Notes on Abstract Algebra

August 22, 2019

Course:
Math 31 – Dartmouth College

Instructor:
Scott M. LaLonde

Contents

Preface

These notes were originally designed to accompany the Summer 2012 and Summer 2013 incarnations of Math 31 at Dartmouth College. This course is the introductory undergraduate-level abstract algebra course at Dartmouth; it is aimed at math majors, especially those who are not necessarily planning to attend graduate school in mathematics. That being said, Math 31 covers all the material one would expect to see in a first abstract algebra course at most schools (and then some).

These notes owe their existence in large part to my inability to choose an appropriate textbook for Math 31. The two that I have thought to be "the one" at various times are:

- *Abstract Algebra* by I. N. Herstein
- *Abstract Algebra: A First Course* by Dan Saracino

I used Herstein's book during the summer 2012 term, and Saracino was the official course textbook for summer 2013. Each book has shortcomings (as does any textbook), and these notes represent an attempt to address such imperfections. In doing so, I have pulled from my own knowledge of algebra, as well as from many supplementary sources. In particular, I have consulted the following books to some extent:

- *A First Course in Abstract Algebra* by John Fraleigh
- *Contemporary Abstract Algebra* by Joseph Gallian
- *Abstract Algebra* by John A. Beachy and William D. Blair
- *A Book of Abstract Algebra* by Charles C. Pinter

Based on my statements thus far, you may think that I despise all of the abstract algebra books that are currently in print. This is not true at all—there are many fine books on the list above, and many others that I have not mentioned. The truth is that I have not been able to find a book that covers the basics of abstract algebra in the exact style and order that I prefer. I would need to deviate greatly from most of these books, thus risking the "flow" of the course. I would also need to take care in assigning homework problems—the propositions and theorems that I choose to

leave as exercises are not always treated that way by the author. These personal prejudices—and nothing else—are what have given me the motivation to write this manuscript.

I should mention that Herstein's book is probably my favorite out of those listed above—it is certainly the best one for emphasizing the *abstract* in *abstract algebra*. I also have a bit of a bias toward Fraleigh, which is a great book for a more lightweight algebra course (though it is not as easy as many people would make it out to be). Therefore, these two books have probably influenced me the most in the design and writing of these notes. I have also attempted to include historical anecdotes in the discussion—many of these are taken from the first chapter of Pinter's book, and from *Unknown Quantity: A Real and Imaginary History of Algebra* by John Derbyshire. (While the author of the last book is of questionable character, he does tell a good story.)

To the Student

If you are reading these notes in order to learn abstract algebra, there are a few things to keep in mind. I have designed this text largely from my lecture notes, albeit with additional examples and explanation peppered throughout. Therefore, it is written in a way that strongly reflects my teaching style. I pose questions to the reader quite often. Since the material can be quite abstract, I try to motivate new concepts through familiar examples and ideas. If I believe that the proof of a particular result would make a good exercise, I have left it as such. Finally, I have included some of my favorite homework problems as exercises at the end of each section.

Perhaps this description appeals to you. If so, then I hope you enjoy reading these notes. They should make you feel as though you are taking a course in abstract algebra with me (though without the pressure of my homework and exams!), so the tone is very conversational. I have also included some anecdotes regarding the history of abstract algebra, as well as some applications and previews of more advanced topics in the field. Altogether, I am trying to tell a story, and thus I hope that the notes are fairly readable.

Finally, I should provide some warnings to you. Even though I have taught this course twice, I still consider these notes to be quite raw. I expect that there will be typographical errors—and possibly other errors—so I hope that you will forgive them. There will also be places where the exposition is a bit shaky; I am still searching for the best way to explain certain topics. Finally, these notes are woefully incomplete when it comes to the overall field of algebra. There are many interesting topics that I have not been able to fit into my courses, so they do not appear here. (There are even some things that I have included that might not fit into an actual course.) However, there are many great books where these topics can be found, and I will try to direct you there when the time comes. In all, I have tried

to write a basic, self-contained introduction to elementary abstract algebra—the fundamentals of group theory, and a treatment of rings and fields that emphasizes polynomials. It will be up to you to decide whether I have succeeded.

Chapter 1

Introduction

To fully appreciate (and understand) any new topic in mathematics, it always helps to relate that topic to one's prior knowledge before fully diving in. This introductory chapter is designed to serve this purpose—we will informally introduce some of the ideas behind abstract algebra from a historical perspective, and then discuss some familiar examples for the purpose of motivation.

1.1 What is Abstract Algebra?

In order to answer the question posed in the title, there are really two questions that we must consider. First, you might ask, "What does 'abstract' mean?" You also probably have some preconceived notions about the meaning of the word "algebra." This should naturally lead you to ask, "How does this course relate to what I already know about algebra?" We will see that these two questions are very much intertwined. The second one is somewhat easier to address right now, so we will start there.

1.1.1 History

If you've spent any time looking at suggested textbooks, you have probably noticed that "abstract algebra" looks very different from the algebra you know. Many of the words in the table of contents are likely unrecognizable, especially in the context of high school algebra. However, this new notion of *abstract* algebra does relate to what you already know—the connections just aren't transparent yet. We will shed some light on these connections by first discussing the history of abstract algebra. This will set the stage for the beginning and the end of the course, and the tools that we develop in between will allow us to link the ideas of modern algebra with your prior knowledge of the subject.

In high school, the word "algebra" often means "finding solutions to equations." Indeed, the Persian poet-mathematician Omar Khayyám defined algebra to be the

"science of solving equations."[1] In high school, this probably meant that you were solving **linear** equations, which look like

$$ax + b = 0,$$

or **quadratic** equations, of the form

$$ax^2 + bx + c = 0$$

Methods for solving these equations were known even in ancient times.[2] Indeed, you learned to solve quadratics by factoring, and also by the **quadratic formula**, which gives solutions in terms of square roots:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

You may have also learned to factor cubic polynomials, which have the form

$$ax^3 + bx^2 + cx + d.$$

Techniques were known to ancient mathematicians, including the Babylonians, for solving certain types of cubic equations. Islamic mathematicians, including Omar Khayyám[3], also made significant progress. However, what about a general formula for the roots? Can we write down a formula, like the quadratic formula, which gives us the roots of *any* cubic in terms of square roots and cube roots? There is a formula, which we won't write down here, but it took quite a longtime for mathematicians to find it.

The general formula for cubics[4] was discovered in Italy during the Renaissance, by Niccoló Fontana **Tartaglia**. As was the case with many mathematicians, Tartaglia led an interesting and somewhat tragic life. Born in 1500, Tartaglia was not his real name—it is actually an Italian word meaning "the stammerer." As a child, he suffered a sabre attack during the French invasion of his hometown of Brescia, which resulted in a speech impediment. He was entirely self-taught, and was not only a great mathematician, but also an expert in ballistics. In 1535, he found a general method for solving a cubic equation of the form

$$x^3 + ax^2 + b = 0,$$

[1] Khayyám was not the first to use the term *algebra*. The Arabic phrase *al-jabr*, meaning "balancing" or "reduction" was first used by Muhammad ibn al-Khwārizmi.

[2] This statement deserves clarification. The ancients were able to solve quadratic equations, provided that the solutions didn't involve complex numbers or even *negative* numbers.

[3] It is quite extraordinary that Khayyám was able to make such progress. He lacked the formal symbolism that we now have, using only words to express problems. Also, negative numbers were still quite mysterious at this time, and his solutions were usually geometric in nature.

[4] Again, this formula does not work in full generality. Tartaglia was only able to deal with cubics which had nonnegative discriminant—for such cubics, the solution did not involve square roots of negative numbers. These cubics are the ones which have exactly one real root.

i.e., one with no x term. As was customary in those days, Tartaglia announced his accomplishment, but he kept the details secret. He eventually entered into a "math duel" with Antonio Fior, who had learned a method for solving cubics of the form

$$x^3 + ax + b = 0$$

from his mentor, Scipio del Ferro. These "duels" were not duels in the more familiar and brutal sense, but public competitions in problem solving. The adversaries would exchange lists of problems, and each competitor would attempt to solve more than the other. A few days beforehand, Tartaglia extended his method to handle Fior's brand of cubic equations as well. Within 2 hours, he solved all of Fior's problems, and Fior solved none of Tartaglia's.

His victory over Fior brought Tartaglia a reasonable amount of fame, and in particular it brought him to the attention of the mathematician (and all-around scoundrel[5]) Gerolamo Cardano. Born in 1501, Cardano was an accomplished mathematician and physician, and he was actually the first person to give a clinical description of typhus fever. He was also a compulsive gambler and wrote a manual for fellow gamblers, which was actually an early book on probability. Eventually, Tartaglia cut a deal with Cardano—he divulged his secret method to Cardano in exchange for help in obtaining a job with the military as a ballistics adviser. Cardano was actually writing an algebra book, titled *Ars Magna* ("The Great Art"), in which he collected all of the algebra that was known in Europe at the time. He published Tartaglia's result, while acknowledging del Ferro for his discovery of the solution for cubics with no x^2 term. Cardano gave Tartaglia the appropriate credit for rediscovering this result. Tartaglia was furious at this blatant breach of trust, and the two had a long feud after that. Despite this, the formula is now known as the **Cardano-Tartaglia formula** in honor of both men.

After the question of solving cubics was resolved, people turned their attention to the quartic equation:

$$x^4 + ax^3 + bx^2 + cx + d = 0.$$

Lodovico Ferrari, who was Cardano's personal servant, found the formula. He had learned mathematics, Latin, and Greek from Cardano, and he had actually bested Tartaglia in a duel in 1548. He reduced the problem to that of solving an equation involving the so-called resolvent cubic, and then used Tartaglia's formula. In fact, this work led Cardano to his decision to publish Tartaglia's work, since he needed it in order to describe Ferrari's result.

After Ferrari's work, the obvious next step was to try to unearth general methods for finding the roots of fifth (and higher) degree polynomials. This was not so easy. It took over 200 years before any real progress was made on this question. In the early 19th century, two young mathematicians independently showed that *there is no general formula* for the roots of a quintic polynomial.

[5] At least one author describes him as a "piece of work."

The first of these two young men was Niels Henrik Abel, who proved his result in 1824. He was Norwegian, and he died from tuberculosis 5 years after publishing his work (at the age of 26). The other young prodigy, who has one of the best-known stories among mathematicians, was a French radical by the name of Évariste Galois. He did much of his work around 1830, at the age of 18, though it wasn't published until about 15 years later. His story was very tragic, beginning in his teenage year's with his father's suicide. He was then refused admission to the prestigious École Polytechnique on the basis that his solutions to the entrance exam questions were too radical and original. He eventually entered the École Normale, but was expelled and imprisoned for political reasons. Even his groundbreaking work was largely ignored by the preeminent French mathematicians of his day, including Cauchy, Fourier, and Poisson. Last but not least, he was killed in a duel (under mysterious circumstances) at the age of 20. Fortunately, Galois entrusted his work to a friend on the night before the duel, and it was published posthumously 15 years later.

In proving that there is no "quintic formula," Abel and Galois both essentially invented what is now known as a **group**. In particular, Galois studied groups of permutations of the roots of a polynomial. In short, he basically studied what happens when you "shuffle" the roots of the polynomial around. Galois' ideas led to a whole field called **Galois theory**, which allows one to determine whether *any* given polynomial has a solution in terms of radicals. Galois theory is regarded as one of the most beautiful branches of mathematics, and it usually makes up a whole course on its own.

This is the context in which groups (in their current form) began to arise. Since then, the study of groups has taken off in many directions, and has become quite interesting in its own right. Groups are used to study permutations, symmetries, and cryptography, to name a few things. We'll start off by studying groups abstractly, and we will consider interesting examples of groups along the way. Hopefully this discussion will give you an idea of where groups come from, and where your study of algebra can eventually lead.

1.1.2 Abstraction

We've just discussed where some of the ideas in a course on abstract algebra come from historically, but where do we go with them? That is, what is the real goal of a course like this? That's where the "abstract" part comes in.

In many other classes that you've taken, namely calculus and linear algebra, things are very concrete with lots of examples and computations. In this class, there will still be many examples, but we will take a much more general approach. We will define groups (and other algebraic structures) via a set of desirable axioms, and we will then try to logically deduce properties of groups from these axioms. To give you a preview, let me say the following regarding group theory. A **group** will basically be a set (often consisting of numbers, or perhaps matrices, but also of other objects) with some sort of operation on it that is designed to play the role of

addition or multiplication. Additionally, we'll require that the operation has certain desirable arithmetic properties. In the second part of the course, we will study objects called **rings**. These will arise when we allow for two different operations on a set, with the requirement that they interact well (via a distributive law). Along the way, we will try to do as many examples as possible. Some will come from things that are familiar to you (such as number systems and linear algebra), but some will be totally new. In particular, we will emphasize the role of groups in the study of symmetry.

Despite the concrete examples, there will still be an overarching theme of classification and structure. That is, once we've defined things such as groups and rings and proven some facts about them, we'll try to answer much broader questions. Namely, we'll try to determine when two groups are the same (a concept called **isomorphism**), and when we can break a group down into smaller groups. This idea of classification is something that occurs in pretty much every branch of math (algebra, topology, analysis, etc.). It may seem overly abstract, but it is paramount to the understanding of algebra.

We will start off slowly in our abstract approach to algebra—we'll begin with a couple of motivating examples that should already be familiar to you. Then we'll study one which is more interesting, and perhaps less familiar. Once we've done these, we'll write down the actual definition of a group and begin to study these objects from an "abstract" viewpoint.

1.2 Motivating Examples

As a precursor to group theory, let's talk a little bit about some structures with which you should already be familiar. We will see shortly that these objects will turn out to be simple examples of groups.

1.2.1 The Integers

Let $\mathbb{Z}$ denote the set of all integers. Of course we have an arithmetic operation on $\mathbb{Z}$, namely the one given by addition. That is, given two integers, we can add them together to obtain a new integer. We'll let $\langle \mathbb{Z}, + \rangle$ denote the set of integers endowed with the operation of addition. (At this point, we'll pretend that we don't know how to do anything else yet, such as multiplying integers.)

What desirable properties does $\langle \mathbb{Z}, + \rangle$ have? First of all, what happens if I try to add three integers, say

$$a + b + c?$$

Formally, addition is an operation that just takes in two integers and produces a new integer. Therefore, to make sense of the above expression, we would have to break things down into steps. We could first add a and b, then add c to the result:

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

On the other hand, we could add b and c, and then add a:

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

Both are legitimate ways of defining $a + b + c$. Fortunately, they turn out to be the same, and it doesn't matter which way we add things. That is to say, addition on $\mathbb{Z}$ is **associative**. You could also point out that addition is **commutative**—we can write our integers in either order when adding them. There is also a special element of $\mathbb{Z}$ that acts as an **identity** with respect to addition: for any $n \in \mathbb{Z}$, we have

$$n + 0 = n.$$

Recall that middle/high school algebra is all about solving simple equations. For example, if I have the equation

$$x + 5 = 7,$$

how do I find x? I need to subtract 5 from both sides. Since we only know how to add integers, it would be more appropriate to say that we should add -5 to both sides, which gives $x = 2$. Now imagine instead that I wrote the equation

$$x + 5 = 0.$$

Then the solution would instead be $x = -5$. What we are really saying here is that every integer n has the property that

$$n + (-n) = 0,$$

and we say that n has an **additive inverse**, $-n$. It is the presence of an additive inverse that lets us solve simple equations like $x+5 = 7$ over the integers. In summary, addition on $\mathbb{Z}$ has three nice properties:

- associativity,
- the existence of an identity element, and
- the existence of additive inverses.

These are the properties that we would really like to have in a group, so that we can perform basic arithmetic. Notice that we've left out commutativity—indeed, we won't always require it to hold. (We will soon see an important example of such an algebraic object.)

What if I now pretend that we can only multiply integers: $\langle \mathbb{Z}, \cdot \rangle$? What properties do we have now? Well, multiplication is still associative. There is also a multiplicative identity:

$$n \cdot 1 = 1 \cdot n = n$$

for all $n \in \mathbb{Z}$. What about inverses? Given $n \in \mathbb{Z}$, is there always an integer x so that

$$n \cdot x = 1?$$

No—it even fails if we take $n = 2$. In fact, it turns out that only 1 and -1 have multiplicative inverses. Thus we have just shown that $\langle \mathbb{Z}, \cdot \rangle$ does not satisfy our list of axioms, since not every integer possesses a multiplicative inverse. Therefore, it will not fit our eventual definition of a group.

1.2.2 Matrices

Let's now turn to another example that you should remember from your linear algebra class. You likely spent a lot of time there studying properties of matrices, particularly those belonging to the set

$$\mathrm{M_n}(\mathbb{R}) = \{n \times n \text{ matrices with entries in } \mathbb{R}\}.$$

You saw that you can add matrices by simply adding corresponding entries, and you also learned how to multiply matrices (which is slightly more complicated).

Let's think about $\langle \mathrm{M_n}(\mathbb{R}), + \rangle$ first. Is the operation associative? The answer is of course yes, since addition of real numbers is associative. Is there an identity? Yes—the zero matrix is an additive identity. How about inverses? Well, if A is a matrix, $-A$ is its additive inverse. Thus $\langle \mathrm{M_n}(\mathbb{R}), + \rangle$ satisfies the axioms. We should also note that the operation in question is commutative.

What about $\langle \mathrm{M_n}(\mathbb{R}), \cdot \rangle$? It should have been pointed out in your linear algebra class that matrix multiplication is associative. There is also an identity, namely the **identity matrix**, which has the property that

$$A \cdot I = I \cdot A = A$$

for all $A \in \mathrm{M_n}(\mathbb{R})$. What about inverses? Given an $n \times n$ matrix A, can I always find a matrix B so that

$$AB = BA = I?$$

We know from linear algebra that the answer is no. Quite a bit of time is spent discussing the properties of singular and nonsingular matrices. In particular, nonsingular matrices turn out to be invertible, in the sense that they have multiplicative inverses. Therefore, if we were to restrict our attention to the set

$$\mathrm{GL_n}(\mathbb{R}) = \{A \in \mathrm{M_n}(\mathbb{R}) : A \text{ is invertible}\},$$

then $\mathrm{GL_n}(\mathbb{R})$ does satisfy the axioms. Note that $\mathrm{GL_n}(\mathbb{R})$ is not commutative, since matrix multiplication isn't. This is exactly why we didn't include commutativity on our list of axoms—we will eventually encounter many interesting examples of groups for which the operation fails to be commutative.

1.3 The Integers mod n

Now let's try to build an interesting example that you may or may not have seen before: the set of integers mod n. This example will serve two purposes: we'll use this as motivation for the study of groups, and it will provide an avenue for introducing some facts about the integers that we will need later on.

So far we've decided that $\mathbb{Z}$ (under addition) and the set of invertible n–by–n matrices (under matrix multiplication) have three nice properties, namely associativity, an identity element, and the presence of inverses. There is another set that has these properties, and it will be built from the integers using an operation called **modular arithmetic** (or "clock" arithmetic). Let's start with a specific example.

Example 1.3.1. Consider the set

$$\{0, 1, 2, 3, 4, 5\}.$$

We can define an "addition operation" $+_6$ on this set in the following way: to "add" two elements a and b of this set,

1. Add a and b as integers: $a + b$.
2. Divide $a + b$ by 6 and take the remainder. This is $a +_6 b$.

(This second step will be called "reduction mod 6.") In other words, we ignore multiples of six and only look at the remainder. For example, to compute $2 +_6 5$, we add 2 and 5 and then take the remainder:

$$2 + 5 = 7 = 6 \cdot 1 + 1,$$

so $2 +_6 5 = 1$. (Another way of writing this last calculation is

$$2 + 5 = 7 \equiv 1 \bmod 6,$$

where the notation "$7 \equiv 1 \bmod 6$" simply means that 7 and 1 leave the same remainder when divided by 6.) By adding and then reducing in this way, we ensure that the given set is closed under the operation $+_6$.

Exercise 1.1. For practice, try computing the following two examples:

(a) $3 +_6 1$

(b) $2 +_6 4$

Aside 1.3.2. The operation that we've defined above is often described as "clock arithmetic." This is an apt name, since we can think of the hours on a clock as the

integers $\{0, 1, \ldots, 11\}$ (where we are interpreting 12 o'clock as 0), and compute time by working mod 12. For example, if it's 11 o'clock right now, in 4 hours it will be

$$11 +_{12} 4 = 15 \equiv 3 \bmod 12,$$

or 3 o'clock. We could do a similar thing for minutes by instead working mod 60, or we could compute with a 24-hour clock by working mod 24.

Note that there is really nothing special about 6 (or 12, or 24, or 60) in the example above—we can do this sort of thing for any integer n. That is, given $n \in \mathbb{Z}$, we can look at the set

$$\mathbb{Z}_n = \{0, 1, 2, \ldots, n-1\},$$

endowed with the operation of "addition mod n." We define this operation $+_n$ in exactly the same way as above: to add two elements $a, b \in \mathbb{Z}_n$, do the following:

1. Add a and b as integers: $a + b$.
2. Divide $a + b$ by n and take the remainder. This is $a +_n b$.

Unfortunately, this definition is not very precise. In fact, it would probably please very few mathematicians. One of the things that you will learn in this class is to be careful and precise in your exposition of mathematical ideas. Now is a good time to start.

In the previous example, we actually employed a fact about the integers that you have probably used many times without really thinking about it. Specifically, we used the fact that we can always divide two integers to obtain a quotient and remainder. You've probably been using this idea since elementary school, when you first learned long division. Therefore, it may seem like second nature, but you should understand that it really is a *theorem* about the integers.

Theorem 1.3.3 (Division algorithm)**.** *Let a and n be integers, with $n > 0$. Then there exist unique integers q and r, with $0 \leq r \leq n-1$, such that*

$$a = qn + r.$$

Aside 1.3.4. In general, theorems are statements which require proof. Therefore, when we state something as a theorem we will usually follow up by proving it. However, it would be better to simply get to the important things, and also to ease you into the idea of writing proofs. Therefore, in the beginning we'll be a little selective about what we choose to prove. In particular, we will not prove the Division Algorithm. You've seen many examples of it in action throughout your lives, and there are many proofs available in the literature (see Saracino, Lemma 2.1, for example) if you're interested in reading it.

As far as the Division Algorithm goes, we're interested in the (unique) remainder. This remainder happens to always lie between 0 and $n-1$, so we can think of it as an element of $\mathbb{Z}_n$. In other words, if we fix $n \in \mathbb{Z}$, then any $a \in \mathbb{Z}$ determines a unique element of $\mathbb{Z}_n$. Let's give this element a name.

Definition 1.3.5. Let $a, n \in \mathbb{Z}$, and write $a = qn + r$. We denote the remainder r by $\overline{a}$, or $[a]_n$, and call it the **remainder of a mod n**.

We'll usually use the notation $[a]_n$, since it emphasizes the role of n.

Example 1.3.6. Let $a = 19$ and $n = 3$. Then

$$a = 19 = 6 \cdot 3 + 1 = 6n + 1,$$

so the remainder is

$$[a]_n = [19]_3 = 1.$$

It will be useful to discuss situations where two integers yield the same element of $\mathbb{Z}_n$. This leads us to define the notion of congruence mod n.

Definition 1.3.7. Let $a, b \in \mathbb{Z}$, and let n be a positive integer. We say that a **is congruent to b mod n**, written

$$a \equiv b \bmod n,$$

if $[a]_n = [b]_n$. In other words, a and b are congruent if they leave the same remainder when divided by n.

Example 1.3.8. Are 19 and 7 congruent mod 3? Yes—we already saw that $[19]_3 = 1$, and

$$7 = 2 \cdot 3 + 1,$$

so $[7]_3 = 1$ as well. Therefore, $7 \equiv 19 \bmod 3$. However, 9 is not congruent to 19 mod 3, since

$$9 = 3 \cdot 3 + 0,$$

so $[9]_3 = 0$.

Now we have all the tools we need in order to properly define addition mod n.

Definition 1.3.9. Let $\mathbb{Z}_n = \{0, 1, 2, \ldots, n-1\}$ denote the set of integers mod n. We define the sum of two elements $a, b \in \mathbb{Z}_n$ by

$$a +_n b = [a + b]_n.$$

Aside 1.3.10. What we are really saying here is that the elements of $\mathbb{Z}_n$ are not really integers per se, but families of integers, all of which have the same remainder. Families like this are what we will eventually call **equivalence classes**. We will eventually see that the correct formulation of $\mathbb{Z}_n$ is done in terms of equivalence classes. However, the definition we have given here is perfectly valid and it gives us a straightforward way of thinking about $\mathbb{Z}_n$ for our current purposes.

You will soon find that, when you are trying to prove mathematical statements, it is very useful to have different formulations of certain mathematical concepts. With this in mind, we will introduce another way to characterize congruence, which will be quite useful. Suppose that $a \equiv b \bmod n$. What can we say about $a - b$? Well, by the Division Algorithm, there are integers p, q such that

$$a = qn + [a]_n \quad \text{and} \quad b = pn + [b]_n.$$

Now

$$a - b = qn + [a]_n - pn - [b]_n = (q - p) \cdot n + ([a]_n - [b]_n).$$

But $[a]_n = [b]_n$, so $a - b = (q - p) \cdot n$. In other words, $a - b$ is a multiple of n. It goes the other way too—if n divides $a - b$, then it has to divide $[a]_n - [b]_n$. But $[a]_n$ and $[b]_n$ both lie strictly between 0 and $n - 1$, so $-n < [a]_n - [b]_n < n$. The only multiple of n in this interval is 0, which forces $[a]_n = [b]_n$. Thus we have essentially proven:

Proposition 1.3.11. *Two integers a and b are congruent mod n if and only if n divides $a - b$.*

By "divide," we mean the following:

Definition 1.3.12. We say that an integer n **divides** another integer m if there exists $c \in \mathbb{Z}$ such that $n = cm$.

Now let's get on with our study of $\mathbb{Z}_n$. We'll look first at $\langle \mathbb{Z}_n, +_n \rangle$. Does it satisfy the axioms that we came up with? We've set up addition in such a way that $\mathbb{Z}_n$ is automatically closed under the operation, so there are no worries there. We need to check that addition on $\mathbb{Z}_n$ is associative—this ought to be true, since addition is associative on $\mathbb{Z}$. However, this is something that we really do need to check. If we let $a, b, c \in \mathbb{Z}_n$, then

$$(a +_n b) +_n c = [a + b]_n +_n c = [[a + b]_n + c]_n,$$

while

$$a +_n (b +_n c) = a +_n [b + c]_n = [a + [b + c]_n]_n.$$

Are these the same? To check this, we will need to see that

$$[a+b]_n + c \equiv a + [b+c]_n \bmod n.$$

To do so, we'll use Proposition 1.3.11. That is, we will check that these two integers differ by a multiple of n. Well, we know that

$$a + b = qn + [a+b]_n$$

for some $q \in \mathbb{Z}$, and

$$b + c = pn + [b+c]_n$$

for some $p \in \mathbb{Z}$, by the Division Algorithm. But then

$$\begin{aligned}([a+b]_n + c) - (a + [b+c]_n) &= (a + b - qn) + c - a - (b + c - pn) \\ &= -qn + pn \\ &= (p-q) \cdot n.\end{aligned}$$

Therefore, $[a+b]_n + c \equiv a + [b+c]_n \bmod n$, so $+_n$ is associative.

The other axioms are much easier to check. The identity element is clearly 0, and the additive inverse of $a \in \mathbb{Z}_n$ is just $n - a$, since

$$a +_n (n-a) = [a + n - a]_n = [0]_n = 0.$$

(Conveniently enough, $n - a = [-a]_n$.) The operation $+_n$ is even commutative, so $\langle \mathbb{Z}_n, +_n \rangle$ satisfies all of the axioms that we came up with.

Now let's look at a different operation on $\mathbb{Z}_n$. It is entirely possible to multiply elements of $\mathbb{Z}_n$ by simply defining

$$a \cdot_n b = [a \cdot b]_n,$$

just as with addition.

Example 1.3.13. Working mod 6, we have

$$2 \cdot 4 = [8]_6 \equiv 2 \bmod 6$$

and

$$3 \cdot 5 = [15]_6 \equiv 3 \bmod 6.$$

Let's consider $\langle \mathbb{Z}_n, \cdot \rangle$. Does it satisfy the axioms? We've constructed it in such a way that it is obviously closed, and associativity works pretty much like it did for addition. Also, the multiplicative identity is just 1, and $\cdot_n$ is commutative. However, what about inverses? I claim that they don't always exist. For example, in $\mathbb{Z}_6$, 2 has no multiplicative inverse:

$$2 \cdot 2 = 4 \equiv 4 \bmod 6$$

$$2 \cdot 3 = 6 \equiv 0 \bmod 6$$
$$2 \cdot 4 = 8 \equiv 2 \bmod 6$$
$$2 \cdot 5 = 10 \equiv 4 \bmod 6$$

We could also phrase this observation in terms of solving equations: the fact that 2 has no multiplicative inverse is equivalent to saying that there is no $x \in \mathbb{Z}_6$ such that

$$2x \equiv 1 \bmod 6.$$

This leads naturally to a couple of questions. Which elements do have inverses? Once we know that something has an inverse, how do we find it? In the case of $\mathbb{Z}_6$, it turns out that only 1 and 5 have inverses. What is so special about 1 and 5 in this example? To help you out, here's another example: the invertible elements of $\mathbb{Z}_{12}$ are,

$$\{1, 5, 7, 11\}.$$

and in $\mathbb{Z}_5$,

$$\{1, 2, 3, 4\}$$

make up the invertible elements. What is special in each case? The key fact is that the invertible elements in $\mathbb{Z}_n$ are exactly the ones that share no divisors (other than 1) with n. It turns out that this actually characterizes the invertible elements in $\mathbb{Z}_n$. In other words, we're about to write down a theorem that will tell us, once and for all, which elements of $\mathbb{Z}_n$ have multiplicative inverses. In order to state this in the most concise way, let's write down some new language first.

Definition 1.3.14. Let $a, b \in \mathbb{Z}$.

- The **greatest common divisor** of a and b, denoted by $\gcd(a, b)$, is the largest positive integer that divides both a and b.
- We say that a and b are **relatively prime** if $\gcd(a, b) = 1$.

This definition tells us that in order to determine whether a number a is invertible in $\langle \mathbb{Z}_n, \cdot \rangle$, we need to check that $\gcd(a, n) = 1$. It would therefore be nice to have an efficient way of computing greatest common divisors. Indeed we have a tool, called the **Euclidean algorithm**, which makes it easy to find $\gcd(a, n)$, and also to compute the inverse itself.

1.3.1 The Euclidean Algorithm

We are going to introduce the Euclidean algorithm for the purpose of finding inverses. However it is extremely useful in many contexts, and we will use it when we begin investigating properties of certain kinds of groups. Because of this, we will

spend some time becoming comfortable with it so that we have it at our disposal from now on.

As mentioned above, the Euclidean algorithm is first and foremost a method for finding the greatest common divisor of two integers. Say, for example, that I want to find the gcd of 24 and 15. You may be able to spot right away that it is 3, but it won't be so easy when the numbers are larger. The Euclidean algorithm will be quite efficient, even when the numbers are big. Let's try something and see where it leads. We begin by dividing 15 into 24:

$$24 = 1 \cdot 15 + 9.$$

Now take the remainder and divide it into 15:

$$15 = 1 \cdot 9 + 6.$$

Divide the current remainder into the previous one,

$$9 = 1 \cdot 6 + 3,$$

and repeat:

$$6 = 2 \cdot 3 + 0.$$

Now stop. When we get a remainder of zero, we stop, and we look back at the previous remainder. In this case, the last nonzero remainder is 3, which happens to be the gcd that we were seeking. It turns out that this works in general—here is the general procedure for the Euclidean algorithm that summarizes what we have just done.

Euclidean algorithm: Given integers n and m (suppose that $m < n$)[a], $\gcd(n, m)$ can be computed as follows:

1. Divide m into n, and use the Division Algorithm to write
$$n = q_0 m + r_0.$$
2. Now divide the remainder r_0 into m, and apply the Division Algorithm again to write
$$m = q_1 r_0 + r_1,$$
i.e., to obtain a new quotient and remainder.
3. Continue this process—divide the current remainder into the previous one using the Division Algorithm.
4. Stop when you obtain a remainder of zero. The previous (nonzero) remainder is $\gcd(n, m)$.

[a] If $n = m$, then we know what the greatest common divisor is. Therefore it is safe to assume that one of the integers is bigger than the other.

Example 1.3.15. Let's try this out on some larger numbers, where it is harder to see what the gcd might be. Let's try to find $\gcd(105, 81)$, say. We divide each remainder into the previous divisor:

$$\begin{aligned} 105 &= 1 \cdot 81 + 24 \\ 81 &= 3 \cdot 24 + 9 \\ 24 &= 2 \cdot 9 + 6 \\ 9 &= 1 \cdot 6 + 3 \\ 6 &= 2 \cdot 3 + 0 \end{aligned}$$

We've reached a remainder of 0, so we stop. The last nonzero remainder is 3, so

$$\gcd(105, 81) = 3.$$

Example 1.3.16. Now try to find $\gcd(343, 210)$.

Solution. Run through the Euclidean algorithm until we hit 0:

$$\begin{aligned} 343 &= 1 \cdot 210 + 133 \\ 210 &= 1 \cdot 133 + 77 \\ 133 &= 1 \cdot 77 + 56 \\ 77 &= 1 \cdot 56 + 21 \\ 56 &= 2 \cdot 21 + 14 \\ 21 &= 1 \cdot 14 + 7 \\ 14 &= 2 \cdot 7 + 0 \end{aligned}$$

Then we see that $\gcd(343, 210) = 7$. □

We mentioned that the Euclidean algorithm not only lets us find gcds, and thus show that a number is invertible in $\mathbb{Z}_n$, but that it can actually help us find the inverse. Note that if we worked backwards in our examples, we could write the gcd as a linear combination (with integer coefficients) of the two numbers:

$$\begin{aligned} \gcd(24, 15) &= 3 \\ &= 1 \cdot 9 - 1 \cdot 6 \\ &= 2 \cdot 9 - 1 \cdot 15 \\ &= 2 \cdot (24 - 1 \cdot 15) - 1 \cdot 15 \\ &= \boxed{2 \cdot 24 - 3 \cdot 15} \end{aligned}$$

We simply work our way up the list of remainders, collecting like terms until we reach the two integers that we started with. Similarly, in our second example we would have

$$\gcd(105, 81) = 3$$

$$\begin{aligned}
&= 1 \cdot 9 - 1 \cdot 6 \\
&= 1 \cdot 9 - (24 - 2 \cdot 9) \\
&= 3 \cdot 9 - 1 \cdot 24 \\
&= 3 \cdot (81 - 3 \cdot 24) - 1 \cdot 24 \\
&= 3 \cdot 81 - 10 \cdot 24 \\
&= 3 \cdot 81 - 10(105 - 1 \cdot 81) \\
&= \boxed{13 \cdot 81 - 10 \cdot 105}
\end{aligned}$$

Hopefully you get the idea, and you could work out the third example.

Exercise 1.2. Check that

$$\gcd(343, 210) = 18 \cdot 210 - 11 \cdot 343.$$

Through these examples, we have (in essence) observed that the following theorem holds. It is sometimes called the *extended* Euclidean algorithm, or Bézout's lemma.

Theorem 1.3.17. *Let $n, m \in \mathbb{Z}$. There exist integers x and y such that*

$$\gcd(n, m) = nx + my.$$

Using this theorem, it's not hard to determine precisely which elements of $\mathbb{Z}_n$ have multiplicative inverses. This also brings us to our first proof.

Theorem 1.3.18. *An element $a \in \mathbb{Z}_n$ has a multiplicative inverse if and only if* $\gcd(a, n) = 1$.

Proof. Suppose that $\gcd(a, n) = 1$. Then by Bézout's lemma, there are integers x and y such that

$$ax + ny = 1.$$

Then, since $ax + ny - ax = ny$ is a multiple of n,

$$\begin{aligned}
ax &\equiv ax + ny \bmod n \\
&\equiv 1 \bmod n.
\end{aligned}$$

This implies that a is invertible. What's the inverse? Why, it's $[x]_n$, since

$$a \cdot_n [x]_n = [a \cdot [x]_n]_n = 1.$$

On the other hand, suppose a is invertible. Then there exists an integer x so

$$ax \equiv 1 \bmod n.$$

But this means that $[a \cdot x]_n = 1$, i.e.,

$$ax = qn + 1,$$

or

$$ax - qn = 1.$$

Since $\gcd(a, n)$ divides both a and n, it divides $ax - qn$, so it divides 1. Since the gcd is always positive, it must be 1. □

As advertised, this theorem tells us not only which elements are invertible, but how to explicitly compute the inverse.

Example 1.3.19. We mentioned before that 3 is invertible in $\mathbb{Z}_5$. What is its inverse? We use the Euclidean algorithm:

$$\begin{aligned} 5 &= 1 \cdot 3 + 2 \\ 3 &= 1 \cdot 2 + 1 \\ 2 &= 2 \cdot 1 + 0 \end{aligned}$$

so of course $\gcd(5, 3) = 1$. We also have

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

so the inverse of 3 is 2.

To summarize this discussion on $\mathbb{Z}_n$ and the Euclidean algorithm, we have shown that every element of the set

$$\mathbb{Z}_n^\times = \{a \in \mathbb{Z}_n : \gcd(a, n) = 1\}$$

has a multiplicative inverse, so $\langle \mathbb{Z}_n^\times, \cdot \rangle$ satisfies the group axioms. With all of these preliminaries out of the way, we are now in a position to give the formal definition of a group in the next chapter, where we'll also begin to investigate some more interesting examples.

Additional Exercises for Chapter 1

1.3. Find $\gcd(a, b)$ and express $\gcd(a, b)$ in the form $ma + nb$ for:

(a) $(116, -84)$

(b) $(85, 65)$

(c) $(72, 26)$

(d) $(72, 25)$

1.4. Verify that the following elements of $\langle \mathbb{Z}_n, \cdot \rangle$ are invertible, and find their multiplicative inverses.

(a) 4 in $\mathbb{Z}_{15}$

(b) 14 in $\mathbb{Z}_{19}$

Chapter 2

Group Theory

This chapter contains the bulk of the material for the first half of the course. We will study groups and their properties, and we will try to answer questions regarding the structure of groups. In particular, we will try to determine when two groups are the same, and when and how a group can be built out of smaller groups. Along the way we will encounter several new and interesting examples of groups.

2.1 Definitions and Examples of Groups

Now that we have given some motivation for the study of groups, and we have gotten some preliminaries out of the way, it is time to nail down the formal definition of a group. We are headed for more abstraction—we take things like addition and multiplication for granted, and we want to be able to talk about more general types of operations on other sets. The reason for doing so is really twofold—by working in the most general setting, we can obtain results about many different objects at once. That is, results about abstract groups filter down to all of the examples that we know (and the ones we don't know yet). Once we've established the abstract definition of a group, we will first see how our motivating examples fit into the framework of group theory, and then we'll start discussing new examples of groups. Before we can write down this definition, however, we need to talk about *binary operations*, which will play the role of addition or multiplication in our groups.

2.1.1 Binary Operations

So far in our motivating examples we've talked about some sets endowed with "operations" that satisfy certain nice properties. The reason we did this was to motivate the definition of a group. At times in this discussion we have been a little vague, and we have waved our hands more than once. This will come to an end now. Since we are learning to think like mathematicians, we need to learn to be precise when we are defining things. Therefore, we're going be very careful in defining groups.

The first thing we need to make precise is this notion of "operation." We haven't been precise at all so far, and we've only given *examples* of the sorts of operations that we're looking for. Let's give a definition of an abstract type of operation which will generalize addition and multiplication.

Definition 2.1.1. Let S be a set. A **binary operation** on S is a mapping

$$* : S \times S \to S,$$

which we will usually denote by $*(a, b) = a * b$.

Let's try to dissect this definition a little. First of all, "binary" simply indicates that the operation takes in *two* elements of S and spits out a new element. Also, we've written $*$ as a **function** from $S \times S$ to S, which means two things in particular:

1. The operation $*$ is **well-defined**: given $a, b \in S$, there is exactly one $c \in S$ such that $a * b = c$. In other words, the operation is defined for *all* ordered pairs, and there is no ambiguity in the meaning of $a * b$.

2. S is **closed** under $*$: for all $a, b \in S$, $a * b$ is again in S.

This tells us two things that we need to keep an eye out for when checking that a mapping $S \times S \to S$ is a binary operation. With that in mind, let's talk about some examples (and nonexamples).

Example 2.1.2. Here are some examples of binary operations.

- Addition and multiplication on $\mathbb{Z}$ are binary operations.
- Addition and multiplication on $\mathbb{Z}_n$ are binary operations.
- Addition and multiplication on $\mathrm{M_n}(\mathbb{R})$ are binary operations.

The following are nonexamples.

- Define $*$ on $\mathbb{R}$ by $a * b = a/b$. This is not a binary operation, since it is not defined everywhere. In particular, $a * b$ is undefined whenever $b = 0$.
- Define $*$ on $\mathbb{R}$ by $a * b = c$, where c is some number larger than $a + b$. This is not well-defined, since it is not clear exactly what $a * b$ should be. This sort of operation is fairly silly, and we will rarely encounter such things in the wild. It's more likely that the given set is not closed under the operation.
- Define $*$ on $\mathbb{Z} - \{0\}$ by $a * b = a/b$. Then $*$ is not a binary operation, since the ratio of two integers need not be an integer.

- Define $*$ on $\mathbb{R}^n$ by $v * w = v \cdot w$, the usual dot product. This is not a binary operation, because $v \cdot w \notin \mathbb{R}^n$ (unless $n = 1$).
- Recall that $\mathrm{GL_n}(\mathbb{R})$ is the set of invertible $n \times n$ matrices with coefficients in $\mathbb{R}$. Define $*$ on $\mathrm{GL_n}(\mathbb{R})$ by $A * B = A + B$, matrix addition. This is not a binary operation, since $\mathrm{GL_n}(\mathbb{R})$ is not closed under addition. For example, if $A \in \mathrm{GL_n}(\mathbb{R})$, then so is $-A$, but $A + (-A) = 0$ is clearly not invertible.
- On the other hand, matrix multiplication *is* a binary operation on $\mathrm{GL_n}(\mathbb{R})$. Recall from linear algebra that the determinant is multiplicative, in the sense that
$$\det(AB) = \det(A)\det(B).$$
Therefore, if A and B both have nonzero determinant, then so does AB, and $\mathrm{GL_n}(\mathbb{R})$ is closed under multiplication. You may even recall the formula for the inverse of a product: $(AB)^{-1} = B^{-1}A^{-1}$.

Exercise 2.1. Define $H = \{n^2 : n \in \mathbb{Z}^+\}$ to be the set of all perfect squares in $\mathbb{Z}$. Determine whether the usual addition and multiplication of integers give binary operations on H.

We've probably spent enough time talking about simple examples of binary operations. It's time to talk about some of the useful properties that will be desirable for our operations to have.

Definition 2.1.3. A binary operation $*$ on a set S is **commutative** if
$$a * b = b * a$$
for all $a, b \in S$.

Example 2.1.4. Let's determine whether some of our known examples of binary operations are actually commutative.

1. $+$ and $\cdot$ on $\mathbb{Z}$ and $\mathbb{Z}_n$ are commutative.
2. Matrix multiplication is not commutative (on both $\mathrm{M_n}(\mathbb{R})$ and $\mathrm{GL_n}(\mathbb{R})$).

As we've pointed out already, commutativity will be "optional" when we define groups. The really important property that we would like to have is **associativity**.

Definition 2.1.5. A binary operation $*$ on a set S is **associative** if
$$(a * b) * c = a * (b * c)$$
for all $a, b, c \in S$.

Why is associativity important to have around? We often take it for granted in $\mathbb{Z}$, but how about other settings? Well, a binary operation only lets you combine *two* elements of a set. What if you wanted to combine three? Say we have three elements $a, b, c \in S$, and we want to write down something of the form

$$a * b * c.$$

Of course if $*$ is simply addition of integers, there is no need to worry. However, in general there are two possible ways to do this. We could compute $a * b$ and then multiply by c,

$$(a * b) * c,$$

or we could multiply b and c, and then multiply by a,

$$a * (b * c).$$

Both