four elements.

Hence D_4 has eight elements, making it smaller than S_4, which has $4! = 24$.

Is D_n always a group?

Theorem 5.44. Let $n \in \mathbb{N}^+$. If $n \geq 3$, then $(D_n, \circ)$ is a group with $2n$ elements, called the **dihedral group**.

It is possible to prove Theorem 5.44 using the following proposition, which could be proved using an argument from matrices, as in Section 2.2.

Proposition 5.45. All the symmetries of a regular n-sided polygon can be generated by a composition of a power of the rotation ρ of angle $2\pi/n$ and a power of the flip φ across the y-axis. In addition, $\varphi^2 = \rho^n = \iota$ (the identity symmetry) and $\varphi\rho = \rho^{n-1}\varphi$.

However, that would be a colossal waste of time. Instead, we prove the theorem *by turning symmetries of the polygon into permutations.*

D_n and S_n

Our strategy is as follows. For arbitrary $n \in \mathbb{N}^+$, we consider a list $(1,2,\ldots,n)$ of vertices of the n-sided polygon, imagine how they can move without violating the rules of symmetry,

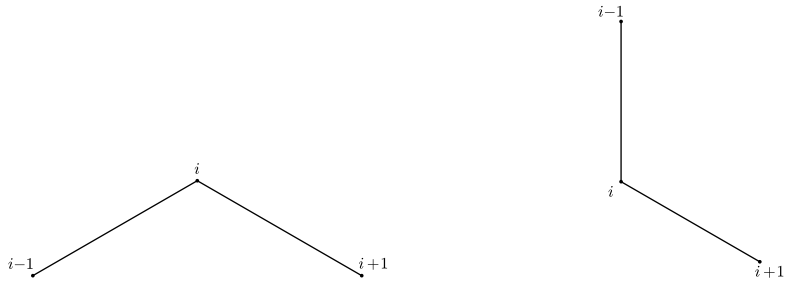


Figure 5.5. To preserve distance bewteen vertices, a permutation of a regular polygon must move vertex i and its neighbors in such a way that they remain neighbors.

and then count how many possible permutations that gives us. We then show that this set of permutations satsifies the requirements of a group.

Proof of Theorem 5.44. Let $n \in \mathbb{N}^+$ and assume $n \geq 3$. Let $V = (1, 2, \ldots, n)$ be a list of the vertices of the n-sided polygon, *in order*. Thus, the distance from vertex $i-1$ to vertex i is precisely the distance from vertex i to vertex $i+1$.

What must be true after we apply any symmetry? While vertices $i-1$, i, and $i+1$ may have moved, the distances between them *may not* change. Thus, we can rearrange them in the order $i-1$, i, and $i+1$, but in different positions, or in the order $i+1$, i, $i-1$, in either the same or different positions. That limits our options. To count the number of possible symmetries, then, we start by counting the number of positions where we can move vertex 1: there are n such positions, one for each vertex. As we just observed, the vertex that follows vertex 1 *must* be vertex 2 or vertex n — if we are to preserve the distances between vertices, we have no other choice! (See Figure 5.5.) That gives us only two choices for the vertex that follows vertex 1! We can in fact create symmetries corresponding to these choices — simply count up or down, as appropriate. By the counting principle, D_n has $2n$ elements. But is it a group?

The associative property follows from the fact that permutations are functions, and composition of functions is associative. The identity symmetry, which moves the vertices onto themselves, corresonds to the identity element $\iota \in D_n$. The inverse property holds because (1) any permutation has an inverse permutation, and (2) Exercise 5.40 shows that this inverse permutation reverses the order of entries, so that the requirement that vertex $i-1$ precede or follow vertex i is preserved.

It remains to show closure. Let $\alpha, \beta \in D_n$, and let $i \in V$. Now, if $\beta(i) = j$, then the preservation of distance between vertices implies that $\beta(i+1)$ either precedes j or succeeds it; that is, $\beta(i+1) = j \pm 1$. If $\alpha(j) = k$, then the preservation of distance between vertices implies that $\alpha(j \pm 1)$ either precedes k or succeeds it; that is, $\alpha(j \pm 1) = k \pm 1$. By substitution,

$$(\alpha \circ \beta)(i) = \alpha(\beta(i)) = \alpha(j) = k$$

and

$$(\alpha \circ \beta)(i+1) = \alpha(\beta(i+1)) = \alpha(j \pm 1) = k \pm 1.$$

We see that $\alpha \circ \beta$ preserves the distance between the vertices, as vertex $i+1$ after the transformation either succeeds or precedes vertex i. Since i was arbitrary in V, this is true for all the vertices of the n-sided polygon. Thus, $\alpha \circ \beta \in D_n$, and D_n is closed.

We have shown that D_n has $2n$ elements, and that it satisfies the four properties of a group. □

The basic argument we followed above gives us the following result, as well.

Corollary 5.46. For any $n \geq 3$, D_n is isomorphic to a subgroup of S_n. If $n = 3$, then $D_3 \cong S_3$ itself.

Proof. You already proved that $D_3 \cong S_3$ in Exercise 5.37. □

What we have seen is that some problems, such as the symmetries of a regular polygon, fall naturally into a group-theoretical context if you can formulate the activity as a set of permutations. The next section shows that this is no accident.

Exercises.

Exercise 5.47. Write all eight elements of D_4 in cycle notation.

Exercise 5.48. Construct the composition table of D_4. Compare this result to that of Exercise 2.84.

Exercise 5.49. Show that the symmetries of any n-sided polygon can be described as a power of ρ and φ, where φ is a flip about the y-axis and ρ is a rotation of $2\pi/n$ radians.

Exercise 5.50. Show that D_n is solvable for all $n \geq 3$.

5.4: Cayley's Theorem

The mathematician Arthur Cayley discovered a lovely fact about the permutation groups. Its effective consequence is that the theory of finite groups is equivalent to the study of groups of permutations.

Theorem 5.51 (Cayley's Theorem)**.** Every group of order n is isomorphic to a subgroup of S_n.

Before we give the proof, we give an example that illustrates how the proof of the theorem works.

Example 5.52. Consider the Klein 4-group; this group has four elements, so Cayley's Theorem tells us that it must be isomorphic to a subgroup of S_4. We will build the isomorphism by looking at the Cayley table for the Klein 4-group:

$\times$	e	a	b	ab
e	e	a	b	ab
a	a	e	ab	b
b	b	ab	e	a
ab	ab	b	a	e

To find a permutation appropriate to each element, we'll do the following. First, we label each element with a certain number:

$$\begin{aligned} e &\leftrightsquigarrow 1, \\ a &\leftrightsquigarrow 2, \\ b &\leftrightsquigarrow 3, \\ ab &\leftrightsquigarrow 4. \end{aligned}$$

We will use this along with tabular notation to determine the isomorphism. Define a map f from the Klein 4-group to S_4 by

$$f(x) = \begin{pmatrix} 1 & 2 & 3 & 4 \\ \ell(x \cdot e) & \ell(x \cdot a) & \ell(x \cdot b) & \ell(x \cdot ab) \end{pmatrix}, \tag{13}$$

where $\ell(y)$ is the label that corresponds to y.

This notation can make things hard to read. Why? Well, f maps an element g of the Klein 4-group to a permutation $f(x) = \sigma$ of S_4. Suppose $\sigma = (1\,2)(3\,4)$. Any permutation of S_4 is a one-to-one function on a list of 4 elements, say $(1,2,3,4)$. By definition, $\sigma(2) = 1$. Since $\sigma = f(x)$, we can likewise write, $(f(x))(2) = 1$. This double-evaluation is hard to look at; is it saying "$f(x)$ times 2" or "$f(x)$ of 2"? In fact, it says the latter. To avoid confusion, we adopt the following notation to emphasize that $f(x)$ is a permutation, and thus a function:

$$f(x) = f_x.$$

It's much easier now to look at $f_x(2)$ and understand that we want $f_x(2) = 1$.

Let's compute f_a:

$$f_a = \begin{pmatrix} 1 & 2 & 3 & 4 \\ \ell(a \cdot e) & \ell(a \cdot a) & \ell(a \cdot b) & \ell(a \cdot ab) \end{pmatrix}.$$

The first entry has the value $\ell(a \cdot e) = \ell(a) = 2$, telling us that

$$f_a = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & \ell(a \cdot a) & \ell(a \cdot b) & \ell(a \cdot ab) \end{pmatrix}.$$

The next entry has the value $\ell(a \cdot a) = \ell(a^2) = \ell(e) = 1$, telling us that

$$f_a = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & \ell(a \cdot b) & \ell(a \cdot ab) \end{pmatrix}.$$

The third entry has the value $\ell(a \cdot b) = \ell(ab) = 4$, telling us that

$$f_a = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & \ell(a \cdot ab) \end{pmatrix}.$$

The final entry has the value $\ell(a \cdot ab) = \ell\left(a^2 b\right) = \ell(b) = 3$, telling us that

$$f_a = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix} = (\,1\;\;2\,)(\,3\;\;4\,).$$

So applying the formula in equation (13) definitely gives us a permutation.

Look closely. We could have filled out the bottom row of the permutation by looking above at the Klein 4-group's Cayley table, locating the row for the multiples of a (the second row of the multiplication table), and filling in the labels for the entries in that row! After all,

the row corresponding to a is *precisely*

the row of products $a \cdot y$ for all elements y of the group!

Doing this or applying equation (13) to the other elements of the Klein 4-group tells us that

$$\begin{aligned} f_e &= \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix} = (1) \\ f_b &= \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \end{pmatrix} = (\,1\;\;3\,)(\,2\;\;4\,) \\ f_{ab} &= \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix} = (\,1\;\;4\,)(\,2\;\;3\,). \end{aligned}$$

The result is a subset of S_4; or, in cycle notation,

$$\begin{aligned} W &= \{f_e, f_a, f_b, f_{ab}\} \\ &= \{(1), (1\,2)(3\,4), (1\,3)(2\,4), (1\,4)(2\,3)\}. \end{aligned}$$

Verifying that W is a group, and therefore a subgroup of S_4, is straightforward; you will do so in the homework. In fact, it is a consequence of the fact that f is a homomorphism. Strictly speaking, f is really an isomorphism. Inspection shows that f is one-to-one and onto; the hard part is the homomorphism property. We will use a little cleverness for this. Let x, y in the Klein 4-group.

- Recall that f_x, f_y, and f_{xy} are permutations, and by definition one-to-one, onto functions on a list of four elements.
- Notice that ℓ is also a one-to-one function, and it has an inverse. Just as $\ell(z)$ is the label of z, $\ell^{-1}(m)$ is the element labeled by the number m. For instance, $\ell^{-1}(b) = 3$.
- Since f_x is a permutation of a list of four elements, we can look at $f_x(m)$ as the position where f_x moves the element in the mth position.
- By definition, f_x moves m to $\ell(z)$ where z is the product of x and the element in the mth position. Written differently, $z = x \cdot \ell^{-1}(m)$, so

$$f_x(m) = \ell\left(x\ell^{-1}(m)\right). \tag{14}$$

  Similar statements hold for f_y and f_{xy}.

- Applying these facts, we observe that

$$\begin{aligned}
\left(f_x \circ f_y\right)(m) &= f_x\left(f_y(m)\right) && \text{(def. of comp.)}\\
&= f_x\left(\ell\left(y\cdot\ell^{-1}(m)\right)\right) && \text{(def. of } f_y\text{)}\\
&= \ell\left(x\cdot\ell^{-1}\left(\ell\left(y\cdot\ell^{-1}(m)\right)\right)\right) && \text{(def. of } f_x\text{)}\\
&= \ell\left(x\cdot\left(y\cdot\ell^{-1}(m)\right)\right) && (\ell^{-1},\ell \text{ inverses})\\
&= \ell\left(xy\cdot\ell^{-1}(m)\right) && \text{(assoc. prop.)}\\
&= f_{xy}(m). && \text{(def. of } f_{xy}\text{)}
\end{aligned}$$

- Since m was arbitrary in $\{1,2,3,4\}$, f_{xy} and $f_x \circ f_y$ are identical functions.
- Since $f_x f_y = f_x \circ f_y$, we have $f_{xy} = f_x f_y$.
- Since x, y were arbitrary in the Klein 4-group, this holds for the entire group.

We conclude that f is a homomorphism; since it is one-to-one and onto, f is an isomorphism.

You should read through Example 5.52 carefully two or three times, and make sure you understand it, since in the homework you will construct a similar isomorphism for a different group, and also because we do the same thing now in the proof of Cayley's Theorem.

Proof of Cayley's Theorem. Let G be a finite group of n elements. Label the elements in any order $G = \{g_1, g_2, \ldots, g_n\}$ and for any $x \in G$ denote $\ell(x) = i$ such that $x = g_i$. Define a relation

$$f : G \to S_n \qquad \text{by} \qquad f(g) = \begin{pmatrix} 1 & 2 & \cdots & n \\ \ell(g\cdot g_1) & \ell(g\cdot g_2) & \cdots & \ell(g\cdot g_n) \end{pmatrix}.$$

By definition, this assigns to each $g \in G$ the permutation whose second row of the tabular notation contains, in order, the labels for each entry in the row of the Cayley table corresponding to g. By this fact, we know that f is one-to-one and onto (see also Theorem 2.13 on page 63). The proof that f is a homomorphism is identical to the proof for Example 5.52: nothing in that argument required x, y, or z to be elements of the Klein 4-group; the proof was for a general group! Hence f is an isomorphism, and $G \cong f(G) < S_n$. □

What's so remarkable about this result? One way of looking at it is the following: since every finite group is isomorphic to a subgroup of a group of permutations, *everything you need to know about finite groups can be learned from studying the groups of permutations!* A more flippant summary is that *the theory of finite groups is all about studying how to rearrange lists.*

In theory, I could go back and rewrite these notes, introducing the reader first to lists, then to permutations, then to S_2, to S_3, to the subgroups of S_4 that correspond to the cyclic group of order 4 and the Klein 4-group, and so forth, making no reference to these other groups, nor to the dihedral group, nor to any other finite group that we have studied. But it is more natural to think in terms other than permutations (geometry for D_n is helpful); and it can be tedious to work only with permutations. While Cayley's Theorem has its uses, it does not suggest that we should always consider groups of permutations in place of the more natural representations.

Exercises.

Exercise 5.53. In Example 5.52 we found W, a subgroup of S_4 that is isomorphic to the Klein 4-group. It turns out that W maps to a subgroup V of D_4, as well. Draw the geometric represen-

tations for each element of V, using a square and writing labels in the appropriate places, as we did in Figures 2.4 on page 68 and 5.3.

Exercise 5.54. Apply Cayley's Theorem to find a subgroup of S_4 that is isomorphic to $\mathbb{Z}_4$. Write the permutations in both tabular and cycle notations.

Exercise 5.55. The subgroup of S_4 that you identified in Exercise 5.54 maps to a subgroup of D_4, as well. Draw the geometric representations for each element of this subgroup, using square with labeled vertices, and arcs to show where the vertices move.

Exercise 5.56. Since S_3 has six elements, we know it is isomorphic to a subgroup of S_6. In fact, it can be isomorphic to more than one subgroup; Cayley's Theorem tells us only that it is isomorphic to *at least* one. Identify a subgroup A of S_6 such that $S_3 \cong A$, yet A is *not* the image of the isomorphism used in the proof of Cayley's Theorem.

5.5: Alternating groups

A special kind of group of permutations, with very important implications for later topics, are the *alternating groups*. To define them, we need to study permutations a little more closely, in particular the cycle notation.

Transpositions

Definition 5.57. Let $n \in \mathbb{N}^+$. An **n-cycle** is a permutation that can be written as one cycle with n entries. A **transposition** is a 2-cycle.

Example 5.58. The permutation $(1\,2\,3) \in S_3$ is a 3-cycle. The permutation $(2\,3) \in S_3$ is a transposition. The permutation $(1\,3)\,(2\,4) \in S_4$ cannot be written as only one n-cycle for any $n \in \mathbb{N}^+$: it is the composition of two disjoint transpositions.

Remark 5.59. *Any transposition is its own inverse.* Why? Consider any transposition $(i\ j)$; it swaps the ith and jth elements of a list. Now consider the product $(i\ j)\,(i\ j)$. The rightmost $(i\ j)$ swaps these two, and the leftmost $(i\ j)$ swaps them back, restoring the list to its original arrangement. Hence $(i\ j)\,(i\ j) = (1)$.

Thanks to 1-cycles, any permutation can be written with many different numbers of cycles: for example,

$$(1\,2\,3) = (1\,2\,3)\,(1) = (1\,2\,3)\,(1)\,(3) = (1\,2\,3)\,(1)\,(3)\,(1) = \cdots.$$

A neat trick allows us to write every permutation as a composition of transpositions.

Example 5.60. Verify that

- $(1\,2\,3) = (1\,3)\,(1\,2)$;
- $(1\,4\,8\,2\,3) = (1\,3)\,(1\,2)\,(1\,8)\,(1\,4)$; and
- $(1) = (1\,2)\,(1\,2)$.

Did you see the relationship between the n-cycle and the corresponding transpositions?

Lemma 5.61. Any permutation can be written as a composition of transpositions.

Proof. You do it! See Exercise 5.72. □

Remark 5.62. Given an expression of σ as a product of transpositions, say $\sigma = \tau_1 \cdots \tau_n$, it is clear from Remark 5.59 that we can write $\sigma^{-1} = \tau_n \cdots \tau_1$, as an application of the associative property yields

$$\begin{aligned}(\tau_1 \cdots \tau_n)(\tau_n \cdots \tau_1) &= (\tau_1 \cdots \tau_{n-1})(\tau_n \tau_n)(\tau_{n-1} \cdots \tau_1) \\ &= (\tau_1 \cdots \tau_{n-1})(\ 1\ )(\tau_{n-1} \cdots \tau_1) \\ &\vdots \\ &= (1)\,.\end{aligned}$$

At this point it is worth looking at Example 5.60 and the discussion before it. Can we write $(\ 1\ \ 2\ \ 3\ )$ with many different numbers of *transpositions*? Yes:

$$\begin{aligned}(1\,2\,3) &= (1\,3)(1\,2) \\ &= (1\,3)(1\,2)(2\,3)(2\,3) \\ &= (1\,3)(1\,2)(1\,3)(1\,3) \\ &= \cdots .\end{aligned}$$

Notice something special about the representation of $(1\,2\,3)$. No matter how you try, you only seem to be able to write it as an *even* number of transpositions. By contrast, consider

$$\begin{aligned}(2\,3) &= (2\,3)(2\,3)(2\,3) \\ &= (2\,3)(1\,2)(1\,3)(1\,3)(1\,2) = \cdots .\end{aligned}$$

No matter how you try, you only seem to be able to write it as an *odd* number of transpositions.

Is this always the case?

Even and odd permutations

Theorem 5.63. Let $\alpha \in S_n$.

- If α can be written as the composition of an even number of transpositions, then it cannot be written as the composition of an odd number of transpositions.
- If α can be written as the composition of an odd number of transpositions, then it cannot be written as the composition of an even number of transpositions.

Proof. Suppose that $\alpha \in S_n$. Consider the polynomials

$$g = \prod_{1 \le i < j \le n} \left(x_i - x_j\right) \qquad \text{and} \qquad g_\alpha = \prod_{1 \le i < j \le n} \left(x_{\alpha(i)} - x_{\alpha(j)}\right).$$

Since the value of g_α depends on the permutation α, and permutations are one-to-one functions, g_α is invariant with respect to the representation of α; that is, it won't change regardless of how we write α in terms of transpositions.

But what, precisely, *is* g_α? Sometimes $g = g_\alpha$; for example, if $\alpha = \begin{pmatrix} 1 & 3 & 2 \end{pmatrix}$ then

$$g = (x_1 - x_2)(x_1 - x_3)(x_2 - x_3)$$

and

$$g_\alpha = (x_3 - x_1)(x_3 - x_2)(x_1 - x_2) = [(-1)(x_1 - x_3)][(-1)(x_2 - x_3)](x_1 - x_2) = g. \tag{15}$$

Is it always the case that $g_\alpha = g$? Not necessarily: if $\alpha = \begin{pmatrix} 1 & 2 \end{pmatrix}$ then $g = x_1 - x_2$ and $g_\alpha = x_2 - x_1 \neq g$. In this case , $g_\alpha = -g$.

Since we cannot guarantee $g_\alpha = g$, can we write g_α in terms of g? Try the following. We know from Lemma 5.61 that α is a composition of transpositions, so let's think about what happens when we compute g_τ for any transposition $\tau = \begin{pmatrix} i & j \end{pmatrix}$. Without loss of generality, we may assume that $i < j$. Let k be another positive integer.

- We know that $x_i - x_j$ is a factor of g. After applying τ, $x_j - x_i$ is a factor of g_τ. This factor of g has changed in g_τ, since $x_j - x_i = -\left(x_i - x_j\right)$.
- If $i < j < k$, then $x_i - x_k$ and $x_j - x_k$ are factors of g. After applying τ, $x_i - x_k$ and $x_j - x_k$ are factors of g_τ. These factors of g have not changed in g_τ.
- If $k < i < j$, then $x_k - x_i$ and $x_k - x_j$ are factors of g. After applying τ, $x_k - x_j$ and $x_k - x_i$ are factors of g_τ.These factors of g have not changed in g_τ.
- If $i < k < j$, then $x_i - x_k$ and $x_k - x_j$ are factors of g. After applying τ, $x_j - x_k$ and $x_k - x_i$ are factors of g_τ. These factors of g have changed in g_τ, but the changes cancel each other out, since

$$\left(x_j - x_k\right)(x_k - x_i) = \left[-\left(x_k - x_j\right)\right][-(x_i - x_k)] = (x_i - x_k)\left(x_k - x_j\right).$$

To summarize: $x_i - x_j$ is the only factor that changes sign *and* does not pair with another factor that changes sign. Thus, $g_\tau = -g$.

Excellent! We have characterized the relationship between g_α and g whenever α is a transposition! Return to the general case, where α is an arbitrary permutation. From Lemma 5.61, α is a composition of transpositions. Choose transpositions $\tau_1, \tau_2, \ldots, \tau_m$ such that $\alpha = \tau_1 \tau_2 \cdots \tau_m$. Using substitution and the observation we just made,

$$g_\alpha = g_{\tau_1 \cdots \tau_m} = -g_{\tau_2 \cdots \tau_m} = (-1)^2 g_{\tau_3 \cdots \tau_m} = \cdots = (-1)^m g.$$

In short,

$$g_\alpha = (-1)^m g. \tag{16}$$

Recall that g_α depends only on α, *and not on its representation.* Assume α can be written as an even number of transpositions; say, $\alpha = \tau_1 \cdots \tau_{2m}$. Formula (16) tells us that $g_\alpha = (-1)^{2m} g = g$. If we could *also* write α as an odd number of transpositions, say, $\alpha = \mu_1 \cdots \mu_{2m+1}$, then $g_\alpha = (-1)^{2k+1} g$. Substitution gives us $(-1)^{2m} g = (-1)^{2k+1} g$; simplification yields $g = -g$, a contradiction. Hence, α cannot be written as an odd number of transpositions.

A similar argument shows that if α can be written as an odd number of transpositions, then it cannot be written as an even number of transpositions. Since $\alpha \in S_n$ was arbitrary, the claim holds. □

Lemma 5.61 tells us that any permutation can be written as a composition of transpositions, and Theorem 5.63 tells us that for any given permutation, this number is always either an even or odd number of transpositions. This relationship merits a definition.

Definition 5.64. If a permutation can be written with an even number of permutations, then we say that the permutation is **even**. Otherwise, we say that the permutation is **odd**.

Example 5.65. The permutation $\rho = (1\,2\,3) \in S_3$ is even, since as we saw earlier $\rho = (1\,3)\,(1\,2)$. So is the permutation $\iota = (1) = (1\,2)\,(1\,2)$.

The permutation $\varphi = (2\,3)$ is odd.

At this point, we are ready to define a new group.

The alternating group

Definition 5.66. Let $n \in \mathbb{N}^+$ and $n \geq 2$. Let $A_n = \{\alpha \in S_n : \alpha \text{ is even}\}$. We call A_n **the set of alternating permutations**.

Remark 5.67. Although A_3 is not the same as "A_3" in Example 3.57 on page 109, the two are isomorphic, because $D_3 \cong S_3$. For this reason, we need not worry about the difference in construction.

Theorem 5.68. For all $n \geq 2$, $A_n < S_n$.

Proof. Let $n \geq 2$, and let $x, y \in A_n$. By the definition of A_n, we can write $x = \sigma_1 \cdots \sigma_{2m}$ and $y = \tau_1 \cdots \tau_{2n}$, where $m, n \in \mathbb{Z}$ and each σ_i or τ_j is a transposition. From Remark 5.62,

$$y^{-1} = \tau_{2n} \cdots \tau_1,$$

so

$$xy^{-1} = (\sigma_1 \cdots \sigma_{2m})\,(\tau_{2n} \cdots \tau_1)\,.$$

Counting the transpositions, we find that xy^{-1} can be written as a product of $2m + 2n = 2(m+n)$ transpositions; in other words, $xy^{-1} \in A_n$. By the Subgroup Theorem, $A_n < S_n$. Thus, A_n is a group. □

How large is A_n, relative to S_n?

Theorem 5.69. For any $n \geq 2$, there are half as many even permutations as there are permutations. That is, $|A_n| = |S_n|\,/2$.

Proof. We show that there are two cosets of $A_n < S_n$, then apply Lagrange's Theorem from page 105.

Let $X \in S_n/A_n$. Let $\alpha \in S_n$ such that $X = \alpha A_n$. If α is an even permutation, then Lemma 3.29 on page 102 implies that $X = A_n$. Otherwise, α is odd. Let β be any other odd permutation. Write out the odd number of transpositions of α^{-1}, followed by the odd number of transpositions of β, to see that $\alpha^{-1}\beta$ is an even permutation. Hence, $\alpha^{-1}\beta \in A_n$, and by Lemma 3.29, $\alpha A_n = \beta A_n$.

We have shown that any coset of A_n is either A_n itself or αA_n for some odd permutation α. Thus, there are only two cosets of A_n in S_n: A_n itself, and the coset of odd permutations. By Lagrange's Theorem,

$$\frac{|S_n|}{|A_n|} = |S_n/A_n| = 2,$$

and a little algebra rewrites this equation as $|A_n| = |S_n|/2$. □

Corollary 5.70. For any $n \geq 2$, $A_n \triangleleft S_n$.

Proof. You do it! See Exercise 5.76. □

There are a number of *exciting* facts regarding A_n that have to wait until later; in particular, A_n has a pivotal effect on whether one can solve polynomial equations by radicals (such as the quadratic formula). In comparison, the facts presented here are relatively dull.

I say that only in comparison, though. The facts presented here are quite striking in their own right: A_n is half the size of S_n, and it is a normal subgroup of S_n. If I call these facts "rather dull", that tells you just how interesting this group can get!

Exercises.

Exercise 5.71. List the elements of A_2, A_3, and A_4 in cycle notation.

Exercise 5.72. Show that any permutation can be written as a product of transpositions.

Exercise 5.73. Show that the inverse of any transposition is a transposition.

Exercise 5.74. Recall the polynomials g and g_α defined in the proof of Theorem 5.63. Compute g_α for the permutations $(1\,3)(2\,4)$ and $(1\,3\,2\,4)$. Use the value of g_α to determine which of the two permutations is odd, and which is even?

Exercise 5.75. Recall the polynomials g and g_α defined in the proof of Theorem 5.63. The **sign function** $\operatorname{sgn}(\alpha)$ is defined to satisfy the property,

$$g = \operatorname{sgn}(\alpha) \cdot g_\alpha.$$

Another way of saying this is that

$$\operatorname{sgn}(\alpha) = \begin{cases} 1, & \alpha \in A_n; \\ -1, & \alpha \notin A_n. \end{cases}$$

Show that for any two cycles α, β,

$$(-1)^{\operatorname{sgn}(\alpha\beta)} = (-1)^{\operatorname{sgn}(\alpha)}(-1)^{\operatorname{sgn}(\beta)}.$$

Exercise 5.76. Show that for any $n \geq 2$, $A_n \triangleleft S_n$.

5.6: The 15-puzzle

The 15-puzzle is similar to a toy you probably played with as a child. It looks like a 4×4 square, with all the squares numbered, except one. The numbering starts in the upper left and proceeds consecutively until the lower right; the only squares that aren't in order are the last two, which are swapped:

1	2	3	4
5	6	7	8
9	10	11	12
13	15	14	

The challenge is to find a way to rearrange the squares so that they are in order, like so:

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	

The only permissible moves are those where one "slides" a square left, right, above, or below the empty square. Given the starting position above, the following first moves are permissible:

1	2	3	4
5	6	7	8
9	10	11	12
13	15		14

or

1	2	3	4
5	6	7	8
9	10	11	
13	15	14	12

.

The following moves are *not*:

1	2	3	4
5	6	7	8
9	10		12
13	15	14	11

or

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	

.

We will use groups of permutations to show that that the challenge is impossible.

How? Since the problem is one of rearranging a list of elements, it is a problem of permutations. Every permissible move consists of transpositions $\tau = (x\ y)$ in S_{16} where:

- $x < y$;
- one of x or y is the position of the empty square in the current list; and
- legal moves imply that either
  - $y = x + 1$ and $x \notin 4\mathbb{Z}$; or
  - $y = x + 4$.

Example 5.77. The legal moves illustrated above correspond to the transpositions

- $(15\ 16)$, because square 14 was in position 15, and the empty space was in position 16: notice that $16 = 15 + 1$; and
- $(12\ 16)$, because square 12 was in position 12, and the empty space was in position 16: notice that $16 = 12 + 4$.

The illegal moves illustrated above correspond to the transpositions

- $(11\ 16)$, because square 11 was in position 11, and the empty space was in position 16: notice that $16 = 11 + 5$; and

- $(13\ 14)$, because in the original configuration, neither 13 nor 14 contains the empty square.

Likewise $(12\ 13)$ would be an illegal move in any configuration, because it crosses rows: even though $y = 13 = 12 + 1 = x + 1$, $x = 12 \in 4\mathbb{Z}$.

How can we use this to show that it is impossible to solve 15-puzzle? We show this in two steps. The first shows that if there is a solution, it must belong to a particular group.

Lemma 5.78. If there is a solution to the 15-puzzle, it is a permutation $\sigma \in A_{16}$, where A_{16} is the alternating group.

Proof. Any permissible move corresponds to a transposition τ as described above. Any solution contains the empty square in the lower right hand corner. As a consequence,

- if $(x\ y)$ is a move left, then the empty square must eventually return to the rightmost row, so there must eventually be a corresponding move $(x'\ y')$ where $[x'] = [x]$ in $\mathbb{Z}_4$ and $[y'] = [y]$ in $\mathbb{Z}_4$; and,
- if $(x\ y)$ is a move up, the empty square must eventually return to the bottom row, so there must eventually be a corresponding move $(x'\ y')$ of the second type.

Thus, moves come in pairs. The upshot is that any solution to the 15-puzzle must be a permutation σ defined by an even number of transpositions. By Theorem 5.63 on page 170 and Definitions 5.64 and 5.66, $\sigma \in A_{16}$. □

We can now show that there is no solution to the 15-puzzle.

Theorem 5.79. The 15-puzzle has no solution.

Proof. By way of contradiction, assume that it has a solution σ. By Lemma 5.78, $\sigma \in A_{16}$. Because A_{16} is a subgroup of S_{16}, and hence a group in its own right, $\sigma^{-1} \in A_{16}$. Notice $\sigma^{-1}\sigma = \iota$, the permutation which corresponds to the configuration of the solution.

Now σ^{-1} is a permutation corresponding to the moves that change the arrangement

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	

into the arrangement

1	2	3	4
5	6	7	8
9	10	11	12
13	15	14	

which corresponds to $(14\ 15)$. Regardless of the transpositions used, the representation must simplify to $\sigma^{-1} = (14\ 15)$. This shows that $\sigma \notin A_{16}$, which contradicts the assumption that we have a contradiction. □

As a historical note, the 15-puzzle was developed in 1878 by an American puzzlemaker, who promised a $1,000 reward to the first person to solve it. Most probably, the puzzlemaker knew

that no one would ever solve it: if we account for inflation, the reward would correspond to $22,265 in 2008 dollars.[15]

The textbook [Lau03] contains a more general discussions of solving puzzles of this sort using algebra.

Exercises

Exercise 5.80. Determine which of these configurations, if any, is solvable by the same rules as the 15-puzzle:

1	2	3	4
5	6	7	8
9	10	12	11
13	14	15	

,

1	2	3	4
5	10	6	8
13	9	7	11
14	15	12	

,

3	6	4	7
1	2	12	8
5	15	10	14
9	13	11	

.

[15] According to the website `http://www.measuringworth.com/ppowerus/result.php`.

Chapter 6:
Number theory

The theory of groups was originally developed to answer questions about the roots of polynomials. From such beginnings, it has grown to many applications that seem at first completely unrelated to this topic. Some of the most widely-used applications in recent decades are in number theory, the study of properties of the integers.

This chapter introduces several of these applications. Section 6.1 fills some background with two of the most important tools in computational algebra and number theory. The first is a fundamental definition; the second is a fundamental algorithm. Both recur throughout the chapter, and later in the notes. Section 6.2 moves us to our first application of group theory, the *Chinese Remainder Theorem*, used thousands of years ago for the task of counting the number of soldiers who survived a battle. We will use it to explain the card trick described on page 1.

The rest of the chapter moves us toward Section 6.6, the RSA cryptographic scheme, a major component of internet communication and commerce. In Section 3.5 you learned of additive clockwork groups; in Section 6.4 you will learn of multiplicative clockwork groups. These allows us to describe in Section 6.5 the theoretical foundation of RSA, Euler's number and Euler's Theorem.

6.1: The Greatest Common Divisor

Until now, we've focused on division with remainder, extending its notion even to cosets of subgroups. Many problems care about divisibility; that is, division with remainder 0.

Common divisors

Recall that we say the integer a divides the integer b when we can find another integer x such that $ax = b$.

Definition 6.1. Let $m, n \in \mathbb{Z}$, not both zero. We say that $d \in \mathbb{Z}$ is a **common divisor of m and n** if $d \mid m$ and $d \mid n$. We say that $d \in \mathbb{N}$ is a **greatest common divisor of m and n** if d is a common divisor *and* any other common divisor d' satisfies $d' < d$.

Example 6.2. Common divisors of 36 and -210 are 1, 2, 3, and 6. The greatest common divisor is 6.

In grade school, you learned how to compute the greatest common divisor of two integers. For example, given the integers 36 and 210, you can find their greatest common divisor, 6. Computing greatest common divisors—not only of integers, but of other objects as well — is an important problem in mathematics, with a large number of important applications. Arguably, it is one of the most important problems in mathematics, and it has an ancient pedigree.

But, do greatest common divisors always exist?

Theorem 6.3. Let $m, n \in \mathbb{Z}$, not both zero. There exists a unique greatest common divisor of m, n.

Algorithm 1. The Euclidean algorithm

```
inputs
    m, n ∈ ℤ
outputs
    gcd(m, n)
do
    Let s = max(m, n)
    Let t = min(m, n)
    repeat while t ≠ 0
        Let q, r ∈ ℤ be the result of dividing s by t
        Let s = t
        Let t = r
    return s
```

Proof. Let D be the set of common divisors of m, n that are also in $\mathbb{N}^+$. Since 1 divides both m and n, we know that $D \neq \emptyset$. We also know that any $d \in D$ must satisfy $d \leq \min(m, n)$; otherwise, the remainder from the Division Algorithm would be nonzero for at least one of m, n. Hence, D is finite. Let d be the largest element of d. By definition of D, d is a common divisor; we claim that it is also the only greatest common divisor. After all, the integers are a linear ordering, so every other common divisor d' of m and n is either

- negative, so that by definition of subtraction, $d - d' \in \mathbb{N}^+$, or (by definition of $<$) $d' < d$; or,
- in D, so that (by definition of d) $d' \leq d$, and $d \neq d'$ implies $d' < d$.

□

How can we compute the greatest common divisor? One way is to make a list of all common divisors, and find the largest. That would require a list of all possible divisors of each integer. In practice, this takes a Very Long Time$^{\text{TM}}$, so we need a different method. One such method was described by the ancient Greek mathematician, Euclid.

The Euclidean Algorithm

Theorem 6.4 (The Euclidean Algorithm)**.** Let $m, n \in \mathbb{Z}$. We can compute the greatest common divisor of m, n in the following way:

1. Let $s = \max(m, n)$ and $t = \min(m, n)$.
2. Repeat the following steps until $t = 0$:
   (a) Let q be the quotient and r the remainder after dividing s by t.
   (b) Assign s the current value of t.
   (c) Assign t the current value of r.

The final value of s is $\gcd(m, n)$.

It is common to write algorithms in a form called *pseudocode*. You can see this done in Algorithm 1.

Before proving that the Euclidean algorithm gives us a correct answer, let's do an example.

Example 6.5. We compute $\gcd(36,210)$. At the outset, let $s = 210$ and $t = 36$. Subsequently:

1. Dividing 210 by 36 gives $q = 5$ and $r = 30$. Let $s = 36$ and $t = 30$.
2. Dividing 36 by 30 gives $q = 1$ and $r = 6$. Let $s = 30$ and $t = 6$.
3. Dividing 30 by 6 gives $q = 5$ and $r = 0$. Let $s = 6$ and $t = 0$.

Now that $t = 0$, we stop, and conclude that $\gcd(36,210) = s = 6$. This agrees with Example 6.2.

To prove that the Euclidean algorithm generates a correct answer, we will number each remainder that we compute; so, the first remainder is r_1, the second, r_2, and so forth. We will then show that the remainders give us a chain of equalities,

$$\gcd(m,n) = \gcd(m,r_1) = \gcd(r_1,r_2) = \cdots = \gcd\left(r_{k-1},0\right),$$

where r_i is the remainder from division of the previous two integers in the chain, and r_{k-1} is the final non-zero remainder from division.

Lemma 6.6. Let $s,t \in \mathbb{Z}$. Let q and r be the quotient and remainder, respectively, of division of s by t, as per the Division Theorem from page 13. Then $\gcd(s,t) = \gcd(t,r)$.

Example 6.7. We can verify Lemma 6.6 using the numbers from Example 6.5. We know that $\gcd(36,210) = 6$. The remainder from division of 36 by 210 is $r = 36$. The lemma claims that $\gcd(36,210) = \gcd(36,30)$; it should be clear to you that $\gcd(36,30) = 6$.

The example also shows that the lemma doesn't care whether $m < n$ or vice versa. We turn to the proof.

Proof of Lemma 6.6. Let $d = \gcd(s,t)$. First we show that d is a divisor of r. From Definition 0.35 on page 15, there exist $a,b \in \mathbb{Z}$ such that $s = ad$ and $t = bd$. By hypothesis, $s = qt + r$ and $0 \leq r < |t|$. Substitution gives us $ad = q(bd) + r$; rewriting the equation, we have

$$r = (a - qb)\,d.$$

By definition of divisibility, $d \mid r$.

Since d is a common divisor of s, t, and r, it is a common divisor of t and r. We claim that $d = \gcd(t,r)$. Let $d' = \gcd(t,r)$; since d is also a common divisor of t and r, the definition of *greatest* common divisor implies that $d \leq d'$. Since d' is a common divisor of t and r, Definition 0.35 again implies that there exist $x,y \in \mathbb{Z}$ such that $t = d'x$ and $r = d'y$. Substituting into the equation $s = qt + r$, we have $s = q\left(d'x\right) + d'y$; rewriting the equation, we have

$$s = (qx + y)\,d'.$$

So $d' \mid s$. We already knew that $d' \mid t$, so d' is a common divisor of s and t.

Recall that $d = \gcd(s,t)$; since d' is also a common divisor of t and r, the definition of *greatest* common divisor implies that $d' \leq d$. Earlier, we showed that $d \leq d'$. Hence $d \leq d' \leq d$, which implies that $d = d'$.

Substitution gives the desired conclusion: $\gcd(s,t) = \gcd(t,r)$. □

We can finally prove that the Euclidean algorithm gives us a correct answer. This requires two stages, necessary for any algorithm.

1. **Correctness.** If the algorithm terminates, we have to guarantee that it terminates with the correct answer.
2. **Termination.** What if the algorithm doesn't terminate? If you look at the Euclidean algorithm, you see that one of its instructions asks us to repeat some steps "while $t \neq 0$." What if t never attains the value of zero? It's conceivable that its values remain positive at all times, or jump over zero from positive to negative values. That would mean that we never receive any answer from the algorithm, let alone a correct one.

We will identify both stages of the proof clearly. In addition, we will refer back to the the Division Theorem as well as the well-ordering property of the integers from Section 10; you may wish to review those.

Proof of the Euclidean Algorithm. We start with termination. The only repetition in the algorithm occurs in line 8. The first time we compute line 9, we compute the quotient q and remainder r of division of s by t. By the Division Theorem,

$$0 \leq r < |t|. \tag{17}$$

Denote this value of r by r_1. In the next lines we set s to t, then t to $r_1 = r$. Thanks to equation (17), the size of $t_{\text{new}} = r$ is smaller than that of $s_{\text{new}} = t_{\text{old}}$. (We measure "size" using absolute value.) If $t \neq 0$, then we return to line 9 and divide s by t, again obtaining a new remainder r. Denote this value of r by r_2; by the Division Theorem, $r_2 = r < t$, so

$$0 \leq r_2 < r_1.$$

Proceeding in this fashion, we generate a strictly decreasing sequence of elements,

$$r_1 > r_2 > r_3 > \cdots.$$

By Exercise 0.31, this sequence is finite. In other words, the algorithm terminates.

We now show that the algorithm terminates *with the correct answer*. If line 9 of the algorithm repeated a total of k times, then $r_k = 0$. Apply Lemma 6.6 repeatedly to the remainders to obtain the chain of equalities

$$\begin{aligned}
r_{k-1} = \gcd(0, r_{k-1}) &= \gcd(r_k, r_{k-1}) && \text{(definition of gcd, substitution)} \\
&= \gcd(r_{k-1}, r_{k-2}) && \text{(Lemma 6.6)} \\
&= \gcd(r_{k-2}, r_{k-3}) && \text{(Lemma 6.6)} \\
&\vdots \\
&= \gcd(r_2, r_1) && \text{(Lemma 6.6)} \\
&= \gcd(r_1, s) && \text{(substitution)} \\
&= \gcd(t, s) && \text{(substitution)} \\
&= \gcd(m, n). && \text{(substitution)}
\end{aligned}$$

The Euclidean Algorithm terminates with the correct answer. □

Bezout's identity

A fundamental fact of number theory is that the greatest common divisor of two integers can be expressed as a simple expression of those integers.

Theorem 6.8 (Bezout's Lemma, or, the Extended Euclidean Algorithm)**.** Let $m, n \in \mathbb{Z}$. There exist $a, b \in \mathbb{Z}$ such that $am + bn = \gcd(m, n)$. Both a and b can be found by reverse-substituting the chain of equations obtained by the repeated division in the Euclidean algorithm.

The expression, $am + bn = \gcd(m, n)$, is important enough to be known by the name, **Bezout's identity**. It can be used to prove a *lot* of properties of greatest common divisors.

Example 6.9. Recall from Example 6.5 the computation of $\gcd(210, 36)$. The divisions gave us a series of equations:

$$210 = 5 \cdot 36 + 30 \tag{18}$$
$$36 = 1 \cdot 30 + 6 \tag{19}$$
$$30 = 5 \cdot 6 + 0.$$

We concluded from the Euclidean Algorithm that $\gcd(210, 36) = 6$. The Extended Euclidean Algorithm gives us a way to find $a, b \in \mathbb{Z}$ such that $6 = 210a + 36b$. Start by rewriting equation (19):

$$36 - 1 \cdot 30 = 6. \tag{20}$$

This looks a little like what we want, but we need 210 instead of 30. Equation (18) allows us to rewrite 30 in terms of 210 and 36:

$$30 = 210 - 5 \cdot 36. \tag{21}$$

Substituting this result into equation (20), we have

$$36 - 1 \cdot (210 - 5 \cdot 36) = 6 \quad \Longrightarrow \quad 6 \cdot 36 + (-1) \cdot 210 = 6.$$

We have found integers $m = 6$ and $n = -1$ such that for $a = 36$ and $b = 210$, $\gcd(a, b) = 6$.

The method we applied in Example (6.9) is what we use both to prove correctness of the algorithm, and to find a and b in general.

Proof of the Extended Euclidean Algorithm. Look back at the proof of the Euclidean algorithm

to see that it computes a chain of k quotients $\{q_i\}$ and remainders $\{r_i\}$ such that

$$\begin{aligned} m &= q_1 n + r_1 \\ n &= q_2 r_1 + r_2 \\ r_1 &= q_3 r_2 + r_3 \\ &\vdots \\ r_{k-3} &= q_{k-1} r_{k-2} + r_{k-1} \qquad (22) \\ r_{k-2} &= q_k r_{k-1} + r_k \qquad (23) \\ r_{k-1} &= q_{k+1} r_k + 0 \\ \text{and } r_k &= \gcd(m,n). \end{aligned}$$

Rewrite equation (23) as

$$r_{k-2} = q_k r_{k-1} + \gcd(m,n).$$

Solving for $\gcd(m,n)$, we have

$$r_{k-2} - q_k r_{k-1} = \gcd(m,n). \qquad (24)$$

Solve for r_{k-1} in equation (22) to obtain

$$r_{k-3} - q_{k-1} r_{k-2} = r_{k-1}.$$

Substitute this into equation (24) to obtain

$$\begin{aligned} r_{k-2} - q_k \left(r_{k-3} - q_{k-1} r_{k-2} \right) &= \gcd(m,n) \\ \left(q_{k-1} + 1 \right) r_{k-2} - q_k r_{k-3} &= \gcd(m,n). \end{aligned}$$

Proceeding in this fashion, we exhaust the list of equations, concluding by rewriting the first equation in the form $am + bn = \gcd(m,n)$ for some integers a, b. □

Pseudocode appears in Algorithm 2. One can also derive a method of computing both $\gcd(m,n)$ and the representation $am + bn = \gcd(m,n)$ simultaneously, which is to say, without having to reverse the process. We will not consider that here.

Exercises.

Exercise 6.10. Compute the greatest common divisor of 100 and 140 by (a) listing all divisors, then identifying the largest; and (b) the Euclidean Algorithm.

Exercise 6.11. Compute the greatest common divisor of $m = 4343$ and $n = 4429$ by the Euclidean Algorithm. Use the Extended Euclidean Algorithm to find $a, b \in \mathbb{Z}$ that satisfy Bezout's identity.

Exercise 6.12. Show that any common divisor of any two integers divides the integers' greatest common divisor.

Algorithm 2. Extended Euclidean Algorithm

inputs
  $m, n \in \mathbb{N}$ such that $m > n$
outputs
  $\gcd(m,n)$ and $a, b \in \mathbb{Z}$ such that $\gcd(m,n) = am + bn$
do
  if $n = 0$
    Let $d = m$, $a = 1$, $b = 0$
  else
    Let $r_0 = m$ and $r_1 = n$
    Let $k = 1$
    repeat while $r_k \neq 0$
      Increment k by 1
      Let q_k, r_k be the quotient and remainder from division of r_{k-2} by r_{k-1}
    Let $d = r_{k-1}$ and $p = r_{k-3} - q_{k-1} r_{k-2}$ (do not simplify p)
    Decrement k by 2
    repeat while $k \geq 2$
      Substitute $r_k = r_{k-2} - q_k r_{k-1}$ into p
      Decrement k by 1
    Let a be the coefficient of r_0 in p, and b be the coefficient of r_1 in p
  return d, a, b

Exercise 6.13. In Lemma 6.6 we showed that $\gcd(m,n) = \gcd(m,r)$ where r is the remainder after division of m by n. Prove the following more general statement: for all $m, n, q \in \mathbb{Z}$ $\gcd(m,n) = \gcd(n, m - qn)$.

Exercise 6.14. Bezout's Identity (Theorem 6.8) states that for any $m, n \in \mathbb{Z}$, we can find $a, b \in \mathbb{Z}$ such that $am + bn = \gcd(m,n)$.

(a) Show that the existence of $a, b, d \in \mathbb{Z}$ such that $am + bn = d$ does *not* imply $d = \gcd(m,n)$.

(b) However, not only does the converse of Bezout's Identity hold, we can specify the relationship more carefully. Fill in each blank of Figure 6.1 with the appropriate justification or statement.

6.2: The Chinese Remainder Theorem

In this section we explain how the card trick on page 1 works. The result is based on an old Chinese observation.[16] Recall from Section 3.5 that for any $m \neq 0$ there exists a group $\mathbb{Z}_m$ of m elements, under the operation of adding, then taking remainder after division by m. Remember that we often write $[x]$ for the elements of $\mathbb{Z}_m$ if we want to emphasize that its elements are cosets.

[16] I asked Dr. Ding what the Chinese call this theorem. He looked it up in one of his books, and told me that they call it Sun Tzu's Theorem. Unfortunately, this is not the same Sun Tzu who wrote *The Art of War*.

Let $m, n \in \mathbb{Z}$, $S = \{am + bn : a, b \in \mathbb{Z}\}$, and $M = S \cap \mathbb{N}$. Since M is a subset of $\mathbb{N}$, the well-ordering property of $\mathbb{Z}$ implies that it has a smallest element; call it d.
Claim: $d = \gcd(m, n)$.
Proof:

1. We first claim that $\gcd(m, n)$ divides d.
   (a) By _____, we can find $a, b \in \mathbb{Z}$ such that $d = am + bn$.
   (b) By _____, $\gcd(m, n)$ divides m and n.
   (c) By _____, there exist $x, y \in \mathbb{Z}$ such that $m = x \gcd(m, n)$ and $n = y \gcd(m, n)$.
   (d) By susbtitution, _____.
   (e) Collect the common term to obtain _____.
   (f) By _____, $\gcd(m, n)$ divides d.
2. A similar argument shows that d divides $\gcd(m, n)$.
3. By _____, $d \leq \gcd(m, n)$ and $\gcd(m, n) \leq d$.
4. By _____, $d = \gcd(m, n)$.

Figure 6.1. Material for Exercise 6.14

The simple Chinese Remainder Theorem

Theorem 6.15 (The Chinese Remainder Theorem, simple version)**.** Let $m, n \in \mathbb{Z}$ such that $\gcd(m, n) = 1$. Let $\alpha, \beta \in \mathbb{Z}$. There exists a solution $x \in \mathbb{Z}$ to the system of linear congruences

$$\begin{cases} [x] = [\alpha] \text{ in } \mathbb{Z}_m; \\ [x] = [\beta] \text{ in } \mathbb{Z}_n; \end{cases}$$

and $[x]$ is unique in $\mathbb{Z}_N$ where $N = mn$.

Before giving a proof, let's look at an example.

Example 6.16 (The card trick)**.** In the card trick, we took twelve cards and arranged them

- once in groups of three; and
- once in groups of four.

Each time, the player identified the *column* in which the mystery card lay. Laying out the cards in rows of three and four corresponds to division by three and four, so that α and β are in fact the remainders from division by three and by four. This corresponds to a system of linear congruences,

$$\begin{cases} [x] = [\alpha] \text{ in } \mathbb{Z}_3 \\ [x] = [\beta] \text{ in } \mathbb{Z}_4 \end{cases},$$

where x is the location of the mystery card. The simple version of the Chinese Remainder Theorem guarantees a solution for x, which is unique in $\mathbb{Z}_{12}$. Since there are only twelve cards, the solution is unique in the game: as long as the dealer can compute x, s/he can identify the card infallibly.

"Well, and good," you think, "but knowing only the existence of a solution seems rather pointless. I also need to know *how* to compute x, so that I can pinpoint the location of the card."

It turns out that Bezout's identity,

$$am + bn = \gcd(m,n),$$

is the key to unlocking the Chinese Remainder Theorem. Before doing so, we need an important lemma about numbers whose gcd is 1.

Lemma 6.17. Let $d, m, n \in \mathbb{Z}$. If $m \mid nd$ and $\gcd(m,n) = 1$, then $m \mid d$.

Proof. Assume that $m \mid nd$ and $\gcd(m,n) = 1$. By definition of divisibility, there exists $q \in \mathbb{Z}$ such that $qm = nd$. Use the Extended Euclidean Algorithm to choose $a, b \in \mathbb{Z}$ such that $am + bn = \gcd(m,n) = 1$. Multiplying both sides of this equation by d, we have

$$\begin{aligned} (am + bn)\, d &= 1 \cdot d \\ amd + b\,(nd) &= d \\ adm + b\,(qm) &= d \\ (ad + bq)\, m &= d. \end{aligned}$$

Hence $m \mid d$. □

Now we prove the Chinese Remainder Theorem. You should study this proof carefully, not only to understand the theorem better, but because the proof tells you how to solve the system.

Proof of the Chinese Remainder Theorem, simple version. Recall that the system is

$$\begin{cases} [x] = [\alpha] \text{ in } \mathbb{Z}_m \\ [x] = [\beta] \text{ in } \mathbb{Z}_n \end{cases}.$$

We have to prove two things: first, that a solution x exists; second, that $[x]$ is unique in $\mathbb{Z}_N$.

Existence: Because $\gcd(m,n) = 1$, the Extended Euclidean Algorithm tells us that there exist $a, b \in \mathbb{Z}$ such that $am + bn = 1$. Rewriting this equation two different ways, we have $bn = 1 + (-a)\, m$ and $am = 1 + (-b)\, n$. In terms of cosets of subgroups of $\mathbb{Z}$, these two equations tell us that $bn \in 1 + m\mathbb{Z}$ and $am \in 1 + n\mathbb{Z}$. In the bracket notation, $[bn]_m = [1]_m$ and $[am]_n = [1]_n$. By Lemmas 3.80 and 3.83 on page 117, $[\alpha]_m = \alpha\,[1]_m = \alpha\,[bn]_m = [\alpha bn]_m$ and likewise $[\beta]_n = [\beta am]_n$. Apply similar reasoning to see that $[\alpha bn]_n = [0]_n$ and $[\beta am]_m = [0]_m$ in $\mathbb{Z}_m$. Hence,

$$\begin{cases} [\alpha bn + \beta am]_m = [\alpha]_m \\ [\alpha bn + \beta am]_n = [\beta]_n \end{cases}.$$

If we let $x = \alpha bn + \beta am$, then the equations above show that x is a solution to the system.

Uniqueness: Suppose that there exist $[x], [y] \in \mathbb{Z}_N$ that both satisfy the system. Since $[x] = [\alpha] = [y]$ in $\mathbb{Z}_m$, $[x - y] = [0]$, and by Lemma 3.86 on page 119, $m \mid (x - y)$. A similar argument shows that $n \mid (x - y)$. By definition of divisibility, there exists $q \in \mathbb{Z}$ such that $mq = x - y$. By substitution, $n \mid mq$. By Lemma 6.17, $n \mid q$. By definition of divisibility, there exists $q' \in \mathbb{Z}$ such that $q = nq'$. By substitution,

$$x - y = mq = mnq' = Nq'.$$

Algorithm 3. Solution to Chinese Remainder Theorem, simple version

inputs
  $m, n \in \mathbb{Z}$ such that $\gcd(m,n) = 1$
  $\alpha, \beta \in \mathbb{Z}$
outputs
  $x \in \mathbb{Z}$ satisfying the Chinese Remainder Theorem
do
  Use the Extended Euclidean Algorithm to find $a, b \in \mathbb{Z}$ such that $am + bn = 1$
  return $[\alpha bn + \beta am]_N$

Hence $N \mid (x-y)$, and again by Lemma 3.86 $[x]_N = [y]_N$, which means that the solution x is unique in $\mathbb{Z}_N$, as desired. □

Pseudocode to solve the Chinese Remainder Theorem appears as Algorithm 3.

Example 6.18. The algorithm of Corollary 3 finally explains the method of the card trick. We have $m = 3$, $n = 4$, and $N = 12$. Suppose that the player indicates that his card is in the first column when they are grouped by threes, and in the third column when they are grouped by fours; then $\alpha = 1$ and $\beta = 3$.

Using the Extended Euclidean Algorithm, we find that $a = -1$ and $b = 1$ satisfy $am + bn = 1$; hence $am = -3$ and $bn = 4$. We can therefore find the mystery card by computing

$$x = 1 \cdot 4 + 3 \cdot (-3) = -5.$$

Its canonical representation in $\mathbb{Z}_{12}$ is

$$[x] = [-5 + 12] = [7],$$

which implies that the player chose the 7th card. In fact, $[7] = [1]$ in $\mathbb{Z}_3$, and $[7] = [3]$ in $\mathbb{Z}_4$, which agrees with the information given.

The Chinese Remainder Theorem can be generalized to larger systems with more than two equations under certain circumstances.

A generalized Chinese Remainder Theorem

What if you have more than just two ways to arrange the groups? You might like to arrange the cards into rows of 3, 4, 5, and 7. What about other groupings? What constraints do there have to be on the groupings, and how would we solve the new problem?

Theorem 6.19 (Chinese Remainder Theorem on $\mathbb{Z}$)**.** Let $m_1, m_2, \ldots, m_n \in \mathbb{Z}$ and assume that $\gcd\left(m_i, m_j\right) = 1$ for all $1 \leq i < j \leq n$. Let $\alpha_1, \alpha_2, \ldots \alpha_n \in \mathbb{Z}$. There exists a solution $x \in \mathbb{Z}$ to the system of linear congruences

$$\begin{cases} [x] &= [\alpha_1] \text{ in } \mathbb{Z}_{m_1}; \\ [x] &= [\alpha_2] \text{ in } \mathbb{Z}_{m_2}; \\ &\vdots \\ [x] &= [\alpha_n] \text{ in } \mathbb{Z}_{m_n}; \end{cases}$$

and $[x]$ is unique in $\mathbb{Z}_N$ where $N = m_1 m_2 \cdots m_n$.

Before we can prove this version of the Chinese Remainder Theorem, we need to make an observation of $m_1, m_2, \ldots, m_n$.

Lemma 6.20. Let $m_1, m_2, \ldots, m_n \in \mathbb{Z}$ such that $\gcd\left(m_i, m_j\right) = 1$ for all $1 \leq i < j \leq n$. For each $i = 1, 2, \ldots, n$ define $N_i = N/m_i$ where $N = m_1 m_2 \cdots m_n$; that is, N_i is the product of all the m's except m_i. Then $\gcd(m_i, N_i) = 1$.

Proof. We show that $\gcd(m_1, N_1) = 1$; for $i = 2, \ldots, n$ the proof is similar.

Use the Extended Euclidean Algorithm to choose $a, b \in \mathbb{Z}$ such that $am_1 + bm_2 = 1$. Use it again to choose $c, d \in \mathbb{Z}$ such that $cm_1 + dm_3 = 1$. Then

$$\begin{aligned} 1 &= (am_1 + bm_2)(cm_1 + dm_3) \\ &= (acm_1 + adm_3 + bcm_2)m_1 + (bd)(m_2 m_3). \end{aligned}$$

Let $x = \gcd(m_1, m_2 m_3)$; since x divides both m_1 and $m_2 m_3$, it divides each term of the right hand side above. That right hand side equals 1, so x also divides 1. The only divisors of 1 are ± 1, so $x = 1$. We have shown that $\gcd(m_1, m_2 m_3) = 1$.

Rewrite the equation above as $1 = a'm_1 + b'm_2 m_3$; notice that $a', b' \in \mathbb{Z}$. Use the Extended Euclidean Algorithm to choose $e, f \in \mathbb{Z}$ such that $em_1 + fm_4 = 1$. Then

$$\begin{aligned} 1 &= \left(a'm_1 + b'm_2 m_3\right)(em_1 + fm_4) \\ &= \left(a'em_1 + a'fm_4 + b'em_2 m_e\right)m_1 + \left(b'f\right)(m_2 m_3 m_4). \end{aligned}$$

An argument similar to the one above shows that $\gcd(m_1, m_2 m_3 m_4) = 1$.

Repeating this process with each m_i, we obtain $\gcd(m_1, m_2 m_3 \cdots m_n) = 1$. Since $N_1 = m_2 m_3 \cdots m_n$, we have $\gcd(m_1, N_1) = 1$. □

We can now prove the Chinese Remainder Theorem on $\mathbb{Z}$.

Proof of the Chinese Remainder Theorem on $\mathbb{Z}$. *Existence:* Write $N_i = N/m_i$ for $i = 1, 2, \ldots, n$. By Lemma 6.20, $\gcd(m_i, N_i) = 1$. Use the Extended Euclidean Algorithm to compute appropri-

ate a's and b's satisfying

$$\begin{aligned} a_1 m_1 + b_1 N_1 &= 1 \\ a_2 m_2 + b_2 N_2 &= 1 \\ &\vdots \\ a_n m_n + b_n N_n &= 1. \end{aligned}$$

Put $x = \alpha_1 b_1 N_1 + \alpha_2 b_2 N_2 + \cdots + \alpha_n b_n N_n$. Now, $b_1 N_1 = 1 + (-a_1)\, m_1$, so $[b_1 N_1] = [1]$ in $\mathbb{Z}_{m_1}$, so $[\alpha_1 b_1 N_1] = [\alpha_1]$ in $\mathbb{Z}_{m_1}$. Moreover, for any $i = 2, 3, \ldots, n$, inspection of N_i verifies that $m_1 \mid N_i$, implying that $[\alpha_i b_i N_i]_{m_1} = [0]_{m_1}$ (Lemma 3.86). Hence

$$\begin{aligned} [x] &= [\alpha_1 b_1 N_1 + \alpha_2 b_2 N_2 + \cdots + \alpha_n b_n N_n] \\ &= [\alpha_1] + [0] + \cdots + [0] \end{aligned}$$

in $\mathbb{Z}_{m_1}$, as desired. A similar argument shows that $[x] = [\alpha_i]$ in $\mathbb{Z}_{m_i}$ for $i = 2, 3, \ldots, n$.

Uniqueness: As in the previous case, let $[x], [y]$ be two solutions to the system in $\mathbb{Z}_N$. Then $[x - y] = [0]$ in $\mathbb{Z}_{m_i}$ for $i = 1, 2, \ldots, n$, implying that $m_i \mid (x - y)$ for $i = 1, 2, \ldots, n$.

Since $m_1 \mid (x - y)$, the definition of divisibility implies that there exists $q_1 \in \mathbb{Z}$ such that $x - y = m_1 q_1$.

Since $m_2 \mid (x - y)$, substitution implies $m_2 \mid m_1 q_1$, and Lemma 6.17 implies that $m_2 \mid q_1$. The definition of divisibility implies that there exists $q_2 \in \mathbb{Z}$ such that $q_1 = m_2 q_2$. Substitution implies that $x - y = m_1 m_2 q_2$.

Since $m_3 \mid (x - y)$, substitution implies $m_3 \mid m_1 m_2 q_2$. By Lemma 6.20, $\gcd(m_1 m_2, m_3) = 1$, and Lemma 6.17 implies that $m_3 \mid q_2$. The definition of divisibility implies that there exists $q_3 \in \mathbb{Z}$ such that $q_2 = m_3 q_3$. Substitution implies that $x - y = m_1 m_2 m_3 q_3$.

Continuing in this fashion, we show that $x - y = m_1 m_2 \cdots m_n q_n$ for some $q_n \in \mathbb{Z}$. By substition, $x - y = N q_n$, so $[x - y] = [0]$ in $\mathbb{Z}_N$, so $[x] = [y]$ in $\mathbb{Z}_n$. That is, the solution to the system is unique in $\mathbb{Z}_N$. □

The algorithm to solve such systems is similar to that given for the simple version, in that it can be obtained from the proof of existence of a solution.

Exercises

Exercise 6.21. Solve the system of linear congruences

$$\begin{cases} [x] = [2] \text{ in } \mathbb{Z}_4 \\ [x] = [3] \text{ in } \mathbb{Z}_9 \end{cases}.$$

Express your answer so that $0 \leq x < 36$.

Exercise 6.22. Solve the system of linear congruences

$$\begin{cases} [x] = [2] \text{ in } \mathbb{Z}_5 \\ [x] = [3] \text{ in } \mathbb{Z}_6 \\ [x] = [4] \text{ in } \mathbb{Z}_7 \end{cases}.$$

Exercise 6.23. Solve the system of linear congruences

$$\begin{cases} [x] = [33] \text{ in } \mathbb{Z}_{16} \\ [x] = [-4] \text{ in } \mathbb{Z}_{33} \\ [x] = [17] \text{ in } \mathbb{Z}_{504} \end{cases}.$$

This problem is a little tougher than the previous, since $\gcd(16,504) \neq 1$ and $\gcd(33,504) \neq 1$. Since you can't use either of the Chinese Remainder Theorems presented here, you'll have to generalize their approaches to get a method for this one.

Exercise 6.24. Give directions for a similar card trick on all 52 cards, where the cards are grouped first by 4's, then by 13's. Do you think this would be a practical card trick?

Exercise 6.25. Is it possible to modify the card trick to work with only ten cards instead of 12? If so, how; if not, why not?

Exercise 6.26. Is it possible to modify the card trick to work with only eight cards instead of 12? If so, how; if not, why not?

6.3: The Fundamental Theorem of Arithmetic

In this section, we address a fundamental result of number theory with algebraic implications.

Definition 6.27. Let $n \in \mathbb{N}^+ \setminus \{1\}$. We say that n is **irreducible** if the only integers that divide n are ± 1 and $\pm n$.

You may read this and think, "Oh, he's talking about prime numbers." Yes and no. We'll say more about that in a moment.

Example 6.28. The integer 36 is not irreducible, because $36 = 6 \times 6$. The integer 7 is irreducible, because the only integers that divide 7 are ± 1 and ± 7.

One useful aspect to irreducible integers is that, aside from ± 1, any integer is divisible by at least one irreducible integer.

Theorem 6.29. Let $n \in \mathbb{Z} \setminus \{\pm 1\}$. There exists at least one irreducible integer p such that $p \mid n$.

Proof. *Case 1:* If $n = 0$, then 2 is a divisor of n, and we are done.

Case 2: Assume that $n \in \mathbb{N}^+ \setminus \{1\}$. If n is not irreducible, then by definition $n = a_1 b_1$ such that $a_1, b_1 \in \mathbb{Z}$ and $a_1, b_1 \neq \pm 1$. Without loss of generality, we may assume that $a_1, b_1 \in \mathbb{N}^+$ (otherwise both a, b are negative and we can replace them with their opposites). Observe further that $a_1 < n$ (this is a consequence of Exercise 0.26 on page 12). If a_1 is irreducible, then we are done; otherwise, we can write $a_1 = a_2 b_2$ where $a_2, b_2 \in \mathbb{N}^+$ and $a_2 < a_1$.

Let $a_0 = n$. As long as a_i is not irreducible, we can find a_{i+1}, $b_{i+1} \in \mathbb{N}^+$ such that $a_i = a_{i+1}b_{i+1}$. By Exercise 0.26, $a_i > a_{i+1}$ for each i. Proceeding in this fashion, we generate a strictly decreasing sequence of elements,

$$a_0 > a_1 > a_2 > \cdots .$$

By Exercise 0.31, this sequence *must* be finite. Let a_m be the final element in the sequence. We claim that a_m is irreducible; after all, if it were not irreducible, then we could extend the sequence further, and we cannot. By substitution,

$$n = a_1 b_1 = a_2 (b_2 b_1) = \cdots = a_m \left(b_{m-1} \cdots b_1\right).$$

That is, a_m is an irreducible integer that divides n.

Case 3: Assume that $n \in \mathbb{Z} \setminus (\mathbb{N} \cup \{-1\})$. Let $m = -n$. Since $m \in \mathbb{N}^+ \setminus \{1\}$, Case 2 implies that there exists an irreducible integer p such that $p \mid m$. By definition, $m = qp$ for some $q \in \mathbb{Z}$. By substitution and properties of arithmetic, $n = -(qp) = (-q)p$, so $p \mid n$. □

Let's turn now to the term you might have expected for the definition given above: a *prime* number. For reasons that you will learn later, we actually associate a different notion with this term.

Definition 6.30. Let $p \in \mathbb{N}^+ \setminus \{1\}$. We say that p is **prime** if for any two integers a, b

$$p \mid ab \quad \Longrightarrow \quad p \mid a \text{ or } p \mid b.$$

Example 6.31. Let $a = 68$ and $b = 25$. It is easy to recognize that 10 divides $ab = 1700$. However, 10 divides neither a nor b, so 10 is not a prime number.

It is also easy to recognize that 17 divides $ab = 1700$. Unlike 10, 17 divides one of a or b; in fact, it divides a. Were we to look at every possible product ab divisible by 17, we would find that 17 always divides one of the factors a or b. Thus, 17 is prime.

If the next-to-last sentence in the example, bothers you, *good.* I've claimed something about every product divisible by 17, but haven't explained why that is true. That's cheating! If I'm going to claim that 17 is prime, I need a better explanation than, "look at every possible product ab." After all, there are an infinite number of products possible, and we can't do that in finite time. We need a *finite* criterion.

To this end, let's return to the notion of an irreducible number. Previously, you were probably taught that a *prime* number was what we have here called *irreducible*. I've now given a definition that seems different.

Could it be that the definitions are *distinctions without a difference?* Indeed, they are equivalent!

Theorem 6.32. An integer is prime if and only if it is irreducible.

Proof. This proof has two parts. You will show in Exercise 6.34 that if an integer is prime, then it is irreducible. Here, we show the converse.

Let $n \in \mathbb{N}^+ \setminus \{1\}$ and assume that n is irreducible. To show that n is prime, we must take arbitrary $a, b \in \mathbb{Z}$ and show that if $n \mid ab$, then $n \mid a$ or $n \mid b$. Therefore, let $a, b \in \mathbb{Z}$ and assume that $n \mid ab$. If $n \mid a$, then we would be done, so assume that $n \nmid a$. We must show that $n \mid b$.

By definition, the common factors of n and a are a subset of the factors of n. Since n is irreducible, its factors are ± 1 and $\pm n$. By hypothesis, $n \nmid a$, so $\pm n$ cannot be common factors of n and a. Thus, the only common factors of n and a are ± 1, which means that $\gcd(n,a) = 1$. By Lemma 6.17, $n \mid b$.

We assumed that if n is irreducible and divides ab, then n must divide one of a or b. By definition, n is prime. □

If the two definitions are equivalent, why would we give a different definition? It turns out that the concepts are equivalent *for the integers,* but not for other sets; you will see this later in Sections 8.4 and 10.1.

The following theorem is a cornerstone of Number Theory.

Theorem 6.33 (The Fundamental Theorem of Arithmetic)**.** Let $n \in \mathbb{N}^+ \setminus \{1\}$. We can **factor n into irreducibles**; that is, we can write

$$n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_r^{\alpha_r}$$

where $p_1, p_2, \ldots, p_r$ are irreducible and $\alpha_1, \alpha_2, \ldots, \alpha_r \in \mathbb{N}$. The representation is unique if we order $p_1 < p_2 < \ldots < p_n$.

Since prime integers are irreducible and vice versa, you can replace "irreducible" by "prime" and obtain the expression of this theorem found more commonly in number theory textboks. We use "irreducible" here to lay the groundwork for Definition 10.16 on page 287.

Proof. The proof has two parts: a proof of existence and a proof of uniqueness.

Existence: We proceed by induction on positive integers.

Inductive base: If $n = 2$, then n is irreducible, and we are finished.

Inductive hypothesis: Assume that the integers 2, 3, …, $n-1$ have a factorization into irreducibles.

Inductive step: If n is irreducible, then we are finished. Otherwise, n is not irreducible. By Lemma 6.29, there exists an irreducible integer p_1 such that $p_1 \mid n$. By definition, there exists $q \in \mathbb{N}^+$ such that $n = qp_1$. Since $p_1 \neq 1$, Exercise 0.44 tells us that $q < n$. By the inductive hypothesis, q has a factorization into irreducibles; say

$$q = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_r^{\alpha_r}.$$

Thus $n = qp = p_1^{\alpha_1+1} p_2^{\alpha_2} \cdots p_r^{\alpha_r}$; that is, n factors into irreducibles.

Uniqueness: Here we use the fact that irreducible numbers are also prime (Lemma 6.32). Assume that $p_1 < p_2 < \cdots < p_r$ and we can factor n as

$$n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_r^{\alpha_r} = p_1^{\beta_1} p_2^{\beta_2} \cdots p_r^{\beta_r}.$$

Without loss of generality, we may assume that $\alpha_1 \leq \beta_1$. It follows that

$$p_2^{\alpha_2} p_3^{\alpha_3} \cdots p_r^{\alpha_r} = p_1^{\beta_1-\alpha_1} p_2^{\beta_2} p_3^{\beta_3} \cdots p_r^{\beta_r}.$$

This equation implies that $p_1^{\beta_1-\alpha_1}$ divides the expression on the left hand side of the equation. Since p_1 is irreducible, hence prime, $\beta_1 - \alpha_1 \neq 0$ implies that p_1 divides one of $p_2, p_3, \ldots, p_r$.

Claim: If p is irreducible, then $\sqrt{p}$ is not rational.
Proof:

1. Assume that p is irreducible.
2. By way of contradiction, assume that $\sqrt{p}$ *is* rational.
3. By _____, there exist $a, b \in \mathbb{N}$ such that $\sqrt{p} = a/b$.
4. Without loss of generality, we may assume that $\gcd(a,b) = 1$.
   (After all, we could otherwise rewrite $\sqrt{p} = (a/d)\,/\,(b/d)$, where $d = \gcd(a,b)$.)
5. By _____, $p = a^2/b^2$.
6. By _____, $pb^2 = a^2$.
7. By _____, $p \mid a^2$.
8. By _____, p is prime.
9. By _____, $p \mid a$.
10. By _____, $a = pq$ for some $q \in \mathbb{Z}$.
11. By _____ and _____, $pb^2 = (pq)^2 = p^2q^2$.
12. By _____, $b^2 = pq^2$.
13. By _____, $p \mid b^2$.
14. By _____, $p \mid b$.
15. This contradicts step _____. Our assumption that $\sqrt{p}$ is rational must have been wrong. Hence, $\sqrt{p}$ is irrational.

Figure 6.2. Material for Exercise 6.36

This contradicts the irreducibility of p_2, p_3, …, p_r. Hence $\beta_1 - \alpha_1 = 0$. A similar argument shows that $\beta_i = \alpha_i$ for all $i = 1, 2, \ldots, r$; hence the representation of n as a product of irreducible integers is unique. □

Exercises.

Exercise 6.34. Show that any prime integer p is irreducible.

Exercise 6.35. Show that there are infinitely many irreducible integers.

Exercise 6.36. Fill in each blank of Figure 6.2 with the justification.

Exercise 6.37. Let $n \in \mathbb{N}^+ \setminus \{1\}$. Modify the proof in Figure 6.2 to show that if p is irreducible, then $\sqrt[n]{p}$ is irrational.

Exercise 6.38. Let $n \in \mathbb{N}^+ \setminus \{1\}$. Modify the proof in Figure 6.2 to show that if there exists an irreducible integer p such that $p \mid n$ but $p^2 \nmid n$, then $\sqrt{n}$ is irrational.

6.4: Multiplicative clockwork groups

Throughout this section, $n \in \mathbb{N}^+ \setminus \{1\}$, unless otherwise stated.

Multiplication in $\mathbb{Z}_n$

Recall that $\mathbb{Z}_n$ is an additive group, but not multiplicative. In this section we find a subset of $\mathbb{Z}_n$ that we can turn into a multiplicative group, where multiplication is "intuitive":

$$[2]_5 \cdot [3]_5 = [2 \cdot 3]_5 = [6]_5 = [1]_5 .$$

Remember, though: cosets can have various representations, and different representations may lead to different results. We have to ask ourselves, is this operation well-defined?

Lemma 6.39. The proposed multiplication of elements of $\mathbb{Z}_n$ as

$$[a]\,[b] = [ab]$$

is well-defined.

This lemma requires no special constraints on n, so it applies even if $n \in \mathbb{Z}$ is arbitrary.

Proof. Let $x, y \in \mathbb{Z}_n$. Choose $a, b, c, d \in \mathbb{Z}$ such that $x = [a] = [c]$ and $y = [b] = [d]$. By definition of the operation,

$$xy = [a]\,[b] = [ab] \quad \text{and} \quad xy = [c]\,[d] = [cd] .$$

We need to show that $[ab] = [cd]$. The best tool for this is Lemma 3.86 on page 119, which tells us that if we can show that $ab - cd \in n\mathbb{Z}$, then we're done.

How can we accomplish this? By assumption, $[a] = [c]$; this notation means that $a + n\mathbb{Z} = c + n\mathbb{Z}$. Lemma 3.86 tells us that $a - c \in n\mathbb{Z}$. By definition, $a - c = nt$ for some $t \in \mathbb{Z}$. Similarly, $b - d = nu$ for some $u \in \mathbb{Z}$. We can build ab using these differences by multiplying $b\,(a-c)$, but this actually equals $ac - bc$. We can cancel bc using these differences by adding $c\,(b-d)$, and that will give us precisely what we need:

$$\begin{aligned} ab - cd &= b\,(a-c) + c\,(b-d) \\ &= b\,(nt) + c\,(nu) \\ &= n\,(bt + cu), \end{aligned}$$

so $ab - cb \in n\mathbb{Z}$. Lemma 3.86 again tells us that $[ab] = [cb]$ as desired, so the proposed multiplication of elements in $\mathbb{Z}_n$ is well-defined. □

Example 6.40. Recall that $\mathbb{Z}_5 = \mathbb{Z}/\langle 5 \rangle$. The elements of $\mathbb{Z}_5$ are cosets; since $\mathbb{Z}$ is an additive group, we were able to define easily an addition on $\mathbb{Z}_5$ that turns it into an additive group in its own right.

Can we also turn it into a multiplicative group? We need to identify an identity, and inverses. Certainly $[0]$ won't have a multiplicative inverse, but what about $\mathbb{Z}_5 \setminus \{[0]\}$? This generates a multiplication table that satisfies the properties of an abelian (but non-additive) group:

×	1	2	3	4
1	1	2	3	4
2	2	4	1	3
3	3	1	4	2
4	4	3	2	1

That is a group! We'll call it $\mathbb{Z}_5^*$.

In fact, $\mathbb{Z}_5^* \cong \mathbb{Z}_4$; they are both the cyclic group of four elements. In $\mathbb{Z}_5^*$, however, the nominal operation is multiplication, whereas in $\mathbb{Z}_4$ the nominal operation is addition.

You might think that this trick of dropping zero and building a multiplication table always works, *but it doesn't*.

Example 6.41. Recall that $\mathbb{Z}_4 = \mathbb{Z}/\langle 4\rangle = \{[0],[1],[2],[3]\}$. Consider the set $\mathbb{Z}_4\backslash\{[0]\} = \{[1],[2],[3]\}$. The multiplication table for this set *is not closed* because

$$[2]\cdot[2] = [4] = [0] \notin \mathbb{Z}_4\backslash\{[0]\}.$$

If you are tempted to think that we made a mistake by excluding zero, think twice: zero has no inverse. So, we must exclude zero; our mistake seems to have been that we must also exclude 2. This finally works out:

×	1	3
1	1	3
3	3	1

That is a group! We'll call it $\mathbb{Z}_4^*$.

In fact, $\mathbb{Z}_4^* \cong \mathbb{Z}_2$; they are both the cyclic group of two elements. In $\mathbb{Z}_4^*$, however, the operation is multiplication, whereas in $\mathbb{Z}_2$, the operation is addition.

You can determine for yourself that $\mathbb{Z}_2\backslash\{[0]\} = \{[1]\}$ and $\mathbb{Z}_3\backslash\{[0]\} = \{[1],[2]\}$ are also multiplicative groups. In this case, as in $\mathbb{Z}_5^*$, we need remove only 0. For $\mathbb{Z}_6$, however, we have to remove nearly all the elements! We only get a group from $\mathbb{Z}_6\backslash\{[0],[2],[3],[4]\} = \{[1],[5]\}$.

Zero divisors

Why do we need to remove more elements of $\mathbb{Z}_n$ for some values of n than others? Aside from zero, which clearly has no inverse under the operation specified, the elements we've had to remove are those whose multiplication would re-introduce zero.

That's strange: didn't we once learn that the product of two nonzero numbers is nonzero? Yet here we have non-zero elements whose product is zero! True, but this is a different set than the one where you learned the zero product property. Here is an instance where $\mathbb{Z}_n$ superficially behaves *very differently* from the integers. This phenomenon is so important that it has a special name.

Definition 6.42. We say that nonzero elements $x, y \in \mathbb{Z}_n$ are **zero divisors** if $xy = [0]$.

In other words, zero divisors are non-zero elements of $\mathbb{Z}_n$ that violate the zero product property.

Can we find a criterion to detect this?

Lemma 6.43. Let $x \in \mathbb{Z}_n$ be nonzero. The following are equivalent:

(A) x is a zero divisor.

(B) For any representation $[a]$ of x, a and n have a common divisor besides ± 1.

Proof. That (B) implies (A): Let $[a]$ be any representation of x, and assume that a and n share a common divisor $d \neq 1$. Use the definition of divisibility to choose $t, q \in \mathbb{Z}\backslash\{0\}$ such that

$n = qd$ and $a = td$. Let $y = [q]$. Substitution and Lemma 6.39 imply that

$$xy = [a][q] = [aq] = [(td)q] = t[qd] = t[n] = [0].$$

Since $d \neq 1$, $-n < q < n$, so $[0] \neq [q] = y$. By definition, x is a zero divisor.

That (A) implies (B): Assume that x is a zero divisor. By definition, we can find nonzero $y \in \mathbb{Z}_n$ such that $xy = [0]$. Choose $a, b \in \mathbb{Z}$ such that $x = [a]$ and $y = [b]$. Since $xy = [0]$, Lemma 3.86 implies that $n \mid (ab - 0)$, so we can find $k \in \mathbb{Z}$ such that $ab = kn$. Let p_0 be any irreducible number that divides n. Then p_0 also divides kn. Since $kn = ab$, we see that $p_0 \mid ab$. Since p_0 is irreducible, hence prime, it must divide one of a or b. If it divides a, then a and n have a common divisor p_0 that is not ± 1, and we are done; otherwise, it divides b. Use the definition of divisibility to find $n_1, b_1 \in \mathbb{Z}$ such that $n = n_1 p_0$ and $a = b_1 p_0$; it follows that $ab_1 = kn_1$. Again, let p_2 be any irreducible number that divides n_2; the same logic implies that p_2 divides ab_2; being prime, p_2 must divide a or b_2.

As long as we can find prime divisors of the n_i that divide b_i but not a, we repeat this process to find triplets $(n_2, b_2, p_2), (n_3, b_3, p_3), \ldots$ satisfying for all i the properties

- $ab_i = kn_i$;
- $b_{i-1} = p_i b_i$ and $n_{i-1} = p_i n_i$; and so, by Exercise 0.44,
- $|n_{i-1}| > |n_i|$.

The sequence $|n|, |n_1|, |n_2|, \ldots$ is a decreasing sequence of elements of $\mathbb{N}$; by Exercise (0.31), it is finite, and so has a least element, call it $|n_r|$. Observe that

$$b = p_1 b_1 = p_1(p_2 b_2) = \cdots = p_1(p_2(\cdots(p_r b_r))) \tag{25}$$

and

$$n = p_1 n_1 = p_1(p_2 n_2) = \cdots = p_1(p_2(\cdots(p_r n_r))).$$

Case 1. If $n_r = \pm 1$, then $n = p_1 p_2 \cdots p_r$. By substitution into equation 25, $b = nb_r$. By the definition of divisibility, $n \mid b$. By the definition of $\mathbb{Z}_n$, $y = [b] = [0]$. This contradicts the hypothesis.

Case 2. If $n_r \neq \{\pm 1\}$, then Theorem 6.29 tells us that n_r has an irreducible divisor p_{r+1}. Since $p_{r+1} \mid kn_r$, it must also divide ab_r. If $p_{r+1} \mid b_r$, then we can construct n_{r+1} and b_{r+1} that satisfy the properties above for $i = r + 1$. As before, $|n_{r+1}| < |n_r|$, which contradicts the choice of n_r. Hence $p_{r+1} \nmid b_r$; since irreducible integers are prime, $p_{r+1} \mid a$.

Hence n and a share a common divisor that is not ± 1. □

Meet $\mathbb{Z}_n^*$

We can now make a *multiplicative* group out of the set of elements of $\mathbb{Z}_n$ that do not violate the zero product rule.

Definition 6.44. Define the set $\mathbb{Z}_n^*$ to be the set of nonzero elements of $\mathbb{Z}_n$ that are not zero divisors. In set builder notation,

$$\mathbb{Z}_n^* := \{X \in \mathbb{Z}_n \setminus \{0\} : \forall Y \in \mathbb{Z}_n \setminus \{0\}\ XY \neq 0\}.$$

By Lemma 6.43, we could also say that $\mathbb{Z}_n^*$ is the set of positive numbers less than n whose only common factors with n are ± 1. This is the usual definition of $\mathbb{Z}_n^*$ in number theory.

We claim that $\mathbb{Z}_n^*$ is a group under multiplication. Keep in mind that, while it is a sub*set* of $\mathbb{Z}_n$, it is not a sub*group*, as the operations are different.

Theorem 6.45. $\mathbb{Z}_n^*$ is an abelian group under its multiplication.

Proof. We showed in Lemma 6.39 that the operation is well-defined. We check each requirement of a group, slightly out of order. Let $X, Y, Z \in \mathbb{Z}_n^*$, and choose $a, b, c \in \mathbb{Z}$ such that $X = [a]$, $Y = [b]$, and $Z = [c]$.

(associative) By substitution and properties of $\mathbb{Z}_n^*$, $\mathbb{Z}_n$, and $\mathbb{Z}$,

$$X(YZ) = [a][bc] = [a(bc)] = [(ab)c] = [ab][c] = (XY)Z.$$

Notice that this applies for elements of $\mathbb{Z}_n$ as well as elements of $\mathbb{Z}_n^*$.

(closed) Since the operation is well-defined, $XY \in \mathbb{Z}_n$. How do we know that $XY \in \mathbb{Z}_n^*$? Assume to the contrary that it is not. That would mean that $XY = [0]$ or XY is a zero divisor; either way, $\gcd(ab, n) \neq 1$. By definition of $\mathbb{Z}_n^*$, neither X nor Y is a zero divisor, so $XY \neq [0]$, which forces us to conclude that XY is a zero divisor. By definition of zero divisor, there must be some $Z \in \mathbb{Z}_n$ such that $(XY)Z = [0]$. By the associative property, $X(YZ) = [0]$; that is, X is a zero divisor. This contradicts the choice of X! Thus, XY *cannot* be a zero divisor; the assumption that $XY \notin \mathbb{Z}_n^*$ must have been wrong.

(identity) We claim that $[1]$ is the identity. Since $\gcd(1, n) = 1$, Lemma 6.43 tells us that $[1] \in \mathbb{Z}_n^*$. By substitution and arithmetic in both $\mathbb{Z}_n^*$ and $\mathbb{Z}_n$,

$$X \cdot [1] = [a \cdot 1] = [a] = X.$$

A similar argument shows that $[1] \cdot X = X$.

(inverse) We need to find an inverse of X. From Lemma 6.43, a and n have no common divisors except ± 1; hence $\gcd(a, n) = 1$. Bezout's Identity tells us that there exist $b, m \in \mathbb{Z}$ such that $ab + mn = 1$. By arithmetic in both $\mathbb{Z}_n^*$ and $\mathbb{Z}$, as well as Lemma 3.86, we deduce that

$$\begin{aligned} ab - 1 &= n(-m) \\ \therefore ab - 1 &\in n\mathbb{Z} \\ \therefore [ab] &= [1] \\ \therefore [a][b] &= [1]. \end{aligned}$$

Let $Y = [b]$; by substitution, the last equation becomes

$$XY = [1].$$

But is $Y \in \mathbb{Z}_n^*$? In fact it is, and the justification is none other than the same Bezout Identity we used above! We had $ab + mn = 1$. I hope you agree that we can't find a positive integer smaller than 1. You will also agree that 1 is the *smallest* positive

integer d for which we can find $w, z \in \mathbb{Z}$ such that $bw + nz = d$, if we can find such $w, z \in \mathbb{Z}$. In fact, we can: the Bezout Identity above provides a solution, where $d = 1$, $w = a$, and $z = m$. Guess what: Exercise 6.14(b) tells us that $\gcd(b, n) = 1$! By definition, then, $Y = [b] \in \mathbb{Z}_n^*$, and X has an inverse in $\mathbb{Z}_n^*$.

(commutative) Use the definition of multiplication in $\mathbb{Z}_n^*$ and the commutative property of integer multiplication to see

$$XY = [ab] = [ba] = YX.$$

□

By removing elements that share non-trivial common divisors with n, we have managed to eliminate those elements that do not satisfy the zero-product rule, and would break closure by trying to re-introduce zero in the multiplication table. We have thereby created a clockwork group for multiplication, $\mathbb{Z}_n^*$.

Example 6.46. Consider $\mathbb{Z}_{10}^*$. To find its elements, collect the elements of $\mathbb{Z}_{10}$ that are not zero divisors. Lemma 6.43 tells us that these are the elements whose representations $[a]$ satisfy $\gcd(a, n) \neq 1$. Thus

$$\mathbb{Z}_{10}^* = \{[1], [3], [7], [9]\}.$$

Theorem 6.45 tells us that $\mathbb{Z}_{10}^*$ is a group. Since it has four elements, it must be isomorphic to either the Klein 4-group, or to $\mathbb{Z}_4$. Which is it? In this case, it's probably easiest to decide the question with a glance at its multiplication table:

×	**1**	**3**	**7**	**9**
1	1	3	7	9
3	3	9	1	7
7	7	1	9	3
9	9	7	3	1

Notice that $3^{-1} \neq 3$. In the Klein 4-group, every element is its own inverse, so $\mathbb{Z}_{10}^*$ cannot be isomorphic to the Klein 4-group. Instead, it must be isomorphic to $\mathbb{Z}_4$.

Exercises.

Exercise 6.47. List the elements of $\mathbb{Z}_7^*$ using their canonical representations, and construct its multiplication table. Use the table to identify the inverse of each element.

Exercise 6.48. List the elements of $\mathbb{Z}_{15}^*$ using their canonical representations, and construct its multiplication table. Use the table to identify the inverse of each element.

6.5: Euler's Theorem

In Section 6.4 we defined the group $\mathbb{Z}_n^*$ for all $n \in \mathbb{N}^+$ where $n > 1$. This group satisfies an important property called *Euler's Theorem,* a result about Euler's φ-function.

Euler's Theorem

Definition 6.49. Euler's φ-function is $\varphi(n) = \left|\mathbb{Z}_n^*\right|$.

In other words, Euler's φ-function counts the number of positive integers smaller than n that share no common factors with it.

Theorem 6.50 (Euler's Theorem)**.** For all $x \in \mathbb{Z}_n^*$, $x^{\varphi(n)} = 1$.

Proofs of Euler's Theorem based only on Number Theory are not very easy. They're not particularly difficult, either; they just aren't easy. See for example the proof on pages 18–19 of [Lau03].

On the other hand, a proof of Euler's Theorem using group theory is short and straightforward.

Proof. Let $x \in \mathbb{Z}_n^*$. By Exercise 3.48, $x^{|\mathbb{Z}_n^*|} = 1$. By substitution, $x^{\varphi(n)} = 1$. □

Corollary 6.51. For all $x \in \mathbb{Z}_n^*$, $x^{-1} = x^{\varphi(n)-1}$.

Proof. You do it! See Exercise 6.60. □

Corollary 6.51 says that we can compute x^{-1} for any $x \in \left|\mathbb{Z}_n^*\right|$ "relatively easily;" all we need to know is $\varphi(n)$.

Computing $\varphi(n)$

The natural followup question is, what is $\varphi(n)$? For an irreducible integer p, this is easy: the only common factors between p and any positive integer less than p are ± 1; there are $p-1$ of these, so $\varphi(p) = p-1$.

For reducible integers, it is not so easy. Checking a few examples, no clear pattern emerges:

n	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$\left\|\mathbb{Z}_n^*\right\|$	1	2	2	4	2	6	4	6	4	10	4	12	6	8

Computing $\varphi(n)$ turns out to be quite hard. It is a major research topic in number theory, and its difficulty makes the RSA algorithm secure (see Section 6.6). One approach, of course, is to factor n and compute all the positive integers that do not share any common factors. For example,

$$28 = 2^2 \cdot 7,$$

so to compute $\varphi(28)$, we could look at all the positive integers smaller than 28 that do not have 2 or 7 as factors. However, this requires us to know first that 2 and 7 are factors of 28, and no one knows a very *efficient* way to do this.

Another way would be to compute $\varphi(m)$ for each factor m of n, then recombine them. But, how? Lemma 6.52 gives us a first step.

Lemma 6.52. Let $a, b, n \in \mathbb{N}^+$. If $n = ab$ and $\gcd(a,b) = 1$, then $\varphi(n) = \varphi(a)\varphi(b)$.

Example 6.53. In the table above, we have $\varphi(15) = 8$. Notice that this satisfies

$$\varphi(15) = \varphi(5 \times 3) = \varphi(5)\varphi(3) = 4 \times 2 = 8.$$

Proof. Assume $n = ab$. Recall from Exercise 2.26 on page 65 that $\mathbb{Z}_a^* \times \mathbb{Z}_b^*$ is a group; the size of this group is $\left|\mathbb{Z}_a^*\right| \times \left|\mathbb{Z}_b^*\right| = \varphi(a)\,\varphi(b)$. We claim that $\mathbb{Z}_n^* \cong \mathbb{Z}_a^* \times \mathbb{Z}_b^*$. If true, this would prove the lemma, since

$$\varphi(n) = \left|\mathbb{Z}_n^*\right| = \left|\mathbb{Z}_a^* \times \mathbb{Z}_b^*\right| = \left|\mathbb{Z}_a^*\right| \times \left|\mathbb{Z}_b^*\right| = \varphi(a)\,\varphi(b).$$

To show that they are indeed isomorphic, let $f : \mathbb{Z}_n^* \to \mathbb{Z}_a^* \times \mathbb{Z}_b^*$ by $f\left([x]_n\right) = \left([x]_a, [x]_b\right)$. First we show that f is a homomorphism: Let $y, z \in \mathbb{Z}_n^*$; then

$$\begin{aligned}
f\left([y]_n\,[z]_n\right) &= f\left([yz]_n\right) && \text{(arithm. in } \mathbb{Z}_n^*\text{)}\\
&= \left([yz]_a, [yz]_b\right) && \text{(def. of } f\text{)}\\
&= \left([y]_a\,[z]_a, [y]_b\,[z]_b\right) && \text{(arithm. in } \mathbb{Z}_a^*, \mathbb{Z}_b^*\text{)}\\
&= \left([y]_a, [y]_b\right)\left([z]_a, [z]_b\right) && \text{(arithm. in } \mathbb{Z}_a^* \times \mathbb{Z}_b^*\text{)}\\
&= f\left([y]_n\right) f\left([z]_n\right). && \text{(def. of } f\text{)}
\end{aligned}$$

It remains to show that f is one-to-one and onto. It is both surprising and delightful that the Chinese Remainder Theorem will do most of the work for us. To show that f is onto, let $\left([y]_a, [z]_b\right) \in \mathbb{Z}_a^* \times \mathbb{Z}_b^*$. We need to find $x \in \mathbb{Z}$ such that $f\left([x]_n\right) = \left([y]_a, [z]_b\right)$. Consider the system of linear congruences

$$\begin{aligned}
[x] &= [y] \text{ in } \mathbb{Z}_a;\\
[x] &= [z] \text{ in } \mathbb{Z}_b.
\end{aligned}$$

The Chinese Remainder Theorem tells us not only that such x exists in $\mathbb{Z}_n$, but that x is unique in $\mathbb{Z}_n$.

We are not quite done; we have shown that a solution $[x]$ exists in $\mathbb{Z}_n$, but what we really need is that $[x] \in \mathbb{Z}_n^*$. To see that $[x] \in \mathbb{Z}_n^*$ indeed, let d be any common divisor of x and n. By way of contradiction, assume $d \neq \pm 1$; by Theorem 6.29, we can find an irreducible divisor r of d; by Exercise 0.46 on page 18, $r \mid n$ and $r \mid x$. Recall that $n = ab$, so $r \mid ab$. Since r is irreducible, hence prime, $r \mid a$ or $r \mid b$. Without loss of generality, we may assume that $r \mid a$. Recall that $[x]_a = [y]_a$; Lemma 3.86 on page 119 tells us that $a \mid (x - y)$. Let $w \in \mathbb{Z}$ such that $wa = x - y$. Rewrite this equation as $x - wa = y$. Recall that $r \mid x$ and $r \mid a$; we can factor r from the left-hand side of $x - wa = y$ to see that $r \mid y$.

What have we done? We showed that if x and n have a common factor besides ± 1, then y and a also have a common, irreducible factor r. The definition of irreducible implies that $r \neq 1$.

Do you see the contradiction? We originally chose $[y] \in \mathbb{Z}_a^*$. By definition, $[y]$ cannot be a zero divisor in $\mathbb{Z}_a$, so by Lemma 6.43, $\gcd(y, a) = 1$. But the definition of greatest common divisor means that

$$\gcd(y, a) \geq r > 1 = \gcd(y, a),$$

a contradiction! Our assumption that $d \neq 1$ must have been false; we conclude that the only common divisors of x and n are ± 1. Hence, $x \in \mathbb{Z}_n^*$. □

Fast exponentiation

Corollary 6.51 gives us an "easy" way to compute the inverse of any $x \in \mathbb{Z}_n^*$. However, it can take a long time to compute $x^{\varphi(n)}$, so let's take a moment to explain how we can compute canonical forms of exponents in this group more quickly. We will take two steps towards a fast exponentiation in $\mathbb{Z}_n^*$.

Lemma 6.54. For any $n \in \mathbb{N}^+$, $[x^a] = [x]^a$ in $\mathbb{Z}_n^*$.

Proof. You do it! See Exercise 6.62 on the next page. □

Example 6.55. In $\mathbb{Z}_{15}^*$ we can determine easily that $\left[4^{20}\right] = [4]^{20} = \left([4]^2\right)^{10} = [16]^{10} = [1]^{10} = [1]$. Notice that this is a *lot* faster than computing $4^{20} = 1099511627776$ and dividing to find the canonical form.

Do you see what we did? The trick is to break the exponent down into "manageable" powers. How exactly can we do that?

Theorem 6.56 (Fast Exponentiation)**.** Let $a \in \mathbb{N}$ and $x \in \mathbb{Z}$. We can compute x^a in the following way:

1. Let b be the largest integer such that $2^b \leq a$.
2. Let $q_0, q_1, \ldots, q_b$ be the bits of the binary representation of a.
3. Let $y = 1$, $z = x$ and $i = 0$.
4. Repeat the following until $i > b$:
   (a) If $q_i \neq 0$ then replace y with the product of y and z.
   (b) Replace z with z^2.
   (c) Replace i with $i + 1$.

This ends with $x^a = y$.

Theorem 6.56 effectively computes the *binary representation* of a and uses this to square x repeatedly, multiplying the result only by those powers that matter for the representation. Its algorithm is especially effective on computers, whose mathematics is based on binary arithmetic. Combining it with Lemma 6.54 gives an added bonus in $\mathbb{Z}_n^*$, which is what we care about most.

Example 6.57. Since $10 = 2^3 + 2^1$, we can compute $\left[4^{10}\right]_7$ following the algorithm of Theorem 6.56:

1. We have $q_3 = 1$, $q_2 = 0$, $q_1 = 1$, $q_0 = 0$.
2. Let $y = 1$, $z = 4$ and $i = 0$.
3. When $i = 0$:
   (a) We do not change y because $q_0 = 0$.
   (b) Put $z = 4^2 = 16 = 2$. (We're in $\mathbb{Z}_7^*$, remember.)
   (c) Put $i = 1$.
4. When $i = 1$:
   (a) Put $y = 1 \cdot 2 = 2$.
   (b) Put $z = 2^2 = 4$.
   (c) Put $i = 2$.
5. When $i = 2$:
   (a) We do not change y because $q_2 = 0$.

(b) Put $z = 4^2 = 16 = 2$.
(c) Put $i = 3$.

6. When $i = 3$:
(a) Put $y = 2 \cdot 2 = 4$.
(b) Put $z = 4^2 = 2$.
(c) Put $i = 4$.

We conclude that $\left[4^{10}\right]_7 = [4]_7$. Hand computation the long way, or a half-decent calculator, will verify this.

Proof of Fast Exponentiation.

Termination: Termination is due to the fact that b is a finite number, and the algorithm assigns to i the values $0, 1, \ldots, b+1$ in succession, stopping when $i > b$.

Correctness: First, the theorem claims that $q_b, \ldots, q_0$ are the bits of the binary representation of x^a, but do we actually know that the binary representation of x^a has $b+1$ bits? By hypothesis, b is the largest integer such that $2^b \leq a$; if we need one more bit, then the definition of binary representation means that $2^{b+1} \leq x^a$, which contradicts the choice of b. Thus, $q_b, \ldots, q_0$ are indeed the bits of the binary representation of x^a. By definition, $q_i \in \{0, 1\}$ for each $i = 0, 1, \ldots, b$. The algorithm multiplies $z = x^{2^i}$ to y only if $q_i \neq 0$, so that the algorithm computes

$$x^{q_b 2^b + q_{b-1} 2^{b-1} + \cdots + q_1 2^1 + q_0 2^0},$$

which is precisely the binary representation of x^a. □

Exercises.

Exercise 6.58. Compute 3^{28} in $\mathbb{Z}$ using fast exponentiation. Show each step.

Exercise 6.59. Compute 24^{28} in $\mathbb{Z}_7^*$ using fast exponentiation. Show each step.

Exercise 6.60. Prove that for all $x \in \mathbb{Z}_n^*$, $x^{\varphi(n)-1} = x^{-1}$.

Exercise 6.61. Prove that for all $x \in \mathbb{N}^+$, if x and n have no common divisors, then $n \mid \left(x^{\varphi(n)} - 1\right)$.

Exercise 6.62. Prove that for any $n \in \mathbb{N}^+$, $[x^a] = [x]^a$ in $\mathbb{Z}_n^*$.

6.6: RSA Encryption

From the viewpoint of practical applications, some of the most important results of group theory and number theory enable security in internet commerce. We described this problem on page 1: when you buy something online, you submit some private information, at least a credit card or bank account number, and usually more. There is no guarantee that, as this information passes through the internet, it will pass only through servers run by disinterested persons. It is quite possible for the information to pass through a computer run by at least one ill-intentioned hacker, and possibly even organized crime. You probably don't want criminals looking at your credit card number.

Given the inherent insecurity of the internet, the solution is to disguise private information so that snoopers cannot understand it. A common method in use today is the RSA encryption algorithm.[17] First we describe the algorithms for encryption and decryption; afterwards we explain the ideas behind each stage, illustrating with an example; finally we prove that it succesfully encrypts and decrypts messages.

Description and example

Theorem 6.63 (RSA algorithm)**.** Let M be a list of positive integers. Let p, q be two irreducible integers such that:

- $\gcd(p,q) = 1$; and
- $(p-1)(q-1) > \max\{m : m \in M\}$.

Let $N = pq$, and let $e \in \mathbb{Z}^*_{\varphi(N)}$, where φ is the Euler phi-function. If we apply the following algorithm to M:

1. Let $e \in \mathbb{Z}^*_{\varphi(N)}$.
2. Let C be a list of positive integers found by computing the canonical representation of $[m^e]_N$ for each $m \in M$.

and subsequently apply the following algorithm to C:

1. Let $d = e^{-1} \in \mathbb{Z}^*_{\varphi(N)}$.
2. Let D be a list of positive integers found by computing the canonical representation of $\left[c^d\right]_N$ for each $c \in C$.

then $D = M$.

Example 6.64. Consider the text message

ALGEBRA RULZ.

We convert the letters to integers in the fashion that you might expect: A=1, B=2, …, Z=26. We also assign 0 to the space. This allows us to encode the message as,

$$M = (1,12,7,5,2,18,1,0,18,21,12,26).$$

Let $p = 5$ and $q = 11$; then $N = 55$. Let $e = 3$. Is $e \in \mathbb{Z}^*_{\varphi(N)}$? We know that

$$\begin{aligned}\gcd(3,\varphi(N)) &= \gcd(3,\varphi(5)\cdot\varphi(11)) = \gcd(3,4\times 10)\\ &= \gcd(3,40) = 1;\end{aligned}$$

Definition 6.44 and Lemma 6.43 show that, yes, $e \in \mathbb{Z}^*_{\varphi(n)}$.

Encrypt by computing m^e for each $m \in M$:

$$\begin{aligned}C &= \left(1^3,12^3,7^3,5^3,2^3,18^3,1^3,0^3,18^3,21^3,12^3,26^3\right)\\ &= (1,23,13,15,8,2,1,0,2,21,23,31).\end{aligned}$$

A snooper who intercepts C and tries to read it as a plain message would have several problems trying to read it. First, it contains 31, a number that does not fall in the range 0 and 26. If he gave that number the symbol _, he would see

[17] RSA stands for Rivest (of MIT), Shamir (of the Weizmann Institute in Israel), and Adleman (of USC).

AWMOHBA BUW_

which is not an obvious encryption of ALGEBRA RULZ.

The inverse of $3 \in \mathbb{Z}^*_{\varphi(N)}$ is $d = 27$. (We could compute this using Corollary 6.51, but it's not hard to see that $3 \times 27 = 81$ and $[81]_{40} = [1]_{40}$.) Decrypt by computing c^d for each $c \in C$:

$$\begin{aligned} D &= \left(1^{27}, 23^{27}, 13^{27}, 15^{27}, 8^{27}, 2^{27}, 1^{27}, 0^{27}, 2^{27}, 21^{27}, 23^{27}, 31^{27}\right) \\ &= (1, 12, 7, 5, 2, 18, 1, 0, 18, 21, 12, 26)\,. \end{aligned}$$

Trying to read this as a plain message, we have

ALGEBRA RULZ.

Doesn't it?

Encrypting messages letter-by-letter is absolutely unacceptable for security. For a stronger approach, letters should be grouped together and converted to integers. For example, the first four letters of the secret message above are

ALGE

and we can convert this to a number using any of several methods; for example

$$\text{ALGE} \quad \rightarrow \quad 1 \times 26^3 + 12 \times 26^2 + 7 \times 26 + 5 = 25{,}785.$$

In order to encrypt this, we would need larger values for p and q. This is too burdensome to compute by hand, so you want a computer to help. We give an example in the exercises.

RSA is an example of a *public-key cryptosystem.* That means that person A broadcasts to the world, "Anyone who wants to send me a secret message can use the RSA algorithm with values $N = \ldots$ and $e = \ldots$." So a snooper knows the method, the modulus, N, and the encryption key, e!

If the snooper knows the method, N, and e, how can RSA be safe? To decrypt, the snooper needs to compute $d = e^{-1} \in \mathbb{Z}^*_{\varphi(N)}$. Corollary 6.51 tells us that computing d is merely a matter of computing $e^{\varphi(N)-1}$, which is easy if you know $\varphi(N)$. The snooper also knows that $N = pq$, where p and q are prime. So, decryption should be a simple matter of factoring $N = pq$ and applying Lemma 6.52 to obtain $\varphi(N) = (p-1)(q-1)$. Right?

Well, yes *and* no. Typical implementations choose *very* large numbers for p and q, many digits long, and there is *no known method* of factoring a large integer "quickly" — *even when you know that it factors as the product of two primes!* To make things worse, there is a careful science to choosing p and q in such a way that makes it hard to determine their values from N and e.

As it is too time-consuming to perform even easy examples by hand, a computer algebra system becomes necessary to work with examples. At the end of this section, after the exercises, we list programs that will help you perform these computations in the Sage and Maple computer algebra systems. The programs are:

- `scramble`, which accepts as input a plaintext message like "ALGEBRA RULZ" and turns it into a list of integers;
- `descramble`, which accepts as input a list of integers and turns it into plaintext;
- `en_de_crypt`, which encrypts or decrypts a message, depending on whether you feed it the encryption or decryption exponent.

Examples of usage:

- in Sage:
  - to determine the list of integers M, type `M = scramble("ALGEBRA RULZ")`
  - to encrypt M, type
    ```
    C = en_de_crypt(M,3,55)
    ```
  - to decrypt C, type
    ```
    en_de_crypt(C,27,55)
    ```
- in Maple:
  - to determine the list of integers M, type `M := scramble("ALGEBRA RULZ");`
  - to encrypt M, type
    ```
    C := en_de_crypt(M,3,55);
    ```
  - to decrypt C, type
    ```
    en_de_crypt(C,27,55);
    ```

Now, *why* does the RSA algorithm work?

Theory

Before reading the proof, let's reexamine the theorem.

Theorem (RSA algorithm)**.** Let M be a list of positive integers. Let p, q be two irreducible integers such that:

- $\gcd(p,q) = 1$; and
  - $(p-1)(q-1) > \max\{m : m \in M\}$.

Theorem. Let $N = pq$, and let $e \in \mathbb{Z}^*_{\varphi(N)}$, where φ is the Euler phi-function. If we apply the following algorithm to M:

1. Let $e \in \mathbb{Z}^*_{\varphi(N)}$.
   (a) Let C be a list of positive integers found by computing the canonical representation of $[m^e]_N$ for each $m \in M$.

Theorem. and subsequently apply the following algorithm to C:

1. Let $d = e^{-1} \in \mathbb{Z}^*_{\varphi(N)}$.
   (a) Let D be a list of positive integers found by computing the canonical representation of $\left[c^d\right]_N$ for each $c \in C$.

Theorem. then $D = M$.

Proof of the RSA algorithm. Let $i \in \{1,2,\ldots,|C|\}$. Let $c \in C$. By definition of C, $c = m^e \in \mathbb{Z}^*_N$ for some $m \in M$. We need to show that $c^d = (m^e)^d = m$.

Since $[e] \in \mathbb{Z}^*_{\varphi(N)}$, which is a group under multiplication, we know that it has an inverse element, $[d]$. That is, $[de] = [d][e] = [1]$. By Lemma 3.86, $\varphi(N) \mid (1-de)$, so we can find $b \in \mathbb{Z}$ such that $b \cdot \varphi(N) = 1 - de$, or $de = 1 - b\varphi(N)$.

We claim that $[m]^{de} = [m] \in \mathbb{Z}_N$. To do this, we will show two subclaims about the behavior of the exponentiation in $\mathbb{Z}_p$ and $\mathbb{Z}_q$.

Claim 1. $[m]^{de} = [m] \in \mathbb{Z}_p$.

If $p \mid m$, then $[m] = [0] \in \mathbb{Z}_p$. Without loss of generality, $d, e \in \mathbb{N}^+$, so

$$[m]^{de} = [0]^{de} = [0] = [m] \in \mathbb{Z}_p.$$

Otherwise, $p \nmid m$. Recall that p is irreducible, so $\gcd(m, p) = 1$. By Euler's Theorem,

$$[m]^{\varphi(p)} = [1] \in \mathbb{Z}_p^*.$$

Recall that $\varphi(N) = \varphi(p)\varphi(q)$; thus,

$$[m]^{\varphi(N)} = [m]^{\varphi(p)\varphi(q)} = \left([m]^{\varphi(p)}\right)^{\varphi(q)} = [1].$$

Thus, in $\mathbb{Z}_p^*$,

$$\begin{aligned}[m]^{de} &= [m]^{1-b\varphi(N)} = [m] \cdot [m]^{-b\varphi(N)} \\ &= [m]\left([m]^{\varphi(N)}\right)^{-b} = [m] \cdot [1]^{-b} = [m].\end{aligned}$$

As p is irreducible, Any element of $\mathbb{Z}_p$ is either zero or in $\mathbb{Z}_p^*$. We have considered both cases; hence,

$$[m]^{de} = [m] \in \mathbb{Z}_p.$$

Claim 2. $[m]^{1-b\varphi(N)} = [m] \in \mathbb{Z}_q$.

The argument is similar to that of the first claim.

Since $[m]^{de} = [m]$ in both $\mathbb{Z}_p$ and $\mathbb{Z}_q$, properties of the quotient groups $\mathbb{Z}_p$ and $\mathbb{Z}_q$ tell us that $\left[m^{de} - m\right] = [0]$ in both $\mathbb{Z}_p$ and $\mathbb{Z}_q$ as well. In other words, both p and q divide $m^{de} - m$. You will show in Exercise 6.67 that this implies that N divides $m^{de} - m$.

From the fact that N divides $m^{de} - m$, we have $[m]_N^{ed} = [m]_N$. Thus, computing $(m^e)^d$ in $\mathbb{Z}_{\varphi(N)}$ gives us m. □

Exercises.

Exercise 6.65. The phrase

$$[574, 1, 144, 1060, 1490, 0, 32, 1001, 574, 243, 533]$$

is the encryption of a message using the RSA algorithm with the numbers $N = 1535$ and $e = 5$. You will decrypt this message.

(a) Factor N.
(b) Compute $\varphi(N)$.
(c) Find the appropriate decryption exponent.
(d) Decrypt the message.

Exercise 6.66. In this exercise, we encrypt a phrase using more than one letter in a number.

(a) Rewrite the phrase GOLDEN EAGLES as a list M of three positive integers, each of which combines four consecutive letters of the phrase.
(b) Find two prime numbers whose product is larger than the largest number you would get from four letters.
(c) Use those two prime numbers to compute an appropriate N and e to encrypt M using RSA.
(d) Find an appropriate d that will decrypt M using RSA.
(e) Decrypt the message to verify that you did this correctly.

Exercise 6.67. Let $m, p, q \in \mathbb{Z}$ and suppose that $\gcd(p, q) = 1$.
(a) Show that if $p \mid m$ and $q \mid m$, then $pq \mid m$.
(b) Explain why this completes the proof of the RSA algorithm; that is, since p and q both divide $m^{de} - m$, then so does N.

Sage programs

The following programs can be used in Sage to help make the amount of computation involved in the exercises less burdensome:

```
def scramble(s):
  result = []
  for each in s:
    if ord(each) >= ord("A") \
        and ord(each) <= ord("Z"):
      result.append(ord(each)-ord("A")+1)
    else:
      result.append(0)
  return result


def descramble(M):
  result = ""
  for each in M:
    if each == 0:
      result = result + " "
    else:
      result = result + chr(each+ord("A") - 1)
  return result


def en_de_crypt(M,p,N):
  result = []
  for each in M:
    result.append((each^p).mod(N))
  return result
```

Maple programs

The following programs can be used in Maple to help make the amount of computation involved in the exercises less burdensome:

```
scramble := proc(s)
  local result, each, ord;
  ord := StringTools[Ord];
  result := [];
  for each in s do
    if ord(each) >= ord("A")
        and ord(each) <= ord("Z") then
      result := [op(result),
        ord(each) - ord("A") + 1];
    else
      result := [op(result), 0];
    end if;
  end do;
  return result;
end proc:


descramble := proc(M)
  local result, each, char, ord;
  char := StringTools[Char];
  ord := StringTools[Ord];
  result := "";
  for each in M do
    if each = 0 then
      result := cat(result, " ");
    else
      result := cat(result,
        char(each + ord("A") - 1));
    end if;
  end do;
  return result;
end proc:


en_de_crypt := proc(M,p,N)
  local result, each;
  result := [];
  for each in M do
    result := [op(result), (each^p) mod N];
  end do;
  return result;
end proc:
```

Part II

Rings

Chapter 7:
Rings

While monoids are defined by one operation, groups are arguably defined by two: addition and subtraction, for example, or multiplication and division. The second operation is so closely tied to the first that we consider groups to have only one operation, for which (unlike monoids) every element has an inverse.

Of course, a set can be closed under more than one operation; for example, $\mathbb{Z}$ is closed under both addition and multiplication. As with subtraction, it is possible to define the multiplication of integers in terms of addition, just as we did with groups. However, this is not possible for all sets where an addition and a multiplication are both defined. Think of the multiplication of polynomials; how would you define $(x+1)(x-1)$ as repeated addition of $x-1$, a total of $x+1$ times? Does that even make sense? This motivates the study of a structure that incorporates common properties of two operations, which are related as loosely as possible.

Section 7.1 of this chapter introduces us to this structure, called a *ring*. A ring has two operations, "addition" and "multiplication". As you should expect from your experience with groups, what we call "addition" and "multiplication" may look nothing at all like the usual addition and multiplication of numbers. In fact, while the multiplication of integers has a natural definition from addition, multiplication in a ring may have absolutely nothing to do with addition, with one exception: the distributive property must still hold.

The rest of the chapter examines special kinds of rings. In Section 7.2 we introduce special kinds of rings that model useful properties of $\mathbb{Z}$ and $\mathbb{Q}$. In Section 7.3 we introduce rings of polynomials. The Euclidean algorithm, which proved so important in chapter 6, serves as the model for a special kind of ring described in Section 7.4.

A concept related to monoids is useful for definitions related to rings.

Definition 7.1. Let S be a set, and $\circ$ an operation. We say that $(S, \circ)$ is a **semigroup** if its operation is closed and associative, although it might not have an identity element.

Notice that

- a monoid is a semigroup,
- a semigroup is almost a monoid, but lacks an identity, and
- the "absorbing subsets" of Section 1.4 are "subsemigroups" of monoids.

A "semigroup" is "half a group", in that it satisfies half of the properties of a group. We will take this up further in Chapter 8.

7.1: A structure for addition and multiplication

What sort of properties do we associate with both addition and multiplication? We typically associate the properties of addition with an abelian group, and the properties of multiplication with a monoid, although it really depends on the set. The most basic properties of multiplication are encapsulated by the notion of a semigroup, so we'll start from there, and add more as needed.

Definition 7.2. Let R be a set *with at least one element*, and $+$ and $\times$ two binary operations on that set. We say that $(R,+,\times)$ is a **ring** if it satisfies the following properties:

(R1) $(R,+)$ is an abelian group.

(R2) $(R,\times)$ is a semigroup.

(R4) R satisfies the distributive property of addition over multiplication: that is,
for all $a,b,c\in R$, $a(b+c)=ab+ac$ and $(a+b)c=ac+bc$.

Notation 7.3. As with groups, we usually refer simply to the ring as R, rather than $(R,+,\times)$. Since $(R,+)$ is an abelian group, the ring has an additive identity, 0. We sometimes write 0_R to emphasize that it is the additive identity of R.

Notice the following:

- While addition is commutative on account of (R1), multiplication need not be.
- There is no requirement that a multiplicative identity exists.
- There is no requirement that multiplicative inverses exist.
- There is no guarantee (yet) that the additive identity interacts with multiplication according to properties you have seen before. In particular, there is *no guarantee* that
  - the zero-product rule holds; or even that
  - $0_R\cdot a=0_R$ for any $a\in R$.

Example 7.4. Let $R=\mathbb{R}^{m\times m}$ for some positive integer m. It turns out that R is a ring under the usual addition and multiplication of matrices. After all, Example 1.8 shows that the matrices satisfy the properties of a monoid under multiplication, and Example 2.4 shows that they are a group under addition, though most of the work was done in Section 0.3. The only part missing is distribution, and while that isn't hard, it is somewhat tedious, so we defer to your background in linear algebra.

However, we do want to point out something that should make you at least a *little* uncomfortable. Let

$$A=\begin{pmatrix}1&0\\0&0\end{pmatrix}\quad\text{and}\quad B=\begin{pmatrix}0&0\\0&1\end{pmatrix}.$$

Routine computation shows that

$$AB=\begin{pmatrix}0&0\\0&0\end{pmatrix},$$

or in other words, $AB=0$. This is true even though $A,B\neq 0$! Hence

Not every ring R satisfies the **zero product property**

$$\forall a,b\in R\qquad ab=0\quad\Longrightarrow\quad a=0\text{ or }b=0.$$

Example 7.4 shouldn't surprise you that much; first, you've seen it in linear algebra, and second, you met zero divisors in Section 6.4. In fact, we will shortly generalize that idea into zero divisors for rings.

Likewise, the sets $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, with which you are long familiar, are also rings. We omit the details, but you should think about them a little bit, and ask your instructor if some part of it isn't clear. You will study other example rings in the exercises. For now, we prove a familiar property of the additive identity.

Proposition 7.5. For all $r \in R$,

$$r \cdot 0_R = 0_R \cdot r = 0_R.$$

If you see that and ask, "Isn't that obvious?" then you *really* need to read the proof. While you read it, ask yourself, "What properties of a ring make this statement true?" The answer to that question will indicate your hidden assumptions. Try to prove the proposition without those properties, and you will see why it is *not* in fact obvious.

Proof. Let $r \in R$. Since $(R, +)$ is an abelian group, we know that $0_R + 0_R = 0_R$. By substitution, $r(0_R + 0_R) = r \cdot 0_R$. By distribution, $r \cdot 0_R + r \cdot 0_R = r \cdot 0_R$. Since $(R, +)$ is an abelian group, $r \cdot 0_R$ has an additive inverse; call it s. Applying the properties of a ring, we have

$$\begin{aligned} s + (r \cdot 0_R + r \cdot 0_R) &= s + r \cdot 0_R && \text{(substitution)} \\ (s + r \cdot 0_R) + r \cdot 0_R &= s + r \cdot 0_R && \text{(associative)} \\ 0_R + r \cdot 0_R &= 0_R && \text{(additive inverse)} \\ r \cdot 0_R &= 0_R. && \text{(additive identity)} \end{aligned}$$

A similar argument shows that $0_R \cdot r = 0_R$. □

We now turn our attention to two properties that, while pleasant, are not necessary for a ring.

Definition 7.6. Let R be a ring. If R has a multiplicative identity 1_R such that

$$r \cdot 1_R = 1_R \cdot r = r \qquad \forall r \in R,$$

we say that R is a **ring with unity**. (Another name for the multiplicative identity is **unity**.)

If R is a ring and the multiplicative operation is commutative, so that

$$rs = sr \quad \forall r \in R,$$

then we say that R is a **commutative ring**.

A ring with unity is

- an abelian group under multiplication, and
- a (possibly commutative) monoid under addition.

Example 7.7. The set of matrices $\mathbb{R}^{m \times m}$ is a ring with unity, where I_m is the multiplicative identity. However, it is not a commutative ring.

You will show in Exercise 7.13 that $2\mathbb{Z}$ is a ring. It is a commutative ring, but not a ring with unity.

For a commutative ring with unity, consider $\mathbb{Z}$.

Remark 7.8. While non-commutative rings are interesting,

Unless we state otherwise,
all rings in these notes are commutative.

As with groups, we can characterize all rings with only two elements.

Example 7.9. Let R be a ring with only two elements. There are two possible structures for R.

Why? Since $(R,+)$ is an abelian group, by Example 2.9 on page 60 the addition table of R has the form

$+$	0_R	a
0_R	0_R	a
a	a	0_R

.

By Proposition 7.5, we know that the multiplication table *must* have the form

$\times$	0_R	a
0_R	0_R	0_R
a	0_R	?

where $a \cdot a$ is undetermined. Nothing in the properties of a ring tell us whether $a \cdot a = 0_R$ or $a \cdot a = a$; in fact, rings exist with both properties:

- if $R = \mathbb{Z}_2$ (see Exercise 7.14 to see that this is a ring), then $a = [1]$ and $a \cdot a = a$; but
- if

$$R = \left\{ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, a = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \right\} \subsetneq (\mathbb{Z}_2)^{2\times 2},$$

  then $a \cdot a = 0 \neq a$.

Just as groups have subgroups, rings have subrings:

Definition 7.10. Let R be a ring, and S a nonempty subset of R. If S is also a ring under the same operations as R, then S is a subring of R.

Example 7.11. Recall from Exercise 7.13 that $2\mathbb{Z}$ is a ring; since $2\mathbb{Z} \subsetneq \mathbb{Z}$, it is a subring of $\mathbb{Z}$.

To show that a subset of a ring is a subring, do we have to show all four ring properties? No: as with subgroups, we can simplify the characterization to two properties:

Theorem 7.12 (The Subring Theorem)**.** Let R be a ring and S be a nonempty subset of R. The following are equivalent:

(A) S is a subring of R.

(B) S is closed under subtraction and multiplication: for all $a, b \in S$
  (S1) $a - b \in S$, and
  (S2) $ab \in S$.

Proof. That (A) implies (B) is clear, so assume (B). From (B) we know that for any $a, b \in S$ we have (S1) and (S2). As (S1) is essentially the Subgroup Theorem, S is an additive subgroup of the additive group R. On the other hand, (S2) only tells us that S satisfies property $(R2)$ of a ring, but any elements of S are elements of R, so the associative and distributive properties follow from inheritance. Thus S is a ring in its own right, which makes it a subring of R. □

Exercises

Exercise 7.13.

(a) Show that $2\mathbb{Z}$ is a ring under the usual addition and multiplication of integers.

(b) Show that for any $n \in \mathbb{Z}$, $n\mathbb{Z}$ is a ring under the usual addition and multiplication of integers.

Exercise 7.14. Recall the definition of multiplication for $\mathbb{Z}_n$ from Section 6.4: for $[a], [b] \in \mathbb{Z}_n$, $[a][b] = [ab]$.

(a) Show that $\mathbb{Z}_2$ is a ring under the addition and multiplication of cosets defined in Section 3.5.

(b) Show that for any $n \in \mathbb{N}^+$ where $n > 1$, $\mathbb{Z}_n$ is a ring under the addition and multiplication of cosets defined in Section 3.5.

(c) Show that there exist a, b, n such that $[a]_n [b_n] = [0]_n$ but $[a]_n, [b]_n \neq [0]_n$.

Exercise 7.15. Let R be a ring.

(a) Show that for all $r, s \in R$, $(-r)s = r(-s) = -(rs)$.

(b) Suppose that R has unity. Show that $-r = -1_R \cdot r$ for all $r \in R$.

Exercise 7.16. Let R be a ring with unity. Show that $1_R = 0_R$ if and only if R has only one element.

Exercise 7.17. Consider the two possible ring structures from Example 7.9. Show that if a ring R has only two elements, one of which is unity, then it can have only one of the structures.

Exercise 7.18. Let $R = \{T, F\}$ with the additive operation $\oplus$ (Boolean xor) and a multiplicative operation $\wedge$ (Boolean and where

$$\begin{aligned} F \oplus F &= F & F \wedge F &= F \\ F \oplus T &= T & F \wedge T &= F \\ T \oplus F &= T & T \wedge F &= F \\ T \oplus T &= F & T \wedge T &= T. \end{aligned}$$

(See also Exercises 2.20 and 2.21 on page 64.) Is $(R, \oplus, \wedge)$ a ring? If it is a ring, then

(a) what is the zero element?

(b) does it have a unity element? if so, what is it?

(c) is it commutative?

Exercise 7.19. Let R and S be rings, with $R \subseteq S$ and $\alpha \in S$. The **extension of R by α** is

$$R[\alpha] = \{r_n \alpha^n + \cdots + r_1 \alpha + r_0 : n \in \mathbb{N},\ r_0, r_1, \ldots, r_n \in R\}.$$

(a) Show that $R[\alpha]$ is also a ring.

(b) Suppose $R = \mathbb{Z}$, $S = \mathbb{C}$, and $\alpha = \sqrt{-5}$.

(i) Explain why every element of $R[\alpha]$ can be written in the form $a + b\alpha$.

(ii) Show that 6 can be factored two distinct ways in $R[\alpha]$: one is the ordinary factorization in $R = \mathbb{Z}$, while the other exploits the difference of squares with $\alpha = \sqrt{-5}$.

Exercise 7.20. In Exercise 7.14, you showed that $\mathbb{Z}_n$ is a ring. A nonzero element r of a ring R is **nilpotent** if we can find $n \in \mathbb{N}^+$ such that $r^n = 0_R$.

(a) Identify the nilpotent elements, *if any*, of $\mathbb{Z}_n$ for $n = 2, 3, 4, 5, 6$. If not, state that.

(b) Do you think there is a relationship between n and the nilpotents of $\mathbb{Z}_n$? If so, state it.

7.2: Integral Domains and Fields

In this section, R is always a commutative ring with unity.

Example 7.4 illustrates an important point: not all rings satisfy properties that we might like to take for granted. Not only does it show that not all rings possess the zero product property, it also demonstrates that multiplicative inverses do not necessarily exist in all rings. Both multiplicative inverses and the zero product property are very useful; we use them routinely to solve equations! Rings with these properties deserve special attention.

Two convenient kinds of rings

We first classify rings that satisfy the zero product property.

Definition 7.21. If the elements of R satisfy the zero product property, then we call R an **integral domain**.

We use the word "integral" here because R is like the ring of "integ"ers, $\mathbb{Z}$. We do *not* mean that you can compute the integrals of calculus.

Whenever R is not an integral domain, we can find two elements of R that *do not* satisfy the zero product property; that is, we can find nonzero $a, b \in R$ such that $ab = 0_R$. Recall that we used a special term for this phenomenon in the group $\mathbb{Z}_n^*$, **zero divisors** (Section 6.4). The ideas are identical, so the term is appropriate, and we will call a and b **zero divisors** in a ring, as well.

Example 7.22. As you might expect, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are integral domains.

In Exercise 7.14, you showed that $\mathbb{Z}_n$ was a ring under clockwork addition and multiplication. However, it need not be an integral domain. For example, in $\mathbb{Z}_6$ we have $[2] \cdot [3] = [6] = [0]$, making $[2]$ and $[3]$ zero divisors. On the other hand, it isn't hard to see that $\mathbb{Z}_2$, $\mathbb{Z}_3$, and $\mathbb{Z}_5$ are integral domains, if only via an exhaustive check. What about $\mathbb{Z}_4$? We leave that, and all of $\mathbb{Z}_n$ to the exercises.

Next, we turn to multiplicative inverses.

Definition 7.23. If every non-zero element of R has a multiplicative inverse, then we call R a **field**.

Example 7.24. The rings $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are fields, while $\mathbb{Z}$ is not.

What about $\mathbb{Z}_n$ and $\mathbb{Z}_n^*$? Again, we leave those to the exercises. For now, we need to notice an important relationship between fields and integral domains.

The examples show that some integral domains are not fields, but all the fields we've listed are also integral domains. It would be great if this turned out to be true in general: that is, if every field is an integral domain. Determining the relationships between different classes of rings, and remembering which class you're working with, is a crucial point of ring theory.

Theorem 7.25. Every field is an integral domain.

Proof. Let $\mathbb{F}$ be a field. We claim that $\mathbb{F}$ is an integral domain: that is, the elements of $\mathbb{F}$ satisfy the zero product property. Let $a, b \in \mathbb{F}$ and assume that $ab = 0$. We need to show that $a = 0$ or

$b = 0$. If $a = 0$, we're done, so assume that $a \neq 0$. Since $\mathbb{F}$ is a field, a has a multiplicative inverse. Apply Proposition 7.5 to obtain

$$b = 1 \cdot b = \left(a^{-1}a\right) b = a^{-1}(ab) = a^{-1} \cdot 0 = 0.$$

Hence $b = 0$.

We had assumed that $ab = 0$ and $a \neq 0$. By concluding that $b = 0$, the fact that a and b are arbitrary show that $\mathbb{F}$ is an integral domain. Since $\mathbb{F}$ is an arbitrary field, every field is an integral domain. □

Not every integral domain is a field, however. The most straightforward example is $\mathbb{Z}$.

The field of fractions

Speaking of $\mathbb{Q}$, it happens to be the smallest field that contains $\mathbb{Z}$, an integral domain. So there's another interesting question: can we form a field from any ring R, simply by adding fractions?

No, of course not — we just saw that a field must be an integral domain, and some rings are not integral domains. Even if you add fractions, the zero divisors remain, so you cannot have a field. So, then, can we form a field from any integral domain in the same way that we form $\mathbb{Q}$ from $\mathbb{Z}$? We need some precision in this discussion, which requires a definition.

Definition 7.26. Let R be an arbitrary ring. **The set of fractions over** a ring R is

$$\operatorname{Frac}(R) := \left\{ \frac{p}{q} : \ p, q \in R \text{ and } q \neq 0 \right\},$$

with addition and multiplication defined in the usual way for "fractions", and equality defined by

$$\frac{a}{b} = \frac{p}{q} \iff aq = bp.$$

The answer to our question turns out to be yes!

Theorem 7.27. If R is an integral domain, then $\operatorname{Frac}(R)$ is a ring.

To prove Theorem 7.27, we need two useful properties of fractions that you should be able to prove yourself.

Proposition 7.28. Let R be a ring, $a, b, r \in R$. If $br \neq 0$, then in $\operatorname{Frac}(R)$

- $\dfrac{a}{b} = \dfrac{ar}{br}$, and
- $\dfrac{0_R}{a} = \dfrac{0_R}{b}$.

Proof. You do it! See Exercise 7.33. □

Watch for these properties in what follows.

Proof of Theorem 7.27. Assume that R is an integral domain. First we show that $\mathrm{Frac}(R)$ is an additive group. Let $f, g, h \in R$; choose $a, b, p, q, r, s \in \mathrm{Frac}(R)$ such that $f = a/b$, $g = p/q$, and $h = r/s$. First we show that $\mathrm{Frac}(R)$ is an abelian group.

closure: This is fairly routine, using common denominators. Since R is a domain and $b, q \neq 0$, we know that $bq \neq 0$. Thus,

$$\begin{aligned} f + g &= \frac{a}{b} + \frac{p}{q} && \text{(substitution)} \\ &= \frac{aq}{bq} + \frac{bp}{bq} && \text{(Proposition 7.28)} \\ &= \frac{aq + bp}{bq} && \text{(definition of addition in } \mathrm{Frac}(R)\text{)} \\ &\in \mathrm{Frac}(R). \end{aligned}$$

Why did we need R do be an integral domain? If not, then it is possible that $bq = 0$, and if so, $f + g \notin \mathrm{Frac}(R)$!

associative: This is the hardest one; watch for Proposition 7.28 to show up in many places. As before, since R is a domain and $b, q, s \neq 0$, we know that bq, $(bq)s$, $b(qs)$, and qs are all non-zero. Thus,

$$\begin{aligned} (f + g) + h &= \frac{aq + bp}{bq} + \frac{r}{s} \\ &= \frac{(aq + bp)s}{(bq)s} + \frac{(bq)r}{(bq)s} \\ &= \frac{((aq)s + (bp)s) + (bq)r}{(bq)s} \\ &= \frac{a(qs) + (b(ps) + b(qr))}{b(qs)} \\ &= \frac{a(qs)}{b(qs)} + \frac{b(ps) + b(qr)}{b(qs)} \\ &= \frac{a}{b} + \frac{ps + qr}{qs} \\ &= \frac{a}{b} + \left(\frac{p}{q} + \frac{r}{s}\right) \\ &= f + (g + h.) \end{aligned}$$

identity: We claim that the additive identity of $\mathrm{Frac}(R)$ is $0_R/1_R$. This is easy to see, since

$$f + \frac{0_R}{1_R} = \frac{a}{b} + \frac{0_R \cdot b}{1_R \cdot b} = \frac{a}{b} + \frac{0_R}{b} = \frac{a}{b} = f.$$

additive inverse: For each $f = p/q$, we claim that $(-p)/q$ is the additive inverse. This is easy

to see, but a little tedious. It is straightforward enough that,

$$f + \frac{-p}{q} = \frac{p}{q} + \frac{-p}{q} = \frac{(p + (-p))}{q} = \frac{0_R}{q}.$$

Don't conclude too quickly that we are done! We have to show that $f + (-p)/q = 0_{\mathrm{Frac}(R)}$, which is $0_R/1_R$. By Proposition 7.28, $0_R/1_R = 0_R/q_R$, so we did in fact compute the identity.

commutative: Using the fact that R is commutative, we have

$$\begin{aligned} f + g &= \frac{a}{b} + \frac{c}{d} = \frac{ad}{bd} + \frac{bc}{bd} \\ &= \frac{ad + bc}{bd} = \frac{cb + da}{db} \\ &= \frac{cb}{db} + \frac{da}{db} = \frac{c}{d} + \frac{a}{b} \\ &= g + f. \end{aligned}$$

Next we have to show that $\mathrm{Frac}(R)$ satisfies the requirements of a ring.

closure: Using closure in R and the fact that R is an integral domain, this is straightforward: $fg = (ap)/(bq) \in \mathrm{Frac}(R)$.

associative: Using the associative property of R, this is straightforward:

$$\begin{aligned} (fg)h &= \left(\frac{ap}{bq}\right)\frac{r}{s} = \frac{(ap)r}{(bq)s} = \frac{a(pr)}{b(qs)} \\ &= \frac{a}{b}\frac{(pr)}{qs} = f(gh). \end{aligned}$$

distributive: We rely on the distributive property of R:

$$\begin{aligned} f(g+h) &= \frac{a}{b}\left(\frac{p}{q} + \frac{r}{s}\right) = \frac{a}{b}\left(\frac{ps + qr}{qs}\right) \\ &= \frac{a(ps + qr)}{b(qs)} = \frac{a(ps) + a(qr)}{b(qs)} \\ &= \frac{a(ps)}{b(qs)} + \frac{a(qr)}{b(qs)} = \frac{ap}{bq} + \frac{ar}{bs} \\ &= fg + fh. \end{aligned}$$

Finally, we show that $\mathrm{Frac}(R)$ is a field. We have to show that it is commutative, that it has a multiplicative identity, and that every non-zero element has a multiplicative inverse.

commutative: We claim that the multiplication of $\mathrm{Frac}(R)$ is commutative. This follows from the fact that R, as an integral domain, has a commutative multiplication, so

$$\begin{aligned} fg &= \frac{a}{b} \cdot \frac{p}{q} = \frac{ap}{bq} = \frac{pa}{qb} \\ &= \frac{p}{q} \cdot \frac{a}{b} = gf. \end{aligned}$$

multiplicative identity: We claim that $\frac{1_R}{1_R}$ is a multiplicative identity for $\mathrm{Frac}(R)$. In fact,

$$f \cdot \frac{1_R}{1_R} = \frac{a}{b} \cdot \frac{1_R}{1_R} = \frac{a \cdot 1_R}{b \cdot 1_R} = \frac{a}{b} = f.$$

multiplicative inverse: Let $f \in \mathrm{Frac}(R)$ be a non-zero element. Let $a, b \in R$ such that $f = a/b$ and $a \neq 0$. Let $g = b/a$; then

$$fg = \frac{a}{b} \cdot \frac{b}{a} = \frac{ab}{ab}.$$

By Proposition 7.28

$$\frac{ab}{ab} = \frac{1_R}{1_R},$$

which we just showed to be the identity of $\mathrm{Frac}(R)$.

□

Definition 7.29. For any integral domain R, we call $\mathrm{Frac}(R)$ the **field of fractions of** R.

Exercises.

Exercise 7.30. Explain why $n\mathbb{Z}$ is not always an integral domain. For what values of n is it an integral domain?

Exercise 7.31. Show that $\mathbb{Z}_n$ is an integral domain if and only if n is irreducible. Is it also a field in these cases?

Exercise 7.32. You might think from Exercise 7.31 that we can turn $\mathbb{Z}_n$ into a field, or at least an integral domain, in the same way that we turned $\mathbb{Z}_n$ into a multiplicative group: that is, working with $\mathbb{Z}_n^*$. Explain that this doesn't work in general, because $\mathbb{Z}_n^*$ isn't even a ring.

Exercise 7.33. Show that if R is an integral domain, then the set of fractions has the following properties for any nonzero $a, b, c \in R$:

$$\frac{ac}{bc} = \frac{ca}{cb} = \frac{a}{b}, \qquad \frac{0_R}{a} = \frac{0_R}{1} = 0_{\mathrm{Frac}(R)},$$
$$\text{and} \qquad \frac{a}{a} = \frac{1_R}{1_R} = 1_{\mathrm{Frac}(R)}.$$

Exercise 7.34. To see concretely why $\mathrm{Frac}(R)$ is not a field if R is not a domain, consider $R = \mathbb{Z}_4$. Find nonzero $b, q \in R$ such that $bq = 0$, using them to find $f, g \in \mathrm{Frac}(R)$ such that $fg \notin \mathrm{Frac}(R)$.

7.3: Polynomial rings

When the average man on the street thinks of "algebra", he typically thinks not of "monoids", "groups", or "rings", but of "polynomials". Polynomials are certainly the focus of high school algebra, and they are also a major focus of higher algebra. The last few chapters of these notes are dedicated to the classical applications of the structural theory to important problems about polynomials.

While one can talk of a monoid or group of polynomials under addition, it is more natural to talk about a ring of polynomials under addition and multiplication. Polynomials helped motivate the distinction between the "two operations" of groups, which we decided was really two sides of one coin, and the "two operations" of rings, which really can be quite different operations. Polynomials provide great examples for the remaining topics. It is time to give them a good, hard look.

Some of the following may seem pedantic and needlessly detailed, and there's some truth to that, but it is important to fix these terms now to avoid confusion later. The difference between a "monomial" and a "term" is of special note; some authors reverse the notions. Similarly, pay attention to the notion of the support $\mathcal{T}_f$ of a polynomial f.

As usual, R is a ring.

Fundamental notions

Definition 7.35. An **indeterminate over** R is a symbol that represents an unknown value of R. A **constant of** R is a symbol that represents a fixed value of R. An **variable over** R is an indeterminate whose value is *not* fixed.

Notice that a constant can be indeterminate, as in the usual use of letters like a, b, and c, or quite explicitly determined, as in 1_R, 0_R, and so forth. Variables are always indeterminate. The main difference is that a constant is *fixed*, while a variable is not.

Definition 7.36. A **monomial over** R is a finite product of variables over R.

The use of "monomial" here is meant to be both consistent with its definition in Section 1.1, and with our needs for future work. Typically, though, we refer simply to "a monomial" rather than "a monomial over R".

By referring to "variables", the definition of a monomial explicitly excludes constants. Even though a^2 looks like a monomial, if a is a constant, we do not consider it a monomial; from our point of view, it is a constant.

Definition 7.37. The **total degree** of a monomial is the number of factors in the product. We say that two monomials are **like monomials** if they have the same variables, and corresponding variables have the same exponents.

A **term of** R is a constant, *or* the product of a monomial over R and a constant of R. The constant in a term is called the **coefficient** of the term. Two terms are **like terms** if their monomials are like monomials.

Now we define *polynomials*.

Definition 7.38. A **polynomial over** R is a finite sum of terms of R. We can write a generic polynomial f as $f = a_1t_1 + a_2t_2 + \cdots + a_mt_m$ where each $a_i \in R$ and each t_i is a monomial.

We call the set of monomials of f with non-zero coefficient its **support**. If we denote the support of f by $\mathcal{T}_f$, then we can write f as

$$f \quad = \quad \sum_{i=1,\ldots,\#\mathcal{T}_f} a_i t_i \quad = \quad \sum_{t\in\mathcal{T}_f} a_t t.$$

We call R the **ground ring** of each polynomial.

We say that two polynomials f and g are equal if $\mathcal{T}_f = \mathcal{T}_g$ and the coefficients of corresponding monomials are equal.

Notation 7.39. We adopt a convention that $\mathcal{T}_f$ is the support of a polynomial f.

Definition 7.40. $R[x]$ is the set of **univariate** polynomials in the variable x over R. That is, $f \in R[x]$ if and only if there exist $m \in \mathbb{N}$ and $a_m, a_{m-1}, \ldots, a_1 \in R$ such that

$$f(x) = a_m x^m + a_{m-1}x^{m-1} + \cdots + a_1 x + a_0.$$

The set $R[x,y]$ is the set of **bivariate** polynomials in the variables x and y whose coefficients are in R.

For $n \geq 2$, the set $R[x_1, x_2, \ldots, x_n]$ is the set of **multivariate** polynomials in the variables $x_1, x_2, \ldots, x_n$ whose coefficients are in R.

The **degree** of a univariate polynomial f, written $\deg f$, is the largest of the total degrees of the monomials of f. We write $\mathrm{lm}(f)$ for the monomial of f with that degree, and $\mathrm{lc}(f)$ for its coefficient. Unless we say otherwise, the degree of a multivariate polynomial is undefined.

Example 7.41. Definition 7.40 tells us that $\mathbb{Z}_6[x,y]$ is the set of bivariate polynomials in x and y whose coefficients are in $\mathbb{Z}_6$. For example,

$$f(x,y) = 5x^3 + 2x \in \mathbb{Z}_6[x,y]$$

and

$$g(x,y) = x^2y^2 - 2x^3 + 4 \in \mathbb{Z}_6[x,y].$$

The ground ring for both f and g is $\mathbb{Z}_6$. Observe that f can be considered a univariate polynomial, in which case $\deg f = 3$.

We also consider constants to be polynomials of degree 0; thus $4 \in \mathbb{Z}_6[x,y]$ and even $0 \in \mathbb{Z}_6[x,y]$.

It is natural to think of a constant as a polynomial. This leads to some unexpected, but interesting and important consequences.

Definition 7.42. Let $f \in R[x_1,\ldots,x_n]$.

We say that f is a **constant polynomial** if $\mathcal{T}_f = \{1\}$ or $\mathcal{T}_f = \emptyset$; in other words, all the non-constant terms have coefficient zero.

We say that f is a **vanishing polynomial** if for all $r_1,\ldots,r_n \in R$, $f(r_1,\ldots,r_n) = 0$. We will see that this can happen even if $f \neq 0_R$.

The definition of vanishing and constant polynomials implies that 0_R satisfies both. However, the definition of equality means that vanishing polynomials need not be zero polynomials!

Example 7.43. Let $f(x) = x^2 + x \in \mathbb{Z}_2[x]$. Since $\mathcal{T}_f \neq \emptyset$, $f \neq 0_R$. However,

$$\begin{aligned} f(0) &= 0^2 + 0 && \text{and} \\ f(1) &= 1^2 + 1 = 0 && (\text{in } \mathbb{Z}_2!). \end{aligned}$$

Here f is a vanishing polynomial *even though it is not zero.*

Properties of polynomials

We can now turn our attention to the properties of $R[x]$ and $R[x_1,\ldots,x_n]$. First up is a question raised by Example 7.43: when must a vanishing polynomial be the constant polynomial 0?

Proposition 7.44. If R is a non-zero integral domain, then the following are equivalent.

(A) 0 is the only vanishing polynomial in $R[x_1,\ldots,x_n]$.

(B) R has infinitely many elements.

As is often the case, we can't answer that question immediately. Before proving Proposition 7.44, we need the following, extraordinary theorem.

Theorem 7.45 (The Factor Theorem)**.** If R is a non-zero integral domain, $f \in R[x]$, and $a \in R$, then $f(a) = 0$ if and only if $x - a$ divides $f(x)$.

To prove Theorem 7.45, we need to make precise our notions of addition and multiplication of polynomials.

Definition 7.46. To **add** two polynomials $f, g \in R[x_1, \ldots, x_n]$, let $\mathcal{T} = \mathcal{T}_f \cup \mathcal{T}_g$. Choose $a_t, b_t \in R$ such that

$$f = \sum_{t \in \mathcal{T}} a_t t \quad \text{and} \quad g = \sum_{t \in \mathcal{T}} b_t t.$$

We add the polynomials by adding like terms; that is,

$$f + g = \sum_{t \in \mathcal{T}} (a_t + b_t)\, t.$$

To **multiply** f and g, compute the sum of all products of terms in the first polynomial with terms in the second; that is,

$$fg = \sum_{t \in \mathcal{T}} \sum_{u \in \mathcal{T}} (a_t b_u)\, (tu).$$

We use u in the second summand to distinguishes the terms of g from those of f. Notice that fg is really the distribution of g to the terms of f, followed by the distribution of each term of f to the terms of g.

Proof of the Factor Theorem. If $x - a$ divides $f(x)$, then there exists $q \in R[x]$ such that $f(x) = (x-a) \cdot q(x)$. By substitution, $f(a) = (a-a) \cdot q(a) = 0_R \cdot q(a) = 0_R$.

Conversely, assume $f(a) = 0$. You will show in Exercise 7.49 that we can write $f(x) = q(x) \cdot (x-a) + r$ for some $r \in R$. Thus

$$0 = f(a) = q(a) \cdot (a-a) + r = r,$$

and substitution yields $f(x) = q(x) \cdot (x-a)$. In other words, $x - a$ divides $f(x)$, as claimed. □

We now turn our attention to proving Proposition 7.44.

Proof of Lemma 7.44. Assume that R is a non-zero integral domain.

(A) ⇒ (B): We proceed by the contrapositive. Assume that R has finitely many elements. We can enumerate them all as $r_1, r_2, \ldots, r_m$. Let

$$f(x_1, \ldots, x_n) = (x_1 - r_1)(x_1 - r_2) \cdots (x_1 - r_m).$$

Let $b_1, \ldots, b_n \in R$. By assumption, R is finite, so $b_1 = r_i$ for some $i \in \{1, 2, \ldots, m\}$. Notice that f is not only multivariate, it is also univariate: $f \in R[x_i]$. By the Factor Theorem, $f = 0$. We have shown that ¬(B) implies ¬(A); thus, (A) implies (B).

(A) ⇐ (B): Assume that R has infinitely many elements. Let f be any vanishing polynomial. We proceed by induction on n, the number of variables in $R[x_1, \ldots, x_n]$.

Inductive base: Suppose $n = 1$. By the Factor Theorem, $x - a$ divides f for every $a \in R$. By definition of polynomial multiplication, each distinct factor of f adds 1 to the degree of f; for example, if $f = (x-0)(x-1)$, then $\deg f = 2$. However, the definition of a polynomial implies that f has finite degree. Hence, if $f \neq 0$, then it can be factored as only finitely many polynomials

of the form $x-a$. If so, then choose $a_1, a_2, \ldots, a_n$ such that

$$f = (x-a_1)(x-a_2)\cdots(x-a_n).$$

Since R has infinitely many elements, we can find $b \in R$ such that $b \neq a_1, \ldots, a_n$. That means $b - a_i \neq 0$ for each $i = 1, \ldots, n$. As R is an integral domain,

$$f(b) = (b-a_1)(b-a_2)\cdots(b-a_n) \neq 0.$$

This contradicts the choice of f as a vanishing polynomial. Hence, $f = 0$.

Inductive hypothesis: Assume for all i satisfying $1 \leq i < n$, if $f \in R[x_1, \ldots, x_i]$ is a zero polynomial, then f is the constant polynomial 0.

Inductive step: Let $n > 1$, and $f \in R[x_1, \ldots, x_n]$ be a vanishing polynomial. Let $a_n \in R$, and substitute $x_n = a_n$ into f. Denote the resulting polynomial as g. The substitution means that $x_n \notin \mathcal{T}_g$. Hence, $g \in R\left[x_1, \ldots, x_{n-1}\right]$.

It turns out that g is also a vanishing polynomial in $R\left[x_1, \ldots, x_{n-1}\right]$. *Why?* By way of contradiction, assume that it is not. Then there exist $a_1, \ldots, a_{n-1} \in R$ such that $f\left(a_1, \ldots, a_{n-1}\right) \neq 0$. However, the definition of g implies that

$$f(a_1, \ldots, a_n) = g\left(a_1, \ldots, a_{n-1}\right) \neq 0.$$

This contradicts the choice of f as a vanishing polynomial. The assumption was wrong; g must be a vanishing polynomial in $R\left[x_1, \ldots, x_{n-1}\right]$, after all. We can now apply the inductive hypothesis, and infer that g is the constant polynomial 0.

We chose a_n arbitrarily, so this argument holds for any $a_n \in R$. Thus, any of the terms of f containing any of the variables $x_1, \ldots, x_{n-1}$ has a coefficient of zero. The only non-zero terms are those whose only variables are x_n, so $f \in R[x_n]$. This time, the inductive base implies that f is zero. □

We come to the main purpose of this section.

Theorem 7.47. The univariate and multivariate polynomial rings over a ring R are themselves rings.

Proof. Let $n \in \mathbb{N}^+$ and R a ring. We claim that $R[x_1, \ldots, x_n]$ is a ring. To consider the requirements of a ring, let f, g, $h \in R[x_1, \ldots, x_n]$, and let $\mathcal{T} = \mathcal{T}_f \cup \mathcal{T}_g \cup \mathcal{T}_h$. For each $t \in \mathcal{T}$, choose $a_t, b_t, c_t \in R$ such that

$$f = \sum a_t t, \quad g = \sum b_t t, \quad h = \sum c_t t.$$

(Naturally, if $t \in \mathcal{T} \backslash \mathcal{T}_f$, then $a_t = 0$; if $t \in \mathcal{T} \backslash \mathcal{T}_g$, then $b_t = 0$, and if $t \in \mathcal{T} \backslash \mathcal{T}_h$, then $c_t = 0$.) Although we do not write it, all the sums below are indexed over $t \in \mathcal{T}$.

(R1) First we show that $R[x_1, \ldots, x_n]$ is an abelian group.

(closure) By the definition of polynomial addition,

$$(f+g)(x) = \sum (a_t + b_t) t.$$

Since R is closed under addition, we conclude that $f+g \in R[x_1,\ldots,x_n]$.

(associative) We rely on the associativity of R:

$$\begin{aligned} f+(g+h) &= \sum a_t t + \left(\sum b_t t + \sum c_t t\right) \\ &= \sum a_t t + \sum (b_t + c_t)\, t \\ &= \sum [a_t + (b_t + c_t)]\, t \\ &= \sum [(a_t + b_t) + c_t]\, t \\ &= \sum (a_t + b_t)\, t + \sum_{t\in T} c_t t \\ &= \left(\sum a_t t + \sum b_t t\right) + \sum c_t t \\ &= (f+g)+h. \end{aligned}$$

(identity) We claim that the constant polynomial 0 is the identity. Recall that 0 is a polynomial whose coefficients are all 0. We have

$$\begin{aligned} f+0 &= \sum a_t t + 0 \\ &= \sum a_t t + \sum 0\cdot t \\ &= \sum (a_t + 0)\, t \\ &= f. \end{aligned}$$

(inverse) Let $p = \sum (-a_t)\, t$. We claim that p is the additive inverse of f. In fact,

$$\begin{aligned} p+f &= \sum (-a_t)\, t + \sum a_t t \\ &= \sum (-a_t + a_t)\, t \\ &= \sum 0\cdot t \\ &= 0. \end{aligned}$$

(commutative) By the definition of polynomial addition, $g+f = \sum (b_t + a_t)\, t$. Since R is commutative under addition, addition of coefficients is commutative, so

$$\begin{aligned} f+g &= \sum a_t t + \sum b_t t \\ &= \sum (a_t + b_t)\, t \\ &= \sum (b_t + a_t)\, t \\ &= \sum b_t t + \sum a_t t \\ &= g+f. \end{aligned}$$

(R2) Next, we show that $R[x_1,\ldots,x_n]$ is a semigroup.

(closed) Applying the definition of polynomial multiplication, we have

$$fg = \sum_{t\in T} \left[\sum_{u\in T} (a_t b_u)\,(tu)\right].$$

Since R is closed under multiplication, each $(a_t b_u)(tu)$ is a term. Thus fg is a sum of sums of terms, or a sum of terms. In other words, $fg \in R[x_1, \ldots, x_n]$.

(associative) We start by applying the product fg, then multiplying the result to h:

$$\begin{aligned}(fg)h &= \left[\sum_{t\in T}\left[\sum_{u\in T}(a_t b_u)(tu)\right]\right]\cdot\sum_{v\in T}c_v v \\ &= \sum_{t\in T}\left[\sum_{u\in T}\left[\sum_{v\in T}[(a_t b_u)c_v][(tu)v]\right]\right].\end{aligned}$$

Now apply the associative property of multiplication in R:

$$(fg)h = \sum_{t\in T}\left[\sum_{u\in T}\left[\sum_{v\in T}[a_t(b_u c_v)][t(uv)]\right]\right].$$

(Notice the associative property of R applies to terms over R, as well, inasmuch as those terms represent undetermined elements of R.) Now unapply the product:

$$\begin{aligned}(fg)h &= \sum_{t\in T}\left[\sum_{u\in T}\left[\sum_{v\in T}[a_t(b_u c_v)][t(uv)]\right]\right] \\ &= \sum_{t\in T}a_t t\cdot\left[\sum_{u\in T}\left[\sum_{v\in T}(b_u c_v)(uv)\right]\right] \\ &= f(gh).\end{aligned}$$

(R3) To show the distributive property, first apply addition, then multiplication:

$$\begin{aligned}f(g+h) &= \sum_{t\in T}a_t t\cdot\left(\sum_{u\in T}b_u u+\sum_{u\in T}c_u u\right) \\ &= \sum_{t\in T}a_t t\cdot\sum_{u\in T}(b_u+c_u)u \\ &= \sum_{t\in T}\left[\sum_{u\in T}[a_t(b_u+c_u)](tu)\right].\end{aligned}$$

Now apply the distributive property in the ring, and unapply the addition and multipli-

cation:

$$\begin{aligned} f(g+h) &= \sum_{t\in T}\left[\sum_{u\in T}(a_t b_u + a_t c_u)(tu)\right] \\ &= \sum_{t\in T}\left[\sum_{u\in T}[(a_t b_u)(tu) + (a_t c_u)(tu)]\right] \\ &= \sum_{t\in T}\left[\sum_{u\in T}(a_t b_u)(tu) + \sum_{u\in T}(a_t c_u)(tu)\right] \\ &= \sum_{t\in T}\left[\sum_{u\in T}(a_t b_u)(tu)\right] + \sum_{t\in T}\left[\sum_{u\in T}(a_t c_u)(tu)\right] \\ &= fg + fh. \end{aligned}$$

(commutative) Since we are working in commutative rings, we must also show that that $R[x_1,\ldots,x_n]$ is commutative. This follows from the commutativity of R:

$$\begin{aligned} fg &= \left(\sum_{t\in T} a_t t\right)\left(\sum_{u\in T} b_u u\right) \\ &= \sum_{t\in T}\sum_{u\in T}(a_t b_u)(tu) \\ &= \sum_{u\in T}\sum_{t\in T}(b_u a_t)(ut) \\ &= gf. \end{aligned}$$

(We can swap the sums because of the commutative and associative properties of addition.)

□

Exercises.

Exercise 7.48. Let $f(x) = x$ and $g(x) = x+1$ in $\mathbb{Z}_2[x]$.

(a) Show that f and g are not vanishing polynomials.
(b) Compute the polynomial $p = fg$.
(c) Show that $p(x)$ is a vanishing polynomial.
(d) Explain why this does *not* contradict Proposition 7.44.

Exercise 7.49. Fill in each blank of Figure 7.1 with the justification.

Exercise 7.50. Pick at random a degree 5 polynomial f in $\mathbb{Z}[x]$. Then pick at random some $a \in \mathbb{Z}$.

(a) Find $q \in \mathbb{Z}[x]$ and $r \in \mathbb{Z}$ such that $f(x) = q(x)\cdot(x-a) + r$.
(b) Explain why you *cannot* pick a nonzero integer b at random and expect willy-nilly to find $q \in \mathbb{Z}[x]$ and $r \in \mathbb{Z}$ such that $f(x) = q(x)\cdot(bx-a) + r$.
(c) Explain why you *can* pick a nonzero integer b at random and expect willy-nilly to find $q \in \mathbb{Z}[x]$ and $r, s \in \mathbb{Z}$ such that $s\cdot f(x) = q(x)\cdot(bx-a) + r$. (Neat, huh?)

Let R be an integral domain, $f \in R[x]$, and $a \in R$.
Claim: There exist $q \in R[x]$ and $r \in R$ such that $f(x) = q(x) \cdot (x-a) + r$.
Proof:

1. Without loss of generality, we may assume that $\deg f = n$.
2. By _____, choose $a_1, \ldots, a_n$ such that $f = \sum_{k=1}^{n} a_k x^k$. We proceed by induction on n.
3. For the *inductive base,* assume that $n = 0$. Then $q(x) =$_____ and $r =$_____.
4. For the *inductive hypothesis,* assume that for all $i \in \mathbb{N}$ satisfying $0 \leq i < n$, there exist $q \in R[x]$ and $r \in R$ such that $f(x) = q(x) \cdot (x-a) + r$.
5. For the *inductive step,*
   (a) Let $p(x) = a_n x^{n-1}$, and $g(x) = f(x) - p(x) \cdot (x-a)$.
   (b) Notice that $\deg g <$_____.
   (c) By _____, there exist $p' \in R[x]$ and $r \in R$ such that $g(x) = p'(x) \cdot (x-a) + r$.
   (d) Let $q = p + p'$. By _____, $q \in R[x]$.
   (e) By _____ and _____, $f(x) = q(x) \cdot (x-a) + r$.
6. We have shown that, for arbitrary n, we can find $q \in R[x]$ and $r \in R$ such that $f(x) = q(x) \cdot (x-a) + r$. The claim holds.

Figure 7.1. Material for Exercise 7.49

(d) If the requirements of (b) were changed to finding $q \in \mathbb{Q}[x]$ and $r \in \mathbb{Q}$, would you then be able to carry out (b)? Why or why not?

Exercise 7.51. Let $R = \mathbb{Z}_3[x]$ and $f(x) = x^3 + 2x + 1 \in R$.
(a) Explain how we can infer that f does not factor in R without performing a brute force search of factorizations.
(b) If we divide $g \in R$ by f, how many possible remainders can we obtain?

Exercise 7.52. Let R be an integral domain.
(a) Show that $R[x]$ is also an integral domain.
(b) How does this not contradict Exercise 7.48? After all, $\mathbb{Z}_2$ is a field, and thus an integral domain!

Exercise 7.53. Let R be a ring, and $f, g \in R[x]$. Show that $\deg(f+g) \leq \max(\deg f, \deg g)$.

Exercise 7.54. Let R be a ring and define

$$R(x) = \mathrm{Frac}(R[x]);$$

for example,

$$\mathbb{Z}(x) = \mathrm{Frac}(\mathbb{Z}[x]) = \left\{ \frac{p}{q} : \ p, q \in \mathbb{Z}[x] \right\}.$$

Is $R(x)$ a ring? is it a field?

Exercise 7.55. Let $R = \mathbb{Q}\left[\sqrt{2}\right]$, an extension of $\mathbb{Q}$ by $\sqrt{2}$. (See Exercise 7.19.)
(a) Find $g \in \mathbb{Q}[x]$ such that g factors with coefficients in R, but not with coefficients in $\mathbb{Q}$.
(b) Let $S = \mathbb{Q}\left[\sqrt{2}+\sqrt{3}\right]$ and $T = R\left[\sqrt{3}\right]$. Show that $S = T$.

(c) Is $\mathbb{Z}\left[\sqrt{2}+\sqrt{3}\right] = \mathbb{Z}\left[\sqrt{2}\right]\left[\sqrt{3}\right]$?

Exercise 7.56. Let $p \in \mathbb{Z}$ be irreducible, and $R = \mathbb{Z}_p[x]$. Show that $\varphi : R \to R$ by $\varphi(f) = f^p$ is a group automorphism. This is called the **Frobenius automorphism**.

7.4: Euclidean domains

In this section we consider an important similarity between the ring of integers and the ring of polynomials. This similarity will motivate us to define a new kind of ring. We will then show that all rings of this type allow us to perform important operations that we find both useful and necessary. What is the similarity? The ability to *divide with remainder*.

Division of polynomials

We start with polynomials, but we will take this a step higher in a moment.

Theorem 7.57 (The Division Theorem for polynomials)**.** Let $\mathbb{F}$ be a field, and consider the polynomial ring $\mathbb{F}[x]$. Let $f, g \in \mathbb{F}[x]$ with $f \neq 0$. There exist unique $q, r \in \mathbb{F}[x]$ satisfying (D1) and (D2) where
(D1) $g = qf + r$;
(D2) $r = 0$ or $\deg r < \deg f$.
We call g the **dividend**, f the **divisor**, q the **quotient**, and r the **remainder**.

Proof. The proof is essentially the procedure of long division of polynomials.

If $g = 0$, let $r = q = 0$. Then $g = qf + r$ and $r = 0$.

Now assume $g \neq 0$. If $\deg g < \deg f$, let $r = g$ and $q = 0$. Then $g = qf + r$ and $\deg r < \deg f$.

Otherwise, $\deg g \geq \deg f$. Let $m = \deg f$ and $n = \deg g - \deg f$. We proceed by induction on n.

For the *inductive base* $n = 0$, we have $\deg g = \deg f = m$. Let $a_m, \ldots, a_1, b_m, \ldots, b_1 \in R$ such that

$$g = a_m x^m + a_{m-1} x^{m-1} + \cdots + a_1 x + a_0$$
$$f = b_m x_m + b_{m-1} x^{m-1} + \cdots + b_1 x + b_0.$$

Let $q = \frac{a_m}{b_m}$ and $r = g - qf$. Since $\mathbb{F}$ is a field and $b_m \neq 0$, we can safely conclude that q is a constant polynomial. Arithmetic shows that $g = qf + r$, but can we guarantee that $r = 0$ or $\deg r < \deg f$? Apply substitution, distribution, and polynomial addition to obtain

$$\begin{aligned} r &= g - qf \\ &= \left(a_m x^m + a_{m-1}x^{m-1} + \cdots + a_1 x + a_0\right) \\ &\quad - \frac{a_m}{b_m}\left(b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0\right) \\ &= \left(a_m - \frac{a_m}{b_m}\cdot b_m\right)x^m + \left(a_{m-1} - \frac{a_m}{b_m}\cdot b_{m-1}\right)x^{m-1} + \cdots + \left(a_0 - \frac{a_m}{b_m}\cdot b_0\right) \\ &= 0x^m + \left(a_{m-1} - \frac{a_m}{b_m}\cdot b_{m-1}\right)x^{m-1} + \cdots + \left(a_0 - \frac{a_m}{b_m}\cdot b_0\right). \end{aligned}$$

Since the coefficient of x^m is zero, we see that if $r \neq 0$, then $\deg r < \deg f$.

For the *inductive hypothesis*, assume that for all $i < n$ there exist $q, r \in R[x]$ such that $g = qf + r$ and $r = 0$ or $\deg r < \deg f$.

For the *inductive step*, let $\ell = \deg g$. Let $a_m, \ldots, a_0, b_\ell, \ldots, b_0 \in R$ such that

$$\begin{aligned} f &= a_m x^m + \cdots + a_0 \\ g &= b_\ell x^\ell + \cdots + b_0. \end{aligned}$$

Let $p = \frac{b_\ell}{a_m}\cdot x^n$ and $r = g - pf$. Once again, since $\mathbb{F}$ is a field and $a_m \neq 0$, we can safely conclude that $p \in \mathbb{F}[x]$. Apply substitution and distribution to obtain

$$\begin{aligned} g' &= g - pf \\ &= g - \frac{b_\ell}{a_m}\cdot x^n\left(a_m x^m + \cdots + a_0\right) \\ &= g - \left(b_\ell x^{m+n} + \frac{b_\ell a_{m-1}}{a_m}\cdot x^{m-1+n} + \cdots + \frac{b_\ell a_0}{a_m}\cdot x^n\right). \end{aligned}$$

Recall that $n = \deg g - \deg f = \ell - m$, so $\ell = m + n$. Apply substitution and polynomial addition to obtain

$$\begin{aligned} g' = g - pf &= \left(b_\ell x^\ell + \cdots + b_0\right) \\ &\quad - \left(b_\ell x^\ell + \frac{b_\ell a_{m-1}}{a_m}\cdot x^{\ell-1} + \cdots + \frac{b_\ell a_0}{a_m}\cdot x^n\right) \\ &= 0x^\ell + \left(b_{\ell-1} - \frac{b_\ell a_{m-1}}{a_m}\right)x^{\ell-1} \\ &\quad + \cdots + \left(b_n - \frac{b_\ell a_0}{a_m}\right)x^n + b_{n-1}x^{n-1}\cdots + b_0. \end{aligned}$$

Since $\mathbb{F}$ is a field and $a_m \neq 0$, we can safely conclude that $g' \in \mathbb{F}[x]$. Observe that $\deg g' < \ell = \deg g$, so $\deg g' - \deg f < n$. Apply the inductive hypothesis to find $p', r \in R[x]$ such that

$g' = p'f + r$ and $r = 0$ or $\deg r < \deg f$. Then

$$\begin{aligned} g &= pf + g' = pf + (p'f + r) \\ &= (p + p')f + r. \end{aligned}$$

Let $q = p + p'$. By closure, $q \in R[x]$, and we have shown the existence of a quotient and remainder.

For uniqueness, assume that there exist $q_1, q_2, r_1, r_2 \in R[x]$ such that $g = q_1 f + r_1 = q_2 f + r_2$ and $\deg r_1, \deg r_2 < \deg f$. Then

$$\begin{aligned} q_1 f + r_1 &= q_2 f + r_2 \\ 0 &= (q_2 - q_1) f + (r_2 - r_1). \end{aligned} \tag{26}$$

If $q_2 - q_1 \neq 0$, then no term of $(q_2 - q_1)\operatorname{lm}(f)$ has degree smaller than $\deg f$. Since every term of $r_2 - r_1$ has degree smaller than $\deg f$, there are no like terms between the two. Thus, there can be no cancellation between $(q_2 - q_1)\operatorname{lm}(f)$ and $r_2 - r_1$, and for similar reasons there can be no cancellation between $(q_2 - q_1)\operatorname{lm}(f)$ and lower-degree terms of $(q_2 - q_1)f$. However, the left hand side of equation 26 is the zero polynomials, so coefficients of $(q_2 - q_1)\operatorname{lm}(f)$ are all 0 on the left hand side. They must likewise be all zero on the right hand side. That implies $(q_2 - q_1)\operatorname{lm}(f)$ is equal to the constant polynomial 0. We are working in an integral domain (Exercise 7.52), and $\operatorname{lm}(f) \neq 0$, so it mus tbe that $q_2 - q_1 = 0$. In other words, $q_1 = q_2$.

Once we have $q_2 - q_1 = 0$, substitution into (26) implies that $0 = r_2 - r_1$. Immediately we have $r_1 = r_2$. We have shown that q and r are unique. □

Notice that the theorem does *not* apply if $R = \mathbb{Z}$, and Exercise 7.50 explains why. That's a shame.

Euclidean domains

Recall from Section 6.1 that the Euclidean algorithm for integers is basically repeated division. You can infer, more or less correctly, that a similar algorithm works for polynomials.

Why stop there? We have a notion of divisibility in rings, and we just found that the Division Theorem for integers can be generalized to any polynomial ring whose ground ring is a field. Can we generalize the Division Theorem beyond a ring of polynomials over a field? We can, but it requires us to generalize the notion of a remainder, as well.

Definition 7.58. Let R be an integral domain and v a function mapping the nonzero elements of R to $\mathbb{N}^+$. We say that R is a **Euclidean Domain** with respect to the **valuation function** v if it satisfies (E1) and (E2) where

(E1) $v(r) \leq v(rs)$ for all nonzero $r, s \in R$.

(E2) For all nonzero $f \in R$ and for all $g \in R$, there exist $q, r \in R$ such that
- $g = qf + r$, and
- $r = 0$ or $v(r) < v(f)$.

Example 7.59. By the Division Theorem, $\mathbb{Z}$ is a Euclidean domain with the valuation function $v(r) = |r|$.

Theorem 7.60. Let $\mathbb{F}$ be a field. Then $\mathbb{F}[x]$ is a Euclidean domain with the valuation function $v(r) = \deg r$.

Proof. You do it! See Exercise 7.70. □

Example 7.61. On the other hand, $\mathbb{Z}[x]$ is *not* a Euclidean domain if the valuation function is $v(r) = \deg r$. If $f = 2$ and $g = x$, we cannot find $q, r \in \mathbb{Z}[x]$ such that $g = qf + r$ and $\deg r < \deg f$. The best we can do is $x = 0 \cdot 2 + x$, but $\deg x > \deg 2$.

If you think back to the Euclidean algorithm, you might remember that it requires only *the ability to perform a division with a unique remainder that was smaller than the divisor*. This means that we can perform the Euclidean algorithm in a Euclidean ring! — But will the result have the same properties as when we perform it in the ring of integers?

Yes and no. We *do* get an object whose properties resemble those of the greatest common divisor of two integers. However, the result *might not be unique!* On the other hand, if we relax our expectation of uniqueness, we can get a greatest common divisor that is… *sort of* unique.

Definition 7.62. Let R be a ring. If $a, b, r \in R$ satisfy $ar = b$ or $ra = b$, then a **divides** b, a is a **divisor of** b, and b is **divisible** by a.

Now suppose that R is a Euclidean domain with respect to v, and let $a, b \in R$. If there exists $d \in R$ such that $d \mid a$ and $d \mid b$, then we call d a **common divisor** of a and b. If in addition all other common divisors d' of a and b divide d, then d is a **greatest common divisor** of a and b.

Two subtle differences with the definition for the integers have profound consequences.

- The definition refers to "a" greatest common divisor, not "the" greatest common divisor. *There can be many great"est" common divisors!*
- Euclidean domains measure "greatness" using divisibility (or multiplication) rather than order (or subtraction). As a consequence, the Euclidean domain R need not have a well ordering, or even a linear ordering — it needs only a valuation function! This is *why* there can be many great"est" common divisors.

Example 7.63. Consider $x^2 - 1, x^2 + 2x + 1 \in \mathbb{Q}[x]$. By Theorem 7.60, $\mathbb{Q}[x]$ is a Euclidean domain with respect to the valuation function $v(p) = \deg p$. Both of the given polynomials factor:

$$x^2 - 1 = (x + 1)(x - 1) \quad \text{and} \quad x^2 + 2x + 1 = (x + 1)^2,$$

so we see that $x + 1$ is a divisor of both. In fact, it is a greatest common divisor, since no polynomial of degree two divides both $x^2 - 1$ and $x^2 + 2x + 1$.

However, $x + 1$ is not the *only* greatest common divisor. Another greatest common divisor is $2x + 2$. It may not be obvious that $2x + 2$ divides both $x^2 - 1$ and $x^2 + 2x + 1$, but it does:

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

and

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

Notice that $2x+2$ divides $x+1$ and vice-versa; also that $\deg(2x+2) = \deg(x+1)$.

Likewise, $\frac{x+1}{3}$ is also a greatest common divisor of x^2-1 and x^2+2x+1.

This new definition will allow more than one greatest common divisor even in $\mathbb{Z}$! For example, for $a = 8$ and $b = 12$, both 4 *and* -4 are greatest common divisors! This happens because each divides the other, emphasizing that in a generic Euclidean domain, the notion of a "greatest" common divisor is relative to divisibility, not to other orderings. However, when speaking of greatest common divisors in the integers, we typically use the ordering, not divisibility.

That said, all greatest common divisors have something in common.

Proposition 7.64. Let R be a Euclidean domain with respect to v, and $a, b \in R$. Suppose that d is a greatest common divisor of a and b. If d' is a common divisor of a and b, then $v(d') \leq v(d)$. If d' is another greatest common divisor of a and b, then $v(d) = v(d')$.

Proof. Since d is a greatest common divisor of a and b, and d' is a common divisor, the definition of a greatest common divisor tells us that d divides d'. Thus there exists $q \in R$ such that $qd' = d$. From property (E1) of a Euclidean domain,

$$v(d') \leq v(qd') = v(d).$$

On the other hand, if d' is also a greatest common divisor of a and b, an argument similar to the one above shows that

$$v(d) \leq v(d') \leq v(d).$$

Hence $v(d) = v(d')$. □

Finally we come to the point of a Euclidean domain: we can use the Euclidean algorithm to compute a gcd of any two elements! Essentially we transcribe the Euclidean Algorithm for integers (Theorem 6.4 on page 178 of Section 6.1).

Theorem 7.65 (The Euclidean Algorithm for Euclidean domains)**.** Let R be a Euclidean domain with valuation v and $m, n \in R \setminus \{0\}$. One can compute a greatest common divisor of m, n in the following way:

1. Let $s = m$ and $t = n$.
2. Repeat the following steps until $t = 0$:
   (a) Let q be the quotient and r the remainder after dividing s by t.
   (b) Assign s the current value of t.
   (c) Assign t the current value of r.

The final value of s is a greatest common divisor of m and n.

Proof. You do it! See Exercise 7.71. □

Just as we could adapt the Euclidean Algorithm for integers to the Extended Euclidean Algorithm in order to compute $a, b \in \mathbb{Z}$ such that Bezout's Identity holds,

$$am + bn = \gcd(m, n),$$

we can do the same in Euclidean domains. You will need this for Exercise 7.71.

Exercises.

Exercise 7.66. Try to devise a division algorithm for $\mathbb{Z}_n$? Does the value of n matter?

Exercise 7.67. Let $f = 2x^2 + 1$ and $g = x^3 - 1$.

(a) Show that 1 is a greatest common divisor of f and g in $\mathbb{Q}[x]$, and find $a, b \in \mathbb{Q}[x]$ such that $1 = af + bg$.

(b) Recall that $\mathbb{Z}_5$ is a field. Show that 1 is a greatest common divisor of f and g in $\mathbb{Z}_5[x]$, and find $a, b \in \mathbb{Z}_5[x]$ such that $1 = af + bg$.

(c) Recall that $\mathbb{Z}[x]$ is not a Euclidean domain. Explain why the result of part (a) cannot be used to show that 1 is a greatest common divisor of f and g in $\mathbb{Z}[x]$. What would you get if you used the Euclidean algorithm on f and g in $\mathbb{Z}[x]$?

Exercise 7.68. Let $f = x^4 + 9x^3 + 27x^2 + 31x + 12$ and $g = x^4 + 13x^3 + 62x^2 + 128x + 96$.

(a) Compute a greatest common divisor of f and g in $\mathbb{Q}[x]$.

(b) Recall that $\mathbb{Z}_{31}$ is a field. Compute a greatest common divisor of f and g in $\mathbb{Z}_{31}[x]$.

(c) Recall that $\mathbb{Z}_3$ is a field. Compute a greatest common divisor of f and g in $\mathbb{Z}_3[x]$.

(d) Even though $\mathbb{Z}[x]$ is not a Euclidean domain, it still has greatest common divisors. What's more, we can compute the greatest common divisors using the Euclidean algorithm! How?

(e) You can even compute the greatest common divisors *without* using the Euclidean algorithm, but by examining the answers to parts (b) and (c) slowly. How?

Exercise 7.69. Show that every field is a Euclidean domain.

Exercise 7.70. Prove Theorem 7.60.

Exercise 7.71. Prove Theorem 7.65, the Euclidean Algorithm for Euclidean domains.

Exercise 7.72. A famous Euclidean domain is the ring of *Gaussian integers*

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

where $i^2 = -1$. Let $v : \mathbb{Z}[i] \to \mathbb{Z}$ by

$$v(a + bi) = a^2 + b^2.$$

(a) Show that $a + bi$ is "orthogonal" to $i(a + bi)$, in the sense that the slope of the line segment connecting 0 and $a + bi$ in the complex plane is orthogonal to the slope of the line segment connecting 0 and $i(a + bi)$.

(b) Assuming the facts given about v, divide:
  (i) 11 by 3;
  (ii) 11 by $3i$;
  (iii) $2 + 3i$ by $1 + 2i$.

(c) Show that v is, in fact, a valuation function suitable for a Euclidean domain.

(d) Describe a method for dividing Gaussian integers. (Again, it helps to think of them as vectors in the plane. See Exercise 0.52 on page 19.)

Chapter 8:
Ideals

This chapter fills two roles. Some sections describe ring analogs to structures that we introduced in group theory:

- Section 8.1 introduces the *ideal*, an analog to a normal subgroup;
- Section 8.3 provides an analog of quotient groups; and
- Section 8.5 decribes ring homomorphisms.

Two of the remaining sections use these ring structures to introduce new kinds of ring structures:

- Section 8.2 describes an important class of rings; and
- Section 8.4 highlights an important class of ideals.

Finally,

- Section 84 discusses the Fundamental Theorem of Algebra: the idea that every polynomial with complex coefficients has a complex root.

8.1: Ideals

Given that normal subgroups were so important to group theory, it will not surprise you that a special kind of subring plays an crucial role in ring theory. But, what sort of characteristic should it have? Rather than take the structural approach that we took last time, and find a criterion on a subring that guarantees we can create a "quotient" that gives us a new ring, let's look at the mathematical applications of rings that interest us.

An application which may strike the reader as more concrete is the question of the roots of polynomials. Start with a ring R, an element $a \in R$, and three univariate polynomials f, g, and p over R. How do the roots of f and/or g behave with respect to ring operations? If a is a root of both f and g, then a is also a root of their sum $h = f + g$, since

$$h(a) = (f+g)(a) = f(a) + g(a) = 0.$$

Also, if a is a root *only of* f, then it is a root of *any* multiple of f, such as $h = fp$. After all,

$$h(a) = (fp)(a) = f(a)\,p(a) = 0 \cdot p(a) = 0.$$

Something subtle is going on here, and you may have missed it, so let's look more carefully. Let S be the subring of R that contains *all* polynomials that have a as a root. By definition, f, g, and h are all in S, but p is not! Compare this to group theory: the product of an element of a subgroup and a element outside the subgroup is *never* in the subgroup. What we are seeing is a property we had studied way back when we looked at monoids: S is an *absorbing subset* of R.

Notice how absorption creates an important difference from group theory. With groups, multiplying an element of a subgroup with an element outside the subgroup *always* gave us another element outside the subgroup! This allowed us to create cosets, and partition the group. We obviously cannot rely on this property to the same thing in rings, because some subrings absorb multiplication from outside the subgroup! You might argue that it still holds for addition, and that is true – in fact, we will use that fact later to create cosets that partition a ring.

Recall our definition of S as the subring of R that contains all polynomials that have a as a root. This definition is quite simple, and clearly important. The fact that S "absorbs" any polynomial that does *not* have a as a root indicates that the absorption property is important. This property likewise occurs with other subrings that have straightforward and obvious definitions; for example, the subring A of $\mathbb{Z}$ that contains all multiples of 4 and 6: $3 \notin A$, but $4 \cdot 3 \in A$ and $6 \cdot 3 \in A$. When a property appears in many contexts that are very different but important, it merits investigation.

Definition and examples

Definition 8.1. Let A be a subring of R that satisfies the **absorption property**:

$$\forall r \in R \quad \forall a \in A \qquad ra \in A.$$

Then A is an **ideal subring** of R, or simply, an **ideal**, and we write $A \triangleleft R$. An ideal A is **proper** if $\{0\} \neq A \neq R$.

Recall that our rings are assumed to be commutative, so if $ra \in A$ then $ar \in A$, also.

Example 8.2. Recall the subring $2\mathbb{Z}$ of the ring $\mathbb{Z}$. We claim that $2\mathbb{Z} \triangleleft \mathbb{Z}$. Why? Let $r \in \mathbb{Z}$, and $a \in 2\mathbb{Z}$. By definition of $2\mathbb{Z}$, there exists $d \in \mathbb{Z}$ such that $a = 2d$. Substitution gives us

$$ra = r \cdot 2d = 2(rd) \in 2\mathbb{Z},$$

so $2\mathbb{Z}$ "absorbs" multiplication by $\mathbb{Z}$. This makes $2\mathbb{Z}$ an ideal of $\mathbb{Z}$.

Naturally, we can generalize this proof to arbitrary $n \in \mathbb{Z}$: see Exercises 8.14 and 8.15.

Ideals in the ring of integers have a nice property that we will use in future examples.

Example 8.3. Certainly $3 \mid 6$ since $3 \cdot 2 = 6$. Look at the ideals generated by 3 and 6:

$$3\mathbb{Z} = \{\ldots, -12, -9, -6, -3, 0, 3, 6, 9, 12, \ldots\}$$
$$6\mathbb{Z} = \{\ldots, -12, -6, 0, 6, 12, \ldots\}.$$

Inspection suggests that $6\mathbb{Z} \subseteq 3\mathbb{Z}$. Is it? Let $x \in 6\mathbb{Z}$. By definition, $x = 6q$ for some $q \in \mathbb{Z}$. By substitution, $x = (3 \cdot 2)q = 3(2 \cdot q) \in 3\mathbb{Z}$. Since x was arbitrary in $6\mathbb{Z}$, we have $6\mathbb{Z} \subseteq 3\mathbb{Z}$.

Lemma 8.4. Let $a, b \in \mathbb{Z}$. The following are equivalent:
(A) $a \mid b$;
(B) $b\mathbb{Z} \subseteq a\mathbb{Z}$.

Proof. You do it! See Exercise 8.16. □

Earlier in the section, we looked at roots of univariate polynomials. The same properties hold when we move to multivariate polynomials. If $a_1, \ldots, a_n \in R$, $f \in R[x_1, \ldots, x_n]$, and $f(a_1, \ldots, a_n) = 0$, then we call $(a_1, \ldots, a_n)$ a **root** of f.

Example 8.5. You showed in Exercise 7.3 that $\mathbb{C}[x, y]$ is a ring. Let $f = x^2 + y^2 - 4$, $g = xy - 1$, and $A = \{hf + kg : h, k \in \mathbb{C}[x, y]\}$. From a geometric perspective what's interesting about A is that *the common roots of f and g are roots of any element of A.* To see this, let (α, β) be a

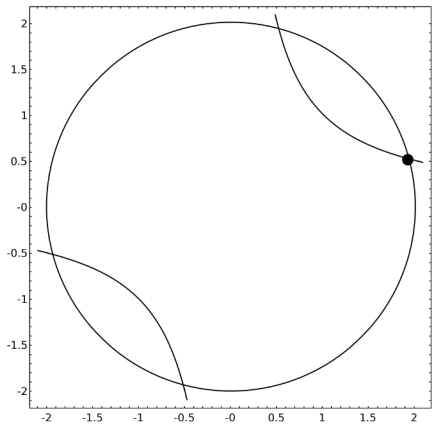


Figure 8.1. A common root of x^2+y^2-4 and $xy-1$

common root of f and g; that is, $f(\alpha,\beta)=g(\alpha,\beta)=0$. Let $p\in A$; by definition, we can write $p=hf+kg$ for some $h,k\in\mathbb{C}[x,y]$. By substitution,

$$\begin{aligned}p(\alpha,\beta)&=(hf+kg)(\alpha,\beta)\\&=h(\alpha,\beta)\cdot f(\alpha,\beta)+k(\alpha,\beta)\cdot g(\alpha,\beta)\\&=h(\alpha,\beta)\cdot 0+k(\alpha,\beta)\cdot 0\\&=0;\end{aligned}$$

that is, (α,β) is a root of p. Figure 8.1 depicts the root

$$(\alpha,\beta)=\left(\sqrt{2+\sqrt{3}},2\sqrt{2+\sqrt{3}}-\sqrt{6+3\sqrt{3}}\right).$$

The remarkable thing is that A is an ideal. To show this, we must show that A is a subring of $\mathbb{C}[x,y]$ that absorbs multiplication.

- Is A a subring? Let $a,b\in A$. By definition, we can find $h_a,h_b,k_a,k_b\in\mathbb{C}[x,y]$ such that $a=h_af+k_ag$ and $b=h_bf+k_bg$. A little arithmetic gives us

$$\begin{aligned}a-b&=(h_af+k_ag)-(h_bf+k_bg)\\&=(h_a-h_b)f+(k_a-k_b)g\in A.\end{aligned}$$

  To show that $ab\in A$, we will distribute over *one* of the two polynomials:

$$\begin{aligned}ab&=a(h_bf+k_bg)\\&=a(h_bf)+a(k_bg)\\&=(ah_b)f+(ak_b)g.\end{aligned}$$

  Let

$$h'=ah_b\quad\text{and}\quad k'=ak_b;$$

then $ab = h'f + k'g$, and by closure, $h', k' \in \mathbb{C}[x,y]$. By definition, $ab \in A$, as well. By the Subring Theorem, A is a subring of $\mathbb{C}[x,y]$.

- Does A absorb multiplication? Let $a \in A$, and $r \in \mathbb{C}[x,y]$. By definition, we can write $a = h_a f + k_a g$, as above. A little arithmetic gives us

$$\begin{aligned} ra &= r(h_a f + k_a g) = r(h_a f) + r(k_a g) \\ &= (rh_a) f + (rk_a) g \in A. \end{aligned}$$

Let

$$h' = rh_a \quad \text{and} \quad k' = rk_a;$$

then $ra = h'f + k'g$, and by closure, $h', k' \in \mathbb{C}[x,y]$. By definition, $ra \in A$, as well. By definition, A satisfies the absorption property.

We have shown that A satisfies the subring and absorption properties; thus, $A \triangleleft \mathbb{C}[x,y]$.

You will show in Exercise 8.23 that the ideal of Example 8.5 can be generalized to other rings and larger numbers of variables.

Remark 8.6. Recall from linear algebra that *vector spaces* are an important tool for the study of systems of linear equations. If we find a *triangular basis* of a system of linear polynomials, we can analyze the subspace of solutions of the system.

Example 8.5 illustrates that ideals are an important analog for non-linear polynomial equations. If we can find a "triangular basis" of an ideal, then we can analyze the solutions of the system in a method very similar to methods for linear systems. We take up this task in Chapter 11.

Properties and elementary theory

Since ideals are fundamental, we would like an analog of the Subring Theorem to decide whether a subset of a ring is an ideal. You might have noticed from the example above that absorption actually implies closure under multiplication. After all, if $rb \in A$ for every $r \in R$, then since $a \in A$ implies $a \in R$, we really have $ab \in A$, too. The Ideal Theorem uses this fact to simplify the criteria for an ideal.

Theorem 8.7 (The Ideal Theorem)**.** Let R be a ring and $A \subseteq R$ with A nonempty. The following are equivalent:

(A) A is an ideal subring of R.

(B) A is closed under subtraction and absorption. That is,

- (I1) for all $a, b \in A$, $a - b \in A$; and
- (I2) for all $a \in A$ and $r \in R$, we have $ar, ra \in A$.

Proof. You do it! See Exercise 8.18. □

We conclude by defining a special kind of ideal, with a notation similar to that of cyclic subgroups, but with a different meaning.

Notation 8.8. Let R be a ring with unity, $m \in \mathbb{N}^+$, and $r_1, r_2, \ldots, r_m \in R$. Define the set $\langle r_1, r_2, \ldots, r_m \rangle$ as the intersection of all the ideals of R that contain all of $r_1, r_2, \ldots, r_m$.

Proposition 8.9. For all $r_1, \ldots, r_m \in R$, $\langle r_1, \ldots, r_m \rangle$ is an ideal.

We will not prove this proposition, as it is a direct consequence of the next:

Proposition 8.10. For every set $\mathcal{I}$ of ideals of a ring R, $\bigcap_{I \in \mathcal{I}} I$ is also an ideal.

Proof. Denote $J = \bigcap_{I \in \mathcal{I}} I$. Observe that $J \neq \emptyset$ because 0_R is an element of every ideal. Let $a, b \in J$ and $r \in R$. Let $I \in \mathcal{I}$. Since J contains only those elements that appear in every element of $\mathcal{I}$, and $a, b \in J$, we know that $a, b \in I$. By the Ideal Theorem, $a - b \in I$, and also $ra \in I$. Since I was an arbitrary ideal in $\mathcal{I}$, *every* element of $\mathcal{I}$ contains $a - b$ and ra. Thus $a - b$ and every ra are in the intersection of these sets, which is J; in other words, $a - b, ra \in J$. By the Ideal Theorem, J is an ideal. □

Since $\langle r_1, \ldots, r_m \rangle$ is defined as the intersection of ideals containing $r_1, \ldots, r_m$, Proposition 8.10 implies that $\langle r_1, \ldots, r_m \rangle$ is an ideal. It is important enough to identify by a special name.

Definition 8.11. We call $\langle r_1, r_2, \ldots, r_m \rangle$ the **ideal generated by** $r_1, r_2, \ldots, r_m$, and $\{r_1, r_2, \ldots, r_m\}$ a **basis** of $\langle r_1, r_2, \ldots, r_m \rangle$.

This ideal is closely related to the ideal we used in Example 8.5.

Proposition 8.12. If R has unity, then $\langle r_1, r_2, \ldots, r_m \rangle$ is precisely the set

$$I = \{h_1 r_1 + h_2 r_2 + \cdots + h_m r_m : \ h_i \in R\}.$$

Proof. First, we show that $I \subseteq \langle r_1, \ldots, r_m \rangle$. Let $p \in I$; by definition, there exist $h_1, \ldots, h_m \in R$ such that $p = \sum_{i=1}^{m} h_i r_i$. Let J be any ideal that contains all of $r_1, \ldots, r_m$. By absorption, $h_i r_i \in J$ for each i. By closure, $p = \sum_{i=1}^{m} h_i r_i \in J$. Since J was an arbitrary ideal containing all of $r_1, \ldots, r_m$, we infer that all the ideals containing all of $r_1, \ldots, r_m$ contain p. Since p is an arbitrary element of I, I is a subset of all the ideals containing all of $r_1, \ldots, r_m$. By definition, $I \subseteq \langle r_1, \ldots, r_m \rangle$.

Now we show that $I \supseteq \langle r_1, \ldots, r_m \rangle$. We claim that I is an ideal that contains each of $r_1, \ldots, r_m$. If true, the definition of $\langle r_1, \ldots, r_m \rangle$ does the rest, as it consists of elements common to *every* ideal that contains all of $r_1, \ldots, r_m$.

But why is I an ideal? We first consider the absorption property. Let $f \in I$. By definition, there exist $h_1, \ldots, h_m \in R$ such that

$$f = h_1 r_1 + \cdots + h_m r_m.$$

Let $p \in R$; we have

$$pf = (ph_1)\, r_1 + \cdots + (ph_m)\, r_m.$$

By closure, $ph_i \in R$ for each $i = 1, \ldots, m$. We have written pf in a form that satisfies the definition of I, so $pf \in I$. As for the closure of subtraction, let $f, g \in I$; then choose $p_i, q_i \in R$ such that

$$f = p_1 r_1 + \cdots + p_m r_m \text{ and}$$
$$g = q_1 r_1 + \cdots + q_m r_m.$$

Using the associative property, the commutative property of addition, the commutative property of multiplication, distribution, and the closure of subtraction in R, we see that

$$\begin{aligned} f-g &= (p_1r_1+\cdots+p_mr_m)-(q_1r_1+\cdots+q_mr_m) \\ &= (p_1r_1-q_1r_1)+\cdots+(p_mr_m-q_mr_m) \\ &= (p_1-q_1)\,r_1+\cdots+(p_m-q_m)\,r_m. \end{aligned}$$

By closure, $p_i-q_i \in R$ for each $i=1,\ldots,m$. We have written $f-g$ in a form that sastisfies the definition of I, so $f-g \in I$. By the Ideal Theorem, I is an ideal.

But, is $r_i \in I$ for each $i=1,2,\ldots,m$? Well,

$$r_i = 1_R \cdot r_i + \sum_{j\neq i} 0 \cdot r_j \in I.$$

Since R has unity, this expression of r_i satisfies the definition of I, so $r_i \in I$.

Hence I is an ideal containing all of $r_1, r_2, \ldots, r_m$. By definition of $\langle r_1,\ldots,r_m\rangle$, $I \supseteq \langle r_1,\ldots,r_m\rangle$.

We have shown that $I \subseteq \langle r_1,\ldots,r_m\rangle \subseteq I$. Hence $I = \langle r_1,\ldots,r_m\rangle$ as claimed. □

As with vector spaces, the basis of an ideal *is not unique*.

Example 8.13. Consider the ring $\mathbb{Z}$, and let $I = \langle 4,6\rangle$. Proposition 8.12 claims that

$$I = \{4m+6n:\ m,n\in\mathbb{Z}\}.$$

Choosing concrete values of m and n, we see that

$$\begin{aligned} 4 &= 4\cdot 1+6\cdot 0 \in I \\ 0 &= 4\cdot 0+6\cdot 0 \in I \\ -12 &= 4\cdot(-3)+6\cdot 0 \in I \\ -12 &= 4\cdot 0+6\cdot(-2) \in I. \end{aligned}$$

Notice that for some elements of I, we can provide representations in terms of 4 and 6 in more than one way.

While we're at it, we claim that we can simplify I as $I = 2\mathbb{Z}$. Why? For starters, it's pretty easy to see that $2 = 4\cdot(-1)+6\cdot 1$, so $2 \in I$. (Even if it wasn't that easy, though, Bezout's Identity would do the trick: $\gcd(4,6) = 4m+6n$ for some $m,n\in\mathbb{Z}$.) Now that we have $2\in I$, let $x \in 2\mathbb{Z}$; then $x = 2q$ for some $q\in\mathbb{Z}$. By substitution and distribution,

$$x = 2q = [4\cdot(-1)+6\cdot 1]\cdot q = 4\cdot(-q)+6\cdot q \in I.$$

Since x was arbitrary, $I \supseteq 2\mathbb{Z}$. On the other hand, let $x\in I$. By definition, there exist $m,n\in\mathbb{Z}$ such that

$$x = 4m+6n = 2(2m+3n) \in 2\mathbb{Z}.$$

Since x was arbitrary, $I \subseteq 2\mathbb{Z}$. We already showed that $I\subseteq 2\mathbb{Z}$, so we conclude that $I = 2\mathbb{Z}$.

So $I = \langle 4,6\rangle = \langle 2\rangle = 2\mathbb{Z}$. If we think of $r_1,\ldots,r_m$ as a "basis" for $\langle r_1,\ldots,r_m\rangle$, then the example above shows that any given ideal can have bases of different sizes.

You might wonder if every ideal can be written as $\langle a \rangle$, the same way that $I = \langle 4, 6 \rangle = \langle 2 \rangle$. As you will see in Section 8.2, the answer is, "Not always." However, the statement is true for the ring $\mathbb{Z}$ (and a number of other rings as well). You will explore this in Exercise 8.17, and Section 8.2.

Exercises.

Exercise 8.14. Show that for any $n \in \mathbb{N}$, $n\mathbb{Z}$ is an ideal of $\mathbb{Z}$.

Exercise 8.15. Show that every ideal of $\mathbb{Z}$ has the form $n\mathbb{Z}$, for some $n \in \mathbb{N}$.

Exercise 8.16.
(a) Prove Lemma 8.4.
(b) More generally, prove that in *any* ring, $a \mid b$ if and only if $\langle b \rangle \subseteq \langle a \rangle$.

Exercise 8.17. In this exercise, we explore how $\langle r_1, r_2, \ldots, r_m \rangle$ behaves in $\mathbb{Z}$. Keep in mind that the results do not necessarily generalize to other rings.
(a) For the following values of $a, b \in \mathbb{Z}$, verify that $\langle a, b \rangle = \langle c \rangle$ for a certain $c \in \mathbb{Z}$.
  (i) $a = 3$, $b = 5$
  (ii) $a = 3$, $b = 6$
  (iii) $a = 4$, $b = 6$
(b) What is the relationship between a, b, and c in part (a)?
(c) Prove the conjecture you formed in part (b).

Exercise 8.18. Prove Theorem 8.7 (the Ideal Theorem).

Exercise 8.19. (a) Suppose R is a ring with unity, and A an ideal of R. Show that if $1_R \in A$, then $A = R$.
(b) Let q be an element of a ring with unity. Show that q has a multiplicative inverse if and only if $\langle q \rangle = \langle 1 \rangle$.

Exercise 8.20. Show that in any field $\mathbb{F}$, the only two distinct ideals are the zero ideal and $\mathbb{F}$ itself.

Exercise 8.21. Let R be a ring and A and I two ideals of R. Decide whether the following subsets of R are also ideals, and explain your reasoning:
(a) $A \cap I$
(b) $A \cup I$
(c) $A + I = \{x + y : x \in A, y \in I\}$
(d) $A \cdot I = \{xy : x \in A, y \in I\}$
(e) $AI = \left\{\sum_{i=1}^{n} x_i y_i : n \in \mathbb{N}, x_i \in A, y_i \in I\right\}$

Exercise 8.22. Let A, B be two ideals of a ring R. The definition of AB appears in Exercise 8.21.
(a) Show that $AB \subseteq A \cap B$.
(b) Show that sometimes $AB \neq A \cap B$; that is, find a ring R and ideals A, B such that $AB \neq A \cap B$.

Exercise 8.23. Let R be a ring with unity. Recall the polynomial ring $P = R[x_1, x_2, \ldots, x_n]$, whose ground ring is R (Section 7.3). Let

$$\langle f_1, \ldots, f_m \rangle = \{h_1 f_1 + \cdots + h_m f_m : h_1, h_2, \ldots, h_m \in P\}.$$

Example 8.5 showed that the set $A = \langle x^2 + y^2 - 4, xy - 1 \rangle$ was an ideal; Proposition 8.12 generalizes this to show that $\langle f_1, \ldots, f_m \rangle$ is an ideal of P. Show that the common roots of $f_1, f_2, \ldots, f_m$ are common roots of all polynomials in the ideal I.

Exercise 8.24. Let A be an ideal of a ring R. Define its **radical** to be

$$\sqrt{A} = \{r \in R : r^n \in A \; \exists r \in \mathbb{N}^+\}.$$

(a) Suppose $R = \mathbb{Z}$. Compute $\sqrt{A}$ for
  (i) $A = 2\mathbb{Z}$
  (ii) $A = 9\mathbb{Z}$
  (iii) $A = 12\mathbb{Z}$

(b) Suppose $R = \mathbb{Q}[x]$. Compute $\sqrt{A}$ for
  (i) $A = \langle x + 1 \rangle$
  (ii) $A = \langle x^2 + 2x + 1 \rangle$
  (iii) $A = \langle x^2 + 1 \rangle$

(c) Show that $\sqrt{A}$ is an ideal.

8.2: Principal Ideal Domains

In the previous section, we described ideals for commutative rings with identity that are generated by a finite set of elements, denoting them by $\langle r_1, \ldots, r_m \rangle$. An important subclass of these ideals consists of ideals generated by only one element.

Principal ideal domains

Definition 8.25. Let A be an ideal of a ring R. If $A = \langle a \rangle$ for some $a \in R$, then A is a **principal ideal**.

Notice that, by Proposition 8.12, we have $\langle a \rangle = \{ra : r \in R\}$.

Many ideals can be rewritten as principal ideals. For example, the zero ideal $\{0\} = \langle 0 \rangle$. If R has unity, we can write $R = \langle 1 \rangle$. On the other hand, not all ideals are principal; we will show that if $A = \langle x, y \rangle$ in the ring $\mathbb{C}[x, y]$, there is no $f \in \mathbb{C}[x, y]$ such that $A = \langle f \rangle$.

The following property of principal ideals is extremely useful.

Lemma 8.26. Let R be a ring with unity, and $a, b \in R$. There exists $q \in R$ such that $qa = b$ if and only if $\langle b \rangle \subseteq \langle a \rangle$. In addition, if R is an integral domain and $a, b \neq 0$, then the same q has a multiplicative inverse if and only if $\langle b \rangle = \langle a \rangle$.

Proof. The first assertion is just Exercise 8.16(b).

For the second, assume first that R is an integral domain, $a, b \neq 0$, and $qa = b$. We first show that if q has a multiplicative inverse, then $\langle b \rangle = \langle a \rangle$. So, assume that q has a multiplicative inverse. The first assertion gives us $\langle b \rangle \subseteq \langle a \rangle$. By definition, q has a multiplicative inverse r iff $rq = 1_R$. By substitution, $rb = r(qa) = a$. By absorption, $a \in \langle b \rangle$. Hence $\langle b \rangle \supseteq \langle a \rangle$. We already had $\langle b \rangle \subseteq \langle a \rangle$, so we conclude that $\langle b \rangle = \langle a \rangle$.

We have shown that if q has a multiplicative inverse, then $\langle b \rangle = \langle a \rangle$. It remains to show the converse; namely, that if $\langle b \rangle = \langle a \rangle$, then q has a multiplicative inverse. So, assume that $\langle b \rangle = \langle a \rangle$. By definition, there exist $r, q \in R$ such that $a = rb$ and $b = qa$. By substitution, $a = r(qa) = (rq)a$, so $a(1 - rq) = 0$. Since R is an integral domain and $a \neq 0$, $1 - rq = 0$. Rewritten as $rq = 1$, it shows that q does have a multiplicative inverse, r. □

Outside an integral domain, a could divide b with an element that has no multiplicative inverse, yet $\langle b \rangle = \langle a \rangle$. For example, in $\mathbb{Z}_6$, we have $[2] \cdot [2] = [4]$, but $\langle [2] \rangle = \{[0], [2], [4]\} = \langle [4] \rangle$.

There are rings in which all ideals are principal.

Definition 8.27. A **principal ideal domain** is an integral domain where every ideal can be written as a principal ideal.

Example 8.28. We claim that $\mathbb{Z}$ is a principal ideal domain, and we can prove this using a careful application of Exercise 8.17. Let A be any ideal of $\mathbb{Z}$. The zero ideal is $\langle 0 \rangle$, so assume that $A \neq \{0\}$. In this case, A contains at least one non-zero element; call it a_1. Without loss of generality, we may assume that $a_1 \in \mathbb{N}^+$ (if not, we could take $-a_1$ instead, since the definition of an ideal requires $-a_1 \in A$ as well).

Is $A = \langle a_1 \rangle$? If not, we can choose $b_1 \in A \setminus \langle a_1 \rangle$. Let $q_1, r_1 \in \mathbb{Z}$ be the quotient and remainder from division of b_1 by a_1; notice that $r_1 = b_1 - q_1 a_1 \in A$. Let $a_2 = \gcd(a_1, r_1)$. By the Extended Euclidean Algorithm, we can find $x, y \in \mathbb{Z}$ such that $xa_1 + yr_1 = a_2$. Since $a_1, r_1 \in A$, absorption and closure imply that $a_2 \in A$. In addition, $b_1 \notin \langle a_1 \rangle$, so Lemma 8.26 implies that $a_1 \nmid b_1$, so $r_1 \neq 0$, so $a_2 = \gcd(a_1, r_1) \neq 0$. We have $0 < a_2 \leq r_1 < a_1$. In fact, since $a_2 \mid a_1$, Exercise 8.17 tells us that $\langle a_1, a_2 \rangle = \langle a_2 \rangle$.

Is $A = \langle a_2 \rangle$? If not, we can repeat the previous process to find $b_2 \in A \setminus \langle a_2 \rangle$, divide b_2 by a_2 to obtain a nonzero remainder $r_2 \in A$, and compute $a_3 = \gcd(a_2, r_2)$. Reasoning similar to that above implies that $0 < a_3 < a_2 < a_1$ and $\langle a_1, a_2, a_3 \rangle = \langle a_3 \rangle$.

Continuing in this fashion, we see that as long as $A \neq \langle a_i \rangle$, we can find nonzero $b_i \in A \setminus \langle a_i \rangle$, a nonzero remainder $r_i \in A$ from division of b_i by a_i, and nonzero $a_{i+1} = \gcd(a_i, r_i)$, so that $0 < a_{i+1} < a_i$ and $\langle a_1, \ldots, a_{i-1} \rangle = \langle a_i \rangle$. This gives us a strictly decreasing chain of integers $a_1 > a_2 > \cdots > a_i > 0$. By Exercise 0.31, this cannot continue indefinitely. Let d be the final a_i computed; since we cannot compute anymore, $A = \langle a_1, \ldots, d \rangle$. As the greatest common divisor of the previously computed a_i, however, we have $a_1, a_2, \ldots \in \langle d \rangle$. Thus, $A = \langle d \rangle$.

Before moving on, let's take a moment to look at how the ideals are related, as well. Let $B_1 = \langle a_1 \rangle$, and $B_2 = \langle a_1, a_2 \rangle$. Lemma 8.26 implies that $B_1 \subsetneq B_2$. Likewise, if we set $B_3 = \langle a_1, a_2, a_3 \rangle = \langle a_3 \rangle$, then $B_2 \subsetneq B_3$. In fact, as long as $A \neq \langle a_i \rangle$, we generate an ascending sequence of ideals $B_1 \subsetneq B_2 \subsetneq \cdots$. In other words, another way of looking at this proof is that it *expands the principal ideal B_i until $B_i = A$*, by adding elements not in B_i. Rather amazingly, the argument above implies that this ascending chain of ideals must stabilize, at least in $\mathbb{Z}$. This property that an ascending chain of ideals must stabilize is one that some rings satisfy, but not all; we return to it in a moment.

We can extend the argument of Example 8.28 to more general rings.

Theorem 8.29. Every Euclidean domain is a principal ideal domain.

Proof. Let R be a Euclidean domain with respect to v, and let A be any non-zero ideal of R. Let $a_1 \in A$. As long as $A \neq \langle a_i \rangle$, do the following:

- find $b_i \in A \setminus \langle a_i \rangle$;
- let r_i be the remainder of dividing b_i by a_i;
  - notice $v(r_i) < v(a_i)$;
- compute a gcd a_{i+1} of a_i and r_i;
  - notice $v(a_{i+1}) \leq v(r_i) < v(a_i)$;
- this means $\langle a_i \rangle \subsetneq \langle a_{i+1} \rangle$; after all,
  - as a gcd, $a_{i+1} \mid a_i$, but
  - $a_i \nmid a_{i+1}$, lest $a_i \mid a_{i+1}$ imply $v(a_i) \leq v(a_{i+1}) < v(a_i)$
- hence, $\langle a_i \rangle \subsetneq \langle a_{i+1} \rangle$ and $v(a_{i+1}) < v(a_i)$.

By Exercise 0.31, the sequence $v(a_1) > v(a_2) > \cdots$ cannot continue indefinitely, which means that we cannot compute a_i's indefinitely. Let d be the final a_i computed. If $A \neq \langle d \rangle$, we could certainly compute another a_i, so it must be that $A = \langle a_i \rangle$. □

Not all integral domains are principal ideal domains; you will show in the exercises that for any field $\mathbb{F}$ and its polynomial ring $\mathbb{F}[x,y]$, the ideal $\langle x, y \rangle$ is not principal.

Noetherian rings and the Ascending Chain Condition

For now, though, we will turn to a phenomenon that appeared in Example 8.28 and Theorem 8.29. In each case, we built a chain of ideals

$$\langle a_1 \rangle \subsetneq \langle a_2 \rangle \subsetneq \langle a_3 \rangle \subsetneq \cdots$$

and were able to show that the procedure we used to find the a_i must eventually terminate.

This property is very useful for a ring. In both Example 8.28 and Theorem 8.29, we relied on the well-ordering of $\mathbb{N}$, but that is not always available to us. So the property might be useful in other settings, even in cases where ideals aren't guaranteed to be principal. For example, eventually we will show that $\mathbb{F}[x,y]$ satisfies this property.

Definition 8.30. Let R be a ring. If for every **ascending chain of ideals** $A_1 \subseteq A_2 \subseteq \cdots$ we can find an integer k such that $A_k = A_{k+1} = \cdots$, then R satisfies the **Ascending Chain Condition**.

Remark 8.31. Another name for a ring that satisfies the Ascending Chain Condition is a **Noetherian ring**, after the German mathematician Emmy Noether.

Theorem 8.32. Each of the following holds.

(A) Every principal ideal domain satisfies the Ascending Chain Condition.
(B) Any field $\mathbb{F}$ satisfies the Ascending Chain Condition.
(C) If a ring R satisfies the Ascending Chain Condition, so does $R[x]$.
(D) If a ring R satisfies the Ascending Chain Condition, so does $R[x_1, x_2, \ldots, x_n]$.

Proof. (A) Let R be a principal ideal domain, and let $A_1 \subseteq A_2 \subseteq \cdots$ be an ascending chain of ideals in R. Let $B = \bigcup_{i=1}^{\infty} A_i$. By Exercise 8.37, B is an ideal. Since R is a principal ideal domain, $B = \langle b \rangle$ for some $b \in R$. By definition of a union, $b \in A_i$ for some $i \in \mathbb{N}$. The definition of an ideal now implies that $rb \in A_i$ for all $r \in R$; since $\langle b \rangle = \{rb : r \in R\}$, we infer that $\langle b \rangle \subseteq A_i$. By substitution, $B \subseteq A_i$. By definition of union, we also have $A_i \subseteq B$. Hence $A_i = B$, and a similar argument shows that $A_j = B$ for all $j \geq i$. In other words, the chain of ideals stabilizes at A_i. Since the chain was arbitrary, every ascending chain of ideals in R stabilizes, so R satisfies the ascending chain condition.

(B) By Exercise 7.69, any field $\mathbb{F}$ is a Euclidean domain, so this follows from (A) and Theorem 8.29. However, it's instructive to look at it from the point of view of a field as well. Recall from Exercise 8.20 that a field has only two distinct ideals: the zero ideal, and the field itself. Hence, any ascending chain of ideals stabilizes either at the zero ideal or at $\mathbb{F}$ itself.

(C) Assume that R satisfies the Ascending Chain Condition. The argument is based on two claims.

Claim 1: Every ideal of $R[x]$ is finitely generated. Let A be any ideal of $R[x]$, and choose $f_1, f_2, \ldots \in A$ in the following way:

- Let $B_0 = \{0\}$, and $k = 0$.
- While $A \neq \langle B_k \rangle$:
  - Let $S_k = \{\deg f : f \in A \setminus \langle B_k \rangle\}$. Since $S_k \subseteq \mathbb{N}$, it has a least element; call it d_k.
  - Let $f_k \in A \setminus \langle B_k \rangle$ be any polynomial of degree d_k. Notice that $f_k \in A \setminus \langle B_k \rangle$ implies that $\langle B_k \rangle \subsetneq \langle B_k \cup \{f_k\} \rangle$.
  - Let $B_{k+1} = B_k \cup \{f_k\}$, and add 1 to k.

Does this process terminate? We built $\langle f_1 \rangle \subsetneq \langle f_1, f_2 \rangle \subsetneq \cdots$ as an ascending chain of ideals. Denote the leading coefficient of f_k by a_k and let $C_k = \langle a_1, a_2, \ldots, a_k \rangle$. Since R satisfies the Ascending Chain Condition, the ascending chain of ideals $C_1 \subseteq C_2 \subseteq \cdots$ stabilizes for some $m \in \mathbb{N}$.

We claim that the chain $\langle B_0 \rangle \subsetneq \langle B_1 \rangle \subsetneq \langle B_2 \rangle \subsetneq \cdots$ has also stabilized at m; that is, we cannot find $f_{m+1} \in A \setminus \langle B_m \rangle$. By way of contradiction, suppose we can find f_{m+1} of minimal degree in $A \setminus \langle B_{m+1} \rangle$. By hypothesis, the chain of ideals C_k has stabilized, so $C_m = C_{m+1}$. Thus, $a_{m+1} \in C_{m+1} = C_m$. That means we can write $a_{m+1} = b_1 a_1 + \cdots + b_m a_m$ for some $b_1, \ldots, b_m \in R$. Write $d_i = \deg_x f_i$, and let

$$p = b_1 f_1 x^{d_{m+1} - d_1} + \cdots + b_m f_m x^{d_{m+1} - d_m}.$$

We chose each f_i to be of minimal degree, so for each i, we have $d_i \leq d_{m+1}$. Thus, $d_{m+1} - d_i \in \mathbb{N}$, and $p \in R[x]$. Moreover, we have set up the sum and products so that $\operatorname{lt}\left(b_i f_i x^{d_{m+1} - d_i}\right) =$

$b_i\left(a_i x^{d_i}\right) x^{d_{m+1}-d_i} = b_i a_i x^{d_{m+1}}$. This implies that the leading term of p is

$$(b_1 a_1 + \cdots + b_m a_m)\, x^{d_{m+1}} = a_{m+1} x^{d_{m+1}}.$$

Let $r = f_{m+1} - p$. Since $\operatorname{lt}(f_{m+1}) = \operatorname{lt}(p)$, the leading terms cancel, and $\deg r < \deg f_{m+1}$. By construction, $p \in B_{m+1}$. If $r \in \langle B_{m+1}\rangle$, we could rewrite $r = f_{m+1} - p$ as $f_{m+1} = r + p$, which would imply that $f_{m+1} \in \langle B_{m+1}\rangle$. This contradicts the choice of $f_{m+1} \in A \setminus \langle B_{m+1}\rangle$. Thus, $r \notin \langle B_{m+1}\rangle$. Since f_{m+1} and p are both in A, we have $r \in A \setminus \langle B_{m+1}\rangle$. However, $\deg r < \deg f_{m+1}$; this contradicts the choice of f_{m+1} as a polynomial with minimal degree in $A \setminus \langle B_{m+1}\rangle$.

The only unfounded assumption was that we could find $f_{m+1} \in A \setminus \langle B_m\rangle$. Apparently, we cannot do so, and the process of choosing elements of $A \setminus \langle B_i\rangle$ must terminate at $i = m$. Since it does not terminate unless $A = \langle B_m\rangle$, we conclude that $A = \langle B_m\rangle = \langle f_1, \ldots, f_m\rangle$. In other words, A is finitely generated.

Claim 2: Every ascending chain of ideals in R eventually stabilizes. Let $I_1 \subseteq I_2 \subseteq \cdots$ be an ascending chain. By Exercise 8.37, the set $I = \cup_{i=1}^{\infty} I_i$ is also an ideal. By Claim 1, I is finitely generated; let $f_1, \ldots, f_m \in I$ such that $I = \langle f_1, \ldots, f_m\rangle$. By definition of union, there exist $k_1, \ldots, k_m \in \mathbb{N}^+$ such that $f_j \in I_{k_j}$. By definition of subset, $f_j \in I_\ell$ for all $\ell > k_j$. Let $\ell \geq \max\{k_1, \ldots, k_m\}$; then $f_1, \ldots, f_m \in I_\ell$, so

$$I = \langle f_1, \ldots, f_m\rangle \subseteq I_\ell \subseteq I,$$

which implies equality. Thus, the chain stabilizes at $\max\{k_1, \ldots, k_m\}$.

(D) follows from (C) by induction on the number of variables n: use R to show $R[x_1]$ satisfies the Ascending Chain Condition; use $R[x_1]$ to show that $R[x_1, x_2] = (R[x_1])[x_2]$ satisfies the Ascending Chain Condition; etc. □

Corollary 8.33 (Hilbert Basis Theorem)**.** For any Noetherian ring R, $R[x_1, x_2, \ldots, x_n]$ satisfies the Ascending Chain Condition. In particular, this is true when R is a field $\mathbb{F}$. Thus, for any ideal I of $\mathbb{F}[x_1, \ldots, x_n]$, we can find $f_1, \ldots, f_m \in I$ such that $I = \langle f_1, \ldots, f_m\rangle$.

Proof. Apply (B) and (D) of Theorem 8.32. □

Exercises

Exercise 8.34. Let $d \in \mathbb{Z}$. Explain why:
(a) $d\mathbb{Z}$ is not a principal ideal domain, but
(b) every ideal is still principal.

Exercise 8.35. Is $\mathbb{F}[x]$ a principal ideal domain for every field $\mathbb{F}$? What about $R[x]$ for every ring R?

Exercise 8.36. Let $\mathbb{F}$ be any field, and consider the polynomial ring $\mathbb{F}[x, y]$. Explain why $\langle x, y\rangle$ cannot be principal.

Exercise 8.37. Let R be a ring and $I_1 \subseteq I_2 \subseteq \cdots$ an ascending chain of ideals. Show that $\mathcal{I} = \bigcup_{i=1}^{\infty} I_i$ is itself an ideal.

Exercise 8.38. Show that $\mathbb{Z}$ satisfies the Ascending Chain Condition.

Exercise 8.39. Let R be a ring and $a, b \in R$.

(a) Show that if R has unity, $\langle a \rangle \langle b \rangle = \langle ab \rangle$.
(b) Show that if R does not have unity, it can happen that $\langle a \rangle \langle b \rangle \neq \langle ab \rangle$.

8.3: Cosets and Quotient Rings

Recall that in group theory, we could use cosets of a subgroup to create equivalence classes in a group. We want to do the same thing for rings. Since a ring has two operations, we need to decide which one we ought to use to do this. The decision isn't very hard; as we saw in Section 8.1, some subrings absorb multiplication — in particular, *ideals* absorb multiplication — so we cannot expect to create cosets using that operation. We will have to try with addition alone.

Definition 8.40. Let R be a ring and S a subring of R. For every $r \in R$, denote

$$r + S := \{r + s : s \in S\},$$

called a **coset**. Then define

$$R/S := \{r + S : r \in R\}.$$

Since a sub*ring* is always a sub*group* under addition — in fact, it is a *normal* subgroup — and subgroups partition a group, we can immediately identify three properties of the cosets of a subring:

- they partition the ring as an additive group;
- they create a set of equivalence classes of the additive group;
- they create a quotient *group* under *addition;* and
- coset equality in rings follows the rules of coset equality in groups, listed in Lemma 3.29 on page 102.

Do they also create a quotient *ring?* In fact, they might not!

The necessity of ideals

The absorption property plays a critical role in guaranteeing that multiplication is well-defined.

Example 8.41. Let $R = \mathbb{Z}[x]$, and S the smallest subring of R that contains $x^2 - 1$. It is not hard to see that $S = \left\{\sum_{i=1}^{n} a_i \left(x^2 - 1\right)^{p_1} : n, p_i \in \mathbb{N}^+, a_i \in \mathbb{Z}\right\}$.

Let $X = x + S$ and $Y = 1 + S$. Notice that we can write $Y = x^2 + S$ as well, because $x^2 - 1 \in S$. However, the value of XY is not equal for both representations of Y! The first gives us

$$XY = (x + S)(1 + S) = x + S,$$

while the second gives us

$$XY = (x+S)\left(x^2+S\right) = x^3+S,$$

and $x^3 - x$ does not have the form necessary for members of S. Thus, R/S is not a ring.

We will see that the absorption property of ideals *does* guarantee that the multiplication of cosets is well-defined, which opens the door to creating quotient rings.

Lemma 8.42. The "natural" addition and multiplication of cosets is well-defined whenever a subring is an ideal.

Proof. First we show that the operations are well-defined. Let $X, Y \in R/A$ and $w, x, y, z \in R$. such that $w+A = x+A = X$ and $y+A = z+A = Y$.

Is addition well-defined? The definition of the operation tells us both $X+Y = (x+y)+A$ and $X+Y = (w+z)+A$. By the hypothesis that $x+A = w+A$ and $y+A = z+A$, Lemma 3.29 implies that $x-w \in A$ and $y-z \in A$. By closure, $(x-w)+(y-z) \in A$. Using the properties of a ring,

$$(x+y)-(w+z) = (x-w)+(y-z) \in A.$$

Again from Lemma 3.29, $(x+y)+A = (w+z)+A$, so, by definition,

$$\begin{aligned}(x+A)+(y+A) &= (x+y)+A \\ &= (w+z)+A = (w+A)+(z+A).\end{aligned}$$

It does not matter, therefore, which representations we use for X and Y; the sum $X+Y$ has the same value, so addition in R/A is well-defined.

Is multiplication well-defined? Observe that $XY = (x+A)(y+A) = xy+A$. As explained above, $x-w \in A$ and $y-z \in A$. Let $a, \widehat{a} \in A$ such that $x-w = a$ and $y-z = \widehat{a}$; **from the absorption property of an ideal**, $ay \in A$, so

$$\begin{aligned}xy-wz &= (xy-xz)+(xz-wz) \\ &= x(y-z)+(x-w)z \\ &= x\widehat{a}+az \in A.\end{aligned}$$

Again from Lemma 3.29, $xy+A = wz+A$, and by definition

$$(x+A)(y+A) = xy+A = wz+A = (w+A)(z+A).$$

It does not matter, therefore, what representations we use for X and Y; the product XY has the same value, so multiplication in R/A is well-defined. □

Using an ideal to create a new ring

We now generalize the notion of *quotient groups* to rings, and prove some interesting properties of certain quotient groups that help explain various phenomena we observed in both group theory and ring theory.

Theorem 8.43. Let R be a ring, and A an ideal. Define addition and multiplication for R/A in the "natural" way: for all $X, Y \in R/A$ denoted as $x+A, y+A$ for some $x, y \in R$,

$$X+Y = (x+y)+A$$
$$XY = (xy)+A.$$

The set R/A is a ring under these operations, called the **quotient ring**.

Example 8.44. Recall that $\mathbb{Z}$ is a ring, and $d\mathbb{Z}$ is an ideal for any $d \in \mathbb{Z}$. Thus, $\mathbb{Z}/d\mathbb{Z}$ is a quotient ring, and $3+d\mathbb{Z}$ is a coset.

Example 8.45. Recall that $\mathbb{Z}_2[x]$ is a ring. Let $A = \langle x^2+1 \rangle$. We construct the addition and multplication tables for $\mathbb{Z}_2[x]/A$.

First, recall that $\mathbb{Z}_2[x]$ is a Euclidean domain, so we can perform division, so any polynomial can be written as $p = q(x^2+1)+r$, where $\deg r < \deg(x^2+1) = 2$. By absorption, $p-r = q(x^2+1) \in A$, so coset equality implies $[p] = [r]$. No remainder has degree more than 1, so every element of $\mathbb{Z}_2[x]$ has the form $[ax+b]=(ax+b)+\mathbb{Z}_2[x]$. That means there are only four elements of the quotient ring:

$$[0], [1], [x], [x+1].$$

Superficially, then, we get the following tables.

$+$	0	1	x	$x+1$
0	0	1	x	$x+1$
1	1	0	$x+1$	x
x	x	$x+1$	0	1
$x+1$	$x+1$	x	1	0

$\times$	0	1	x	$x+1$
0	0	0	0	0
1	0	1	x	$x+1$
x	0	x	x^2	x^2+x
$x+1$	0	$x+1$	x^2+x	x^2+1

(Notice that $2x = 0$ in $\mathbb{Z}_2$, which is why $(x+1)^2 = x^2+1$.)

While the multiplication table is accurate, it is unsatisfactory, because every element of the table can be written as a *linear* polynomial. Applying division again, we get

$$x^2 = 1\cdot(x^2+1)+1, \qquad x^2+1 = 1\cdot(x^2+1)+0, \qquad x^2+x = 1\cdot(x^2+1)+(x+1).$$

Thus, the multiplication table can be written *in canonical form* as follows.

$\times$	0	1	x	$x+1$
0	0	0	0	0
1	0	1	x	$x+1$
x	0	x	1	$x+1$
$x+1$	0	$x+1$	$x+1$	0

Notation 8.46. When we consider elements of $X \in R/A$, we refer to the "usual representation" of X as $x+A$ for appropriate $x \in R$; that is, "big" X is represented by "little" x. Likewise, if $X = 3+d\mathbb{Z}$, we often write $x = [3]$ or even $x = 3$.

You may remember that, when working in quotient rings, we made heavy use of Lemma 3.29 on page 102. You will see that here, too.

Proof of Theorem 8.43. We have already shown that addition and multiplication are well-defined in R/A, so we turn to showing that R/A is a ring. First we show the properties of a group under addition:

closure: Let $X, Y \in R/A$, with the usual representation. By substitution, $X+Y = (x+y)+A$. Since R, a ring, is closed under addition, $x+y \in R$. Thus $X+Y \in R/A$.

associative: Let $X, Y, Z \in R/A$, with the usual representation. Applying substitution and the associative property of R, we have

$$\begin{aligned}(X+Y)+Z &= ((x+y)+A)+(z+A)\\ &= ((x+y)+z)+A\\ &= (x+(y+z))+A\\ &= (x+A)+((y+z)+A)\\ &= X+(Y+Z).\end{aligned}$$

identity: We claim that $A = 0+A$ is itself the identity of R/A; that is, $A = 0_{R/A}$. Let $X \in R/A$ with the usual representation. Indeed, substitution and the additive identity of R demonstrate this:

$$\begin{aligned}X+A &= (x+A)+(0+A)\\ &= (x+0)+A\\ &= x+A\\ &= X.\end{aligned}$$

inverse: Let $X \in R/A$ with the usual representation. We claim that $-x+A$ is the additive inverse of X. Indeed,

$$\begin{aligned}X+(-x+A) &= (x+(-x))+A\\ &= 0+A\\ &= A\\ &= 0_{R/A}.\end{aligned}$$

Hence $-x+A$ is the additive inverse of X.

Now we show that R/A satisfies the ring properties. Each property falls back on the corresponding property of R.

closure: Let $X, Y \in R/A$ with the usual representation. By definition and closure in R,

$$\begin{aligned}XY &= (x+A)(y+A)\\ &= (xy)+A\\ &\in R/A.\end{aligned}$$

associative: Let $X, Y, Z \in R/A$ with the usual representation. By definition and the associative

property in R,

$$\begin{aligned}(XY)Z &= ((xy)+A)(z+A)\\ &= ((xy)z)+A\\ &= (x(yz))+A\\ &= (x+A)((yz)+A)\\ &= X(YZ).\end{aligned}$$

distributive: Let $X, Y, Z \in R/A$ with the usual representation. By definition and the distributive property in R,

$$\begin{aligned}X(Y+Z) &= (x+A)((y+z)+A)\\ &= (x(y+z))+A\\ &= (xy+xz)+A\\ &= ((xy)+A)+((xz)+A)\\ &= XY+XZ.\end{aligned}$$

Hence R/A is a ring. □

We conclude with an obvious property of quotient rings.

Proposition 8.47. If R is a ring with unity, then R/A is also a ring with unity, which is $1_R + A$.

Proof. You do it! See Exercise 8.51. □

In Section 3.5 we showed that one could define a group using the quotient group $\mathbb{Z}_n = \mathbb{Z}/n\mathbb{Z}$. Since $\mathbb{Z}$ is a ring and $n\mathbb{Z}$ is an ideal of $\mathbb{Z}$ by Exercise 8.14, it follows that $\mathbb{Z}_n$ is also a ring. Of course, you had already argued this in Exercise 7.14.

Exercises.

Exercise 8.48. Compute addition and multiplication tables for
(a) $\mathbb{Z}_2[x]/\langle x\rangle$;
(b) $\mathbb{Z}_2[x]/\langle x^2+x\rangle$;
(c) $\mathbb{Z}_2[x,y]/\langle x^2, y^2\rangle$.

Exercise 8.49. The example at the beginning of the section came from the ring $\mathbb{Z}[x]$. Show that in the ring of integers, *any* subring creates a quotient ring.

Exercise 8.50. Let $R = \mathbb{Z}_5[x]$ and $I = \langle x^2+2x+2\rangle$.
(a) Explain why $(x^2+x+3)+I = (4x+1)+I$.
(b) Find a factorization of x^2+2x+2 in R.
(c) Find two non-zero elements of R/I whose product is the zero element of R/I.
(d) Explain why R/I is, therefore, not an integral domain, and, therefore, not a field.

Exercise 8.51. Prove Proposition 8.47.

8.4: When is a quotient ring an integral domain or a field?

You found in Exercise 7.31 that $\mathbb{Z}_n$ is not, in general, an integral domain, let alone a field. The curious thing is that we started with an integral domain $\mathbb{Z}$, computed a quotient ring by an ideal $n\mathbb{Z}$ that satisfies the zero product property, yet we *still* didn't end up with an integral domain! Why did this happen? We found that it occurred when n was not irreducible, which also means it was not prime.

We can view this as a relationship not just of divisibility, but of ideals. From the definition of an irreducible integer, we know that n is irreducible if its only divisors are ± 1 and $\pm n$. Lemma 8.26 translates this into the language of ideals as this remarkable statement:

The only ideals "larger" than $\langle n\rangle$ are $\mathbb{Z}$ (of course) and $\langle n\rangle$ itself.

In other words, the ideal generated by an irreducible number is the "largest" sort of proper ideal in $\mathbb{Z}$. We ought to generalize that.

On the other hand, while the notions of "prime" and "irreducible" are equivalent in the integers, they may not mean the same thing in all rings. For example, in $\mathbb{Z}_6$,

- 2 is *not* irreducible, since $2 = 20 = 4 \cdot 5$, but
- 2 seems to be "prime", since from the 36 products possible in $\mathbb{Z}_6$, the only ones where 2 does not divide one of the factors are

$$1 \times 1,\ 1 \times 3,\ 1 \times 5,\ 3 \times 3,\ 3 \times 5,\ 5 \times 5,$$

  and 2 divides none of those products, either.

This observation raises a number of questions, but looking at them carefully would lead us astray for now, so we delay them until Chapter 10; for now, however, we prefer to focus on ideals. Recall that in algebra, n is prime if any time $n \mid ab$, then $n \mid a$ or $n \mid b$. Lemma 8.26 translates this into the language of ideals as this equally remarkable statement:

If $\langle n\rangle$ contains $\langle ab\rangle$, then it must contain $\langle a\rangle$ or $\langle b\rangle$.

Maximal and prime ideals

Let R be a ring.

Definition 8.52. A proper ideal A of R is a **maximal ideal** if no other proper ideal of R contains A.

Another way of expressing that A is maximal is the following: for any other ideal I of R, $A \subseteq I$ implies that $A = I$ or $I = R$.

Example 8.53. In Exercise 8.15 you showed that all ideals of $\mathbb{Z}$ have the form $n\mathbb{Z}$ for some $n \in \mathbb{Z}$. Are any of these (or all of them) maximal ideals?

Let $n \in \mathbb{Z}$ and suppose that $n\mathbb{Z}$ is maximal. Certainly $n \neq 0$, since $2\mathbb{Z} \not\subseteq \{0\}$. We claim that $|n|$ is irreducible; in other words, n is divisible only by $\pm 1, \pm n$. To see this, recall Lemma 8.4: $m \in \mathbb{Z}$ is a divisor of n iff $n\mathbb{Z} \subseteq m\mathbb{Z}$. Since $n\mathbb{Z}$ is maximal, either $m\mathbb{Z} = \mathbb{Z}$ or $m\mathbb{Z} = n\mathbb{Z}$. In the first case, $m = \pm 1$; in the second case, $m = \pm n$. Hence $|n|$ is irreducible.

For prime ideals, you need to recall from Exercise 8.21 that for any two ideals A, B of R, AB is also an ideal.

Definition 8.54. A proper ideal P of R is a **prime ideal** if for every two ideals A, B of R we know that

$$\text{if } AB \subseteq P \text{ then } A \subseteq P \text{ or } B \subseteq P.$$

Definition 8.54 might remind you of our definition of prime integers from page 6.30. Indeed, the two are connected.

Example 8.55. Let $n \in \mathbb{Z}$ be a prime integer. Let $a, b \in \mathbb{Z}$ such that $p \mid ab$. Hence $p \mid a$ or $p \mid b$. Suppose that $p \mid a$.

Let's turn our attention to the corresponding ideals. Since $p \mid ab$, Lemma 8.4 tells us that $(ab)\mathbb{Z} \subseteq p\mathbb{Z}$. It is routine to show that $(ab)\mathbb{Z} = (a\mathbb{Z})(b\mathbb{Z})$, but in case you think otherwise, it's also Exercise 8.39. Put $A = a\mathbb{Z}$, $B = b\mathbb{Z}$, and $P = p\mathbb{Z}$; thus $AB \subseteq P$.

Recall that $p \mid a$; applying Lemma 8.4 again, we have $A = a\mathbb{Z} \subseteq p\mathbb{Z} = P$.

Conversely, if n is not prime, $n\mathbb{Z}$ is not a prime ideal: for example, $6\mathbb{Z}$ is not a prime ideal because $(2\mathbb{Z})(3\mathbb{Z}) \subseteq 6\mathbb{Z}$ but by Lemma 8.16 neither $2\mathbb{Z} \subseteq 6\mathbb{Z}$ nor $3\mathbb{Z} \subseteq 6\mathbb{Z}$. This can be generalized easily to all integers that are not prime: see Exercise 8.63.

Let's summarize our examples. We found in Example 8.53 that an ideal in $\mathbb{Z}$ is maximal iff it is generated by a prime integer, and in Example 8.55 we argued that an ideal is prime iff it is generated by a prime integer. We learned in Theorem 6.32 that an integer is prime if and only if it is irreducible. Thus, an ideal is maximal if and only if it is prime — in the ring of integers, anyway.

What about other rings? Showing a maximal ideal is prime doesn't require too many additional constraints.

Theorem 8.56. Let R be a ring. If R has unity, then every maximal ideal is prime.

Proof. Assume that R has unity. We want to show that every maximal ideal is prime, so let M be a maximal ideal of R. Let A, B be any two ideals of R such that $AB \subseteq M$. We claim that $A \subseteq M$ or $B \subseteq M$.

If $A \subseteq M$, then we are done, so assume that $A \nsubseteq M$. Recall from Exercise 8.21 that $A + M$ is also an ideal. In addition, it should be clear that $M \subseteq A + M$. (If it isn't clear, try it. It really isn't hard.) Since M is maximal, $A + M = M$ or $A + M = R$. Which is it?

We claim that $A + M = R$. To see why, observe that if $A + M = M$, then for any $a \in A$ and any $m \in M$ we could find $m' \in M$ such that $a + m = m'$. We can rewrite this as $a = m' - m$; closure tells us $m' - m \in M$, and substitution gives us $a \in M$. But a was arbitrary, implying that $A \subseteq M$, contradicting the hypothesis that $A \subsetneq M$. Thus, $A + M \neq M$, which means $A + M = R$.

Since R has unity, $1_R \in R = A + M$, so there exist $a \in A$, $m \in M$ such that

$$1_R = a + m. \tag{27}$$

Let $b \in B$. Multiply both sides of (27) by b; we have

$$\begin{aligned} 1_R \cdot b &= (a + m)\, b \\ b &= ab + mb. \end{aligned}$$

Recall that $AB \subseteq M$; since $ab \in AB$, $ab \in M$. Likewise, absorption implies that $mb \in M$. Closure of addition implies that $ab + mb \in M$. Substitution implies that $b \in M$. Since b was arbitrary in B, $B \subseteq M$.

We assumed that $AB \subseteq M$, and found that $A \subseteq M$ or $B \subseteq M$. Thus, M is prime. □

Is the requirement that the ring have unity that important? Yes, even in simple rings.

Theorem 8.57. If R is a ring without unity, then maximal ideals might not be prime.

Proof. The proof is by counterexample: we use $2\mathbb{Z}$, a ring without unity. We claim that $4\mathbb{Z}$ is a maximal ideal of $R = 2\mathbb{Z}$ that is not prime:

closed under subtraction? Let $x, y \in 4\mathbb{Z}$. By definition of $4\mathbb{Z}$, $x = 4a$ and $y = 4b$ for some $a, b \in \mathbb{Z}$. Using the distributive property and substitution, we have $x - y = 4a - 4b = 4(a - b) \in 4\mathbb{Z}$.

absorbs multiplication? Let $x \in 4\mathbb{Z}$ and $r \in 2\mathbb{Z}$. By definition of $4\mathbb{Z}$, $x = 4q$ for some $q \in \mathbb{Z}$. By substitution, the associative property, and the commutative property of integer multiplication, $rx = 4(rq) \in 4\mathbb{Z}$.

maximal? Let A be any ideal of $2\mathbb{Z}$ such that $4\mathbb{Z} \subseteq A$. Choose $a \in \mathbb{Z}$ such that $A = a\mathbb{Z}$. (We can do this thanks to Exercise 8.34.) By Lemma 8.4, $a \mid 4$, so the only possible values of a are ± 1, ± 2, and ± 4. Certainly $a \neq \pm 1$; after all, $\pm 1 \notin R$. If $a = \pm 4$, then $A = 4\mathbb{Z}$. If $a = \pm 2$, then $A = R$. We took an arbitrary ideal A such that $4\mathbb{Z} \subseteq A$, and found that $A = 4\mathbb{Z}$ or $A = 2\mathbb{Z}$, the entire ring. Hence, $4\mathbb{Z}$ is maximal.

prime? An easy counterexample does the trick: $(2\mathbb{Z})(2\mathbb{Z}) = 4\mathbb{Z}$, but $2\mathbb{Z} \not\subseteq 4\mathbb{Z}$.

□

The situation with prime ideals is less… well, to be cute about it, "less than ideal".

Theorem 8.58. A prime ideal is not necessarily maximal, even in a ring with unity.

Proof. Recall that $R = \mathbb{C}[x, y]$ is a ring with unity, and that $I = \langle x \rangle$ is an ideal of R.

We claim that I is a prime ideal of R. Let A, B be ideals of R such that $AB \subseteq I$. If $A \subseteq I$, then we are done, so suppose that $A \not\subseteq I$. We need to show that $B \subseteq I$. Let $a \in A \backslash I$. For any $b \in B$, $ab \in AB \subseteq I = \langle x \rangle$, so $ab \in \langle x \rangle$. This implies that $x \mid ab$; let $q \in R$ such that $qx = ab$. Write $a = f \cdot x + a'$ and $b = g \cdot x + b'$ where $a', b' \in R \backslash I$; that is, a' and b' are polynomials with *no* terms that are multiples of x. By substitution,

$$\begin{aligned} ab &= (f \cdot x + a')(g \cdot x + b') \\ qx &= (f \cdot x) \cdot (g \cdot x) + a' \cdot (g \cdot x) \\ &\quad + b' \cdot (f \cdot x) + a' \cdot b' \\ (q - fg - a'g - b'f)\, x &= a'b'. \end{aligned}$$

Hence $a'b' \in \langle x \rangle$. However, no term of a' or b' is a multiple of x, so no term of $a'b'$ is a multiple of x. The only element of $\langle x \rangle$ that satisfies this property is 0. Hence $a'b' = 0$, which by the zero product property of complex numbers implies that $a' = 0$ or $b' = 0$.

Which is it? If $a' = 0$, then $a = f \cdot x + 0 \in \langle x \rangle = I$, which contradicts the assumption that $a \in A \backslash I$. Hence $a' \neq 0$, implying that $b' = 0$, so $b = gx + 0 \in \langle x \rangle = I$. Since b is arbitrary, this holds for all $b \in B$; that is, $B \subseteq I$.

We took two arbitrary ideals such that $AB \subseteq I$ and showed that $A \subseteq I$ or $B \subseteq I$; hence $I = \langle x \rangle$ is prime. However, I is not maximal, since

- $y \notin \langle x \rangle$, implying that $\langle x \rangle \subsetneq \langle x, y \rangle$; and
- $1 \notin \langle x, y \rangle$, implying that $\langle x, y \rangle \notin \mathbb{C}[x, y]$.

□

So prime and maximal ideals need not be equivalent. In Chapter 10, we will find conditions on a ring that ensure that prime and maximal ideals are equivalent.

A criterion that determines when a quotient ring is an integral domain or a field

We can now answer the question that opened this section.

Theorem 8.59. If R is a ring with unity and M is a maximal ideal of R, then R/M is a field. The converse is also true.

Proof. ($\Rightarrow$) Assume that R is a ring with unity and M is a maximal ideal of R. Let $X \in R/M$ and assume that $X \neq M$; that is, X is non-zero. Since $X \neq M$, $X = x + M$ for some $x \notin M$. By Exercise 8.21, $\langle x \rangle + M$ is also an ideal. Since $x \notin M$, we know that $M \subsetneq \langle x \rangle + M$. Since M is a maximal ideal, $M \subsetneq \langle x \rangle + M = R$. Since R is a ring with unity, $1 \in R$ by definition. Substitution implies that $1 \in \langle x \rangle + M$, so there exist $h \in R$, $m \in M$ such that $1 = hx + m$. Rewrite this as $1 - hx = m \in M$; by Lemma 3.29,

$$1 + M = hx + M = (h + M)(x + M).$$

In other words, $h + M$ is a multiplicative inverse of $X = x + M$ in R/M. Since X was an arbitrary non-zero element of R/M, every element of R/M has a multiplicative inverse, and R/M is a field.

($\Leftarrow$) For the converse, assume that R/M is a field. We want to show that M is maximal, so let N be any ideal of R such that $M \subseteq N \subseteq R$. If $M = N$, then we are done, so assume that $M \neq N$. We want to show that $N = R$. Let $x \in N \backslash M$; then $x + M \neq M$, and since R/M is a field, $x + M$ has a multiplicative inverse; call it $Y = y + M$. That is,

$$1 + M = (x + M)(y + M) = (xy) + M,$$

which by Lemma 3.29 implies that $xy - 1 \in M$. Let $m \in M$ such that $xy - 1 = m$; then $1 = xy - m$. Now, $xy \in N$ by absorption, and $m \in N$ by inclusion. (After all, $x \in N$ and $m \in M \subsetneq N$.) Closure of the subring N implies that $1 = xy - m \in N$, and Exercise 8.19 implies that $N = R$. Since N was an arbitrary ideal that contained M properly, M is maximal. □

A similar property holds true for prime ideals.

Theorem 8.60. If R is a ring with unity and P is a prime ideal of R, then R/P is an integral domain. The converse is also true.

Proof. ($\Rightarrow$) Assume that R is a ring with unity and P is a prime ideal of R. Let $X, Y \in R/P$ with the usual representation, and assume that $XY = 0_{R/P} = P$. By definition of the operation, $XY = (xy) + P$; by Lemma 3.29, $xy \in P$. We claim that this implies that $x \in P$ or $y \in P$.

Assume to the contrary that $x, y \notin P$. For any $z \in \langle x \rangle \langle y \rangle$, we have $z = \sum_{k=1}^{m} (h_k x)(q_k y)$ for an appropriate choice of $m \in \mathbb{N}^+$ and $h_k, q_k \in R$. Recall that R is commutative, which means $z = xy \sum (h_k q_k)$. We determined above that $xy \in P$, so by absorption, $z \in P$. Since z was arbitrary in $\langle x \rangle \langle y \rangle$, we conclude that $\langle x \rangle \langle y \rangle \subseteq P$. Now P is a prime ideal, so $\langle x \rangle \subseteq P$ or $\langle y \rangle \subseteq P$; without loss of generality, $\langle x \rangle \subseteq P$. Since R has unity, $x \in \langle x \rangle$, and thus $x \in P$. Lemma 3.29 now implies that $x + P = P$. Thus $X = 0_{R/P}$.

We took two arbitrary elements of R/P, and showed that if their product was the zero element of R/P, then one of those elements had to be P, the zero element of R/P. That is, R/P is an integral domain.

($\Leftarrow$) For the converse, assume that R/P is an integral domain. Let A, B be two ideals of R, and assume that $AB \subseteq P$. Assume that $A \nsubseteq P$ and let $a \in A \backslash P$; by coset equality, $a + P \neq P$. Let $b \in B$ be arbitrary. By hypothesis, $ab \in AB \subseteq P$, so here coset equality implies that

$$(a + P)(b + P) = (ab) + P = P \quad \forall b \in B.$$

Since R/P is an integral domain, $P = 0_{R/P}$, and $a + P \neq P$, we conclude that $b + P = P$. By coset equality, $b \in P$. Since b was arbitrary, this holds for all $b \in B$; hence, $B \subseteq P$.

We took two arbitrary ideals of R, and showed that if their product was a subset of P, then one of them had to be a subset of P. Thus P is a prime ideal. □

Have you noticed that this gives us an alternate proof of Theorem 8.56?

Corollary 8.61. In a ring with unity, every maximal ideal is prime, but the converse is not necessarily true.

Proof. Let R be a ring with unity, and M a maximal ideal. By Theorem 8.59, R/M is a field. By Theorem 7.25, R/M is an integral domain. By Theorem 8.60, M is prime.

The converse is not necessarily true, as not every integral domain is a field. □

Chapter Exercises.

Exercise 8.62. Determine necessary and sufficient conditions on a ring R such that in $R[x, y]$:
(a) the ideal $I = \langle x \rangle$ is prime;
(b) the ideal $I = \langle x, y \rangle$ is maximal.

Exercise 8.63. Let $n \in \mathbb{Z}$ be an integer that is not prime. Show that $n\mathbb{Z}$ is not a prime ideal.

Exercise 8.64. Show that $\{[0], [4]\}$ is a proper ideal of $\mathbb{Z}_8$, but that it is not maximal. Then find a maximal ideal of $\mathbb{Z}_8$.

Exercise 8.65. Find all the maximal ideals of $\mathbb{Z}_{12}$. Are they prime? How do you know?

Exercise 8.66. Let $\mathbb{F}$ be a field, and $a_1, a_2 \ldots, a_n \in \mathbb{F}$.
(a) Show that the ideal $\langle x_1 - a_1, x_2 - a_2, \ldots, x_n - a_n \rangle$ is both a prime ideal and a maximal ideal of $\mathbb{F}[x_1, x_2, \ldots, x_n]$.

(b) Use Exercise 8.23 to describe the common root(s) of this ideal.

Exercise 8.67. Consider the ideal $I = \langle x^2+1 \rangle$ in $R = \mathbb{R}[x]$. The purpose of this exercise is to show that I is maximal.

(a) Explain why $(x^2+x)+I = (x-1)+I$.
(b) Explain why every $f \in R/I$ has the form $r+I$ for some $r \in R$ such that $\deg r < 2$.
(c) Part (b) implies that every element of R/I can be written in the form $f = (ax+b)+I$ where $a, b \in \mathbb{R}$. Show that if $f+I$ is a nonzero element of R/I, then $a^2+b^2 \neq 0$.
(d) Let $f+I \in R/I$ be nonzero, and find $g+I \in R/I$ such that $g+I = (f+I)^{-1}$; that is, $(fg)+I = 1_{R/I}$.
(e) Explain why part (d) shows that I is maximal.
(f) Explain why $\langle x^2+1 \rangle$ is not even prime if $R = \mathbb{C}[x]$, let alone maximal. Show further that this is because the observation in part (c) no longer holds in $\mathbb{C}$.

Exercise 8.68. Let $\mathbb{F}$ be a field, and $f \in \mathbb{F}[x]$ be any polynomial that does not factor in $\mathbb{F}[x]$. Show that $\mathbb{F}[x] / \langle f \rangle$ is a field.

Exercise 8.69. Recall the ideal $I = \langle x^2+y^2-4, xy-1 \rangle$ of Exercise 8.5. We want to know whether this ideal is maximal. The purpose of this exercise is to show that it is not so "easy" to accomplish this as it was in Exercise 8.67.

(a) Explain why someone might think naïvely that every $f \in R/I$ has the form $r+I$ where $r \in R$ and $r = bx + p(y)$, for appropriate $b \in \mathbb{C}$ and $p \in \mathbb{C}[y]$; in the same way, someone might think naïvely that every distinct polynomial r of that form represents a distinct element of R/I.
(b) Show that, to the contrary, $1+I = (x+y^3-4y+1)+I$.

8.5: Ring isomorphisms

As with groups and rings, it is often useful to show that two rings have the same ring structure. With monoids and groups, we defined *isomorphisms* to do this. We will do the same thing with rings. However, ring homomorphisms are a little more complicated, as rings have two operations, rather than one.

Ring homomorphisms and their properties

Definition 8.70. Let R and S be rings. A function $f : R \to S$ is a **ring homomorphism** if for all $a, b \in R$

$$f(a+b) = f(a) + f(b)$$

and

$$f(ab) = f(a) f(b).$$

If, in addition, f is one-to-one and onto, we call it a **ring isomorphism**.

Right away, you should see that a ring homomorphism is a special type of group homomorphism with respect to addition. Even if the ring has unity, however, *it might not* be a monoid homomorphism with respect to multiplication, because there is no guarantee that $f(1_R) = 1_S$.

Example 8.71. Let $f : \mathbb{Z} \to \mathbb{Z}_2$ by $f(x) = [x]$. The homomorphism properties are satisfied:

$$f(x+y) = [x+y] = [x] + [y] = f(x) + f(y)$$

and

$$f(xy) = [xy] = [x][y] = f(x) f(y).$$

Notice that f is onto, but it is certainly not one-to-one, inasmuch as $f(0) = f(2)$.

On the other hand, consider Example 8.72.

Example 8.72. Let $f : \mathbb{Z} \to 2\mathbb{Z}$ by $f(x) = 4x$. In Example 4.3 on page 126, we showed that this was a homomorphism of groups. However, it is *not* a homomorphism of rings, because it does not preserve multiplication:

$$f(xy) = 4xy \quad \text{but} \quad f(x) f(y) = (4x)(4y) \neq f(xy).$$

Example 8.72 drives home the point that rings are more complicated than groups on account of having two operations. It is harder to show that two rings are homomorphic, and therefore harder to show that they are isomorphic. This is especially interesting in this example, since we had shown earlier that $\mathbb{Z} \cong n\mathbb{Z}$ *as groups* for all nonzero n. If this is the case with rings, then we have to find some other function between the two. Theorem 8.73 shows that this is not possible, in a way that should not surprise you.

Theorem 8.73. Let R be a ring with unity. If there exists an onto homomorphism between R and another ring S, then S is also a ring with unity.

Proof. Let S be a ring such that there exists a homomorphism f between R and S. We claim that $f(1_R)$ is an identity for S.

Let $y \in S$; the fact that R is onto implies that $f(x) = y$ for some $x \in R$. Applying the homomorphism property,

$$y = f(x) = f(x \cdot 1_R) = f(x) f(1_R) = y \cdot f(1_R).$$

A similar argument shows that $y = f(1_R) \cdot y$. Since y was arbitrary in S, $f(1_R)$ is an identity for S. □

We can deduce from this that $\mathbb{Z}$ and $n\mathbb{Z}$ are not isomorphic as rings whenever $n \neq 1$:

- to be isomorphic, there would have to exist an onto function from $\mathbb{Z}$ to $n\mathbb{Z}$;
- $\mathbb{Z}$ has a multiplicative identity;
- by Theorem 8.73, $n\mathbb{Z}$ would also have to have a multiplicative identity;
- but $n\mathbb{Z}$ does not have a multiplicative identity when $n \neq 1$.

Here are more useful properties of a ring homomorphism.

Theorem 8.74. Let R and S be rings, and f a ring homomorphism from R to S. Each of the following holds:
(A) $f(0_R) = 0_S$;
(B) for all $x \in R$, $f(-x) = -f(x)$;
(C) for all $x \in R$, if x has a multiplicative inverse and f is onto, then $f(x)$ has a multiplicative inverse, and $f(x^{-1}) = f(x)^{-1}$.

Proof. You do it! See Exercise 8.83. □

We have not yet encountered an example of a ring isomorphism, so let's consider one.

Example 8.75. Let $\mathbb{F}$ be any field, and $p = ax + b \in \mathbb{F}[x]$, where $a \neq 0$. Recall from Exercise 8.68 that $\langle p \rangle$ is maximal in $\mathbb{F}[x]$. For convenience, we will write $R = \mathbb{F}[x]$ and $I = \langle p \rangle$; by Theorem 8.59, R/I is a field.

Are $\mathbb{F}$ and R/I isomorphic? Let $f : \mathbb{F} \to R/I$ in the following way: let $f(c) = c + I$ for every $c \in \mathbb{F}$. Is f a homomorphism?

Homomorphism property? Let $c, d \in \mathbb{F}$; using the definition of f and the properties of coset addition,

$$\begin{aligned} f(c+d) &= (c+d) + I \\ &= (c+I) + (d+I) = f(c) + f(d). \end{aligned}$$

Similarly,

$$f(cd) = (cd) + I = (c+I)(d+I) = f(c) f(d).$$

One-to-one? Let $c, d \in \mathbb{F}$ and suppose that $f(c) = f(d)$. Then $c + I = d + I$; by Lemma 3.29, $c - d \in I$. By closure, $c - d \in \mathbb{F}$, while $I = \langle ax + b \rangle$ is the set of all multiples of $ax + b$. Since $a \neq 0$, the only rational number in I is 0, which implies that $c - d = 0$, so $c = d$.

Onto? Let $X \in R/I$; let $p \in R$ such that $X = p + I$. Divide p by $ax + b$ to obtain

$$p = q(ax + b) + r$$

where $q, r \in R$ and $\deg r < \deg(ax + b) = 1$. Since $ax + b \in I$, absorption tells us that $q(ax + b) \in I$, so

$$\begin{aligned} p + I &= [q(ax+b) + r] + I \\ &= [q(ax+b) + I] + (r + I) \\ &= I + (r + I) \\ &= r + I. \end{aligned}$$

Now, $\deg r < 1$ implies that $\deg r = 0$, or in other words, r is a constant. The constants of $R = \mathbb{F}[x]$ are elements of $\mathbb{F}$, so $r \in \mathbb{F}$. Hence

$$f(r) = r + I = p + I,$$

and f is onto.

We have shown that there exists a one-to-one, onto ring homomorphism from $\mathbb{F}$ to R/I; as a consequence, $\mathbb{F}$ and R/I are isomorphic as rings.

The isomorphism theorem for rings

We now consider the isomorphism theorem for groups (Theorem 4.46) in the context of rings. To do this, we need to revisit the definition of a kernel.

Definition 8.76. Let R and S be rings, and $f : R \to S$ a homomorphism of rings. The **kernel** of f, denoted $\ker f$, is the set of all elements of R that map to 0_S. That is,

$$\ker f = \{x \in R : f(x) = 0_S\}.$$

You will show in Exercise 8.85 that $\ker f$ is an ideal of R, and that the function $g : R \to R/\ker f$ by $g(x) = x + \ker f$ is a homomorphism of rings.

Theorem 8.77. Let R, S be rings, and $f : R \to S$ an onto homomorphism. Let $g : R \to R/\ker f$ be the natural homomorphism $g(r) = r + \ker f$. There exists an isomorphism $h : R/\ker f \to S$ such that $f = h \circ g$.

Proof. Define h by $h(X) = f(x)$ where $X = x + \ker f$. Is f an isomorphism? Since its domain consists of cosets, we must show first that it's well-defined:

well-defined? Let $X \in R/\ker f$ and let $x, y \in R$ such that $X = x + \ker f = y + \ker f$ — that is, $x + \ker f$ and $y + \ker f$ are two representations of the same coset, X. We must show that $h(X)$ has the same value regardless of which representation we use. By Lemma 3.29, $x - y \in \ker f$. From the definition of the kernel, $f(x - y) = 0_S$. We can apply Theorem 8.74 to see that

$$\begin{aligned} 0_S &= f(x - y) \\ &= f(x + (-y)) \\ &= f(x) + f(-y) \\ &= f(x) + [-f(y)] \\ f(y) &= f(x). \end{aligned}$$

By substitution, we have $h(y + \ker f) = f(y) = f(x) = h(x + \ker f)$. In other words, the representation of X does not affect the value of h, and h is well-defined.

homomorphism property? Let $X, Y \in R/\ker f$ and consider the representations $X = x + \ker f$

and $Y = y + \ker f$. Since f is a ring homomorphism,

$$\begin{aligned} h(X+Y) &= h((x+\ker f)+(y+\ker f)) \\ &= h((x+y)+\ker f) \\ &= f(x+y) \\ &= f(x)+f(y) \\ &= h(x+\ker f)+f(y+\ker f) \\ &= h(X)+h(Y). \end{aligned}$$

Similarly,

$$\begin{aligned} h(XY) &= h((x+\ker f)\cdot(y+\ker f)) \\ &= h((xy)+\ker f) \\ &= f(xy) \\ &= f(x)f(y) \\ &= h(x+\ker f)\cdot f(y+\ker f) \\ &= h(X)\cdot h(Y). \end{aligned}$$

Thus h is a ring homomorphism.

one-to-one? Let $X, Y \in R/\ker f$ and suppose that $h(X) = h(Y)$. Let $x, y \in R$ such that $X = x + \ker f$ and $Y = y + \ker y$. By the definition of h, $f(x) = f(y)$. Applying Theorem 8.74, we see that

$$\begin{aligned} f(x) = f(y) &\Longrightarrow f(x) - f(y) = 0_S \\ &\Longrightarrow f(x-y) = 0_S \\ &\Longrightarrow x - y \in \ker f \\ &\Longrightarrow x + \ker f = y + \ker f, \end{aligned}$$

so $X = Y$. We have shown that if $h(X) = h(Y)$, then $X = Y$. By definition, h is one-to-one.

onto? Let $y \in S$. Since f is onto, there exists $x \in R$ such that $f(x) = y$. Then $h(x + \ker f) = f(x) = y$. We have shown that an arbitrary element of the range S has a preimage in the domain. By definition, h is onto.

We have shown that h is a well-defined, one-to-one, onto homomorphism of rings. Thus h is an isomorphism from $R/\ker f$ to S. □

Example 8.78. Let $f : \mathbb{Q}[x] \to \mathbb{Q}$ by $f(p) = p(2)$ for any polynomial $p \in \mathbb{Q}[x]$. That is, f maps any polynomial to the value that polynomial gives for $x = 2$. For example, if $p = 3x^3 - 1$, then $p(2) = 3(2)^3 - 1 = 23$, so $f(3x^3 - 1) = 23$.

Is f a homomorphism? For any polynomials $p, q \in \mathbb{Q}[x]$, we have

$$f(p+q) = (p+q)(2);$$

applying a property of polynomial addition, we have

$$f(p+q) = (p+q)(2) = p(2) + q(2) = f(p) + f(q).$$

A similar property of polynomial multiplication gives

$$f(pq) = (pq)(2) = p(2) \cdot q(2) = f(p) f(q),$$

so f is a homomorphism.

Is f onto? Let $a \in \mathbb{Q}$; we need a polynomial $p \in \mathbb{Q}[x]$ such that $p(2) = a$. The easiest way to do this is to use a linear polynomial, and $p = x + (a-2)$ will work, since

$$f(p) = p(2) = 2 + (a-2) = a.$$

We took an arbitrary element of the range $\mathbb{Q}$, and showed that it has a preimage in the domain. By definition, f is onto.

Is f one-to-one? The answer is *no*. We already saw that $f\left(3x^3 - 1\right) = 23$, and from our work showing that f is onto, we deduce that $f(x+21) = 23$, so f is not one-to-one.

Let's apply Theorem 8.77 to obtain an isomorphism. First, identify $\ker f$: it consists of all the polynomials $p \in \mathbb{Q}[x]$ such that $p(2) = 0$. The Factor Theorem (7.45) implies that $x-2$ must be a factor of any such polynomial. In other words,

$$\ker f = \{p \in \mathbb{Q}[x] : (x-2) \text{ divides } p\} = \langle x-2 \rangle.$$

Since $\ker f = \langle x-2 \rangle$, Theorem 8.77 tells us that there exists an isomorphism between the quotient ring $\mathbb{Q}[x] / \langle x-2 \rangle$ and $\mathbb{Q}$.

Notice, as in Example 8.75, that $x-2$ is a linear polynomial. Linear polynomials do not factor. By Exercise 8.68, $\langle x-2 \rangle$ is a maximal ideal; so $\mathbb{Q}[x] / \langle x-2 \rangle$ must be a field—as is $\mathbb{Q}$.

A construction of the complex numbers

We conclude this section by showing that the complex numbers can be viewed not only as an "abstract" extension of $\mathbb{R}$ by an "imaginary" number $i = \sqrt{-1}$, but also as a "concrete" construction: a quotient ring of $\mathbb{R}[x]$. This not only gives you an exciting new view of the complex numbers, but also suggests how we can "solve" polynomial equations in general.

I assume you already know the basics of the complex number system: namely, $i^2 = -1$, and any complex number number takes the form $a + bi$ for some $a, b \in \mathbb{R}$. Addition and multiplication of complex numbers follow very simple rules:

$$(a+bi) + (c+di) = (a+c) + (b+d)i \quad \text{and} \quad (a+bi)(c+di) = (ac-bd) + (ad+bc)i.$$

We can get the same behavior out of a quotient ring. Let $I = \langle x^2 + 1 \rangle$ and $\mathcal{C} = \mathbb{R}[x]/I$. We claim that $\mathcal{C} \cong \mathbb{C}$.

We will construct an explicit isomorphism in just a moment, but first let's look at how arithmetic in $\mathcal{C}$ mimicks the properties of $\mathbb{C}$. Start off by considering the elements of R.

Proposition 8.79. Let $P \in R$; by definition of a quotient ring, P has the form $p + I$, where $p \in \mathbb{R}[x]$. Without loss of generality, we may assume that $\deg p < 2$.

Proof. Since $\mathbb{R}[x]$ is a Euclidean domain, we can find $q, r \in \mathbb{R}[x]$ such that $p = q\left(x^2+1\right) + r$ and $r = 0$ or $\deg r < 2$. Rewrite the equation as $r = p - q\left(x^2+1\right)$. By substitution,

$$p + I = \left(q\left(x^2+1\right) + r\right) + I.$$

Arithmetic in a quotient ring allows us to rewrite this as

$$p + I = \left[q\left(x^2+1\right) + I\right] + (r + I).$$

Recall that $I = \left\langle x^2+1 \right\rangle$. By absorption, $q\left(x^2+1\right) \in I$, so $I = q\left(x^2+1\right) + I$. Since $I = 0_{\mathcal{C}}$, we can rewrite the above equation as $p + I = r + I$. In other words, we can write r in place of p, and obtain the same result as using p. Thus, we can assume that $\deg p = \deg r < 2$. □

Thanks to Proposition 8.79, we can write any element of $\mathcal{C}$ as $(bx + a) + I$, where $a, b \in \mathbb{R}$: its degree is less than 2, and its coefficients are real. Even this is a bit much work, though. To simplify the writing further, we notice that one element of $\mathcal{C}$ has a very nice property.

Proposition 8.80. $(x + I)^2 = -1 + I$.

Proof. You do it! See Exercise 8.84. □

This motivates us to adopt a highly suggestive notation.

Notation 8.81. We will write each $(bx + a) + I \in \mathcal{C}$ as $a + b\mathbf{i}$. If $b = 0$, we will write a and understand that we mean $a + 0\mathbf{i}$ or $a + I$.

This means that we can write $\mathbf{i} = x + I$ and $\mathbf{i}^2 = -1$. Exploring the resulting arithmetic, we find some astonishing parallels to complex arithmetic:

$$\begin{aligned}(a + b\mathbf{i}) + (c + d\mathbf{i}) &= [(bx + a) + I] + [(dx + c) + I] \\ &= [(b + d)x + (a + c)] \\ &= (a + c) + (b + d)\mathbf{i}\end{aligned}$$

and

$$\begin{aligned}(a + b\mathbf{i})(c + d\mathbf{i}) &= [(bx + a) + I][(dx + c) + I] \\ &= (bx + a)(dx + c) + I \\ &= \left[bdx^2 + (bc + ad)x + ac\right] + I \\ &= bd\left(x^2 + I\right) + [((bc + ad)x + ac) + I] \\ &= (-bd + I) + [((bc + ad)x + ac) + I] \\ &= [(bc + ad)x + (ac - bd)] + I \\ &= (ac - bd) + (ad + bc)\mathbf{i}.\end{aligned}$$

Things are looking rather encouraging at this point, so let's try to build an explicit isomorphism. We will build on the notation we have used thus far.

Theorem 8.82. $\mathbb{R}[x] / \langle x^2+1 \rangle \cong \mathbb{C}$.

Proof. Let I, $\mathcal{C}$, and $\mathbf{i}$ hold the same meanings as above. To use the isomorphism theorem, start by defininining a map $f : \mathbb{R}[x] \to \mathbb{C}$ in the following way:

1. Let $p \in \mathbb{R}[x]$.
2. Let $bx+a$ be the remainder of division of p by x^2+1.
3. Let $f(p) = a+bi$.

We claim that f is a homomorphism. To see why, let $p, q \in \mathbb{R}[x]$. Let $bx+a$ and $cx+d$ be the remainders of division of p and q (respectively) by x^2+1. It is pretty clear that

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

and

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

It's a little harder to show that these equal $f(p+q)$ and $f(pq)$, respectively. To see that they do, consider $f(p+q)$ first. Since the remainders of division were $bx+a$ and $dx+c$, we know that there exist $h_p, h_q \in \mathbb{R}[x]$ such that

$$\begin{aligned} p+q &= \left[h_p\left(x^2+1\right)+(bx+a)\right]+\left[h_q\left(x^2+1\right)+(dx+c)\right] \\ &= \left(h_p+h_q\right)\left(x^2+1\right)+\left[(b+d)x+(a+c)\right]. \end{aligned}$$

We see that the remainder of division of $p+q$ by x^2+1 is $(b+d)x+(a+c)$, so by definition,

$$f(p+q) = (a+c)+(b+d)i = f(p)+f(q).$$

As for multiplication,

$$\begin{aligned} pq &= \left[h_p\left(x^2+1\right)+(bx+a)\right]\left[h_q\left(x^2+1\right)+(dx+c)\right] \\ &= h'\left(x^2+1\right)+\left[bdx^2+(bc+ad)x+ac\right], \end{aligned}$$

where

$$h' = h_p h_q\left(x^2+1\right)+h_p(dx+c)+h_q(bx+a).$$

(We don't really care much for the details of h', but there they are.) We can rewrite this again as

$$\begin{aligned} pq &= \left(h'+bd\right)\left(x^2+1\right)+\left[bdx^2+(bc+ad)x+ac\right]-bd\left(x^2+1\right) \\ &= h''\left(x^2+1\right)+\left[(bc+ad)x+(ac-bd)\right], \end{aligned}$$

where $h'' = h'+bd$. (Again, we don't really care much for the details of h''.) We have now written pq in a form that allows us to apply the definition of f:

$$f(pq) = (ac-bd)+(bc+ad)i = f(p)f(q).$$

We have shown that f is indeed a ring homomorphism. It is *not* an isomorphism, since $f\left(x^2\right) = i = f\left(2x^2+1\right)$ (and a bunch more, besides). However, did you notice something? We also have

$$\ker f = \left\langle x^2+1 \right\rangle = I,$$

since the remainder of division of p by x^2+1 is zero if and only if p is a multiple of x^2+1, and hence, in its principal ideal! By the isomorphism theorem, then, there exists an isomorphism from $\mathcal{C} = \mathbb{R}[x] / \ker f$ to $\mathbb{C}$, as claimed by the theorem. □

Exercises.

Exercise 8.83. Prove Theorem 8.74.

Exercise 8.84. Prove Proposition 8.80.

Exercise 8.85. Let R and S be rings, and $f : R \to S$ a homomorphism of rings.

(a) Show that $\ker f$ is an ideal of R.

(b) Show that the function $g : R \to R/\ker f$ by $g(x) = x + \ker f$ is a homomorphism of rings.

Exercise 8.86. Let R be a ring and $a \in R$. The **evaluation map with respect to** a is $\varphi_a : R[x] \to R$ by $\varphi_a(f) = f(a)$; that is, φ_a maps a polynomial to its value at a.

(a) Suppose $R = \mathbb{Q}[x]$ and $a = 2/3$, find $\varphi_a\left(2x^2-1\right)$ and $\varphi_a(3x-2)$.

(b) Show that the evaluation map is a ring homomorphism.

(c) Recall from Example 8.75 that $\mathbb{Q}$ is isomorphic to the quotient ring $\mathbb{Q}[x] / \langle ax+b \rangle$ where $ax+b \in \mathbb{Q}[x]$ is non-zero. Use Theorem 8.77 to show this a different way.

Exercise 8.87. Use Theorem 8.77 to show that $\mathbb{Q}[x] / \left\langle x^2 \right\rangle$ is isomorphic to

$$\left\{ \begin{pmatrix} a & b \\ 0 & a \end{pmatrix} \right\} \subset \mathbb{Q}^{2\times 2}.$$

Note: $\mathbb{Q}^{2\times 2}$ is not commutative! However, $\mathbb{Q}[x] / \left\langle x^2 \right\rangle$ is commutative, so this isomorphism shows that the given susbet of $\mathbb{Q}^{2\times 2}$ is, too. (It might not be the most efficient way of showing that, of course.)

Exercise 8.88. In this exercise we show that $\mathbb{R}$ is not isomorphic to $\mathbb{Q}$ as rings, and $\mathbb{C}$ is not isomorphic to $\mathbb{R}$ as rings.

(a) Assume to the contrary that there exists an isomorphism f from $\mathbb{R}$ to $\mathbb{Q}$.

  (i) Use the properties of an onto homomorphism to find $f(1)$.

  (ii) Use the properties of a homomorphism with the result of (i) to find $f(2)$.

  (iii) Use the properties of a homomorphism to obtain a contradiction with $f\left(\sqrt{2}\right)$.

(b) Find a similar proof that $\mathbb{C}$ and $\mathbb{R}$ are not isomorphic.

Exercise 8.89. Show that if R is an integral domain, then $\mathrm{Frac}(R)$ is isomorphic to the intersection of all fields containing R as a subring.

Part III
Applications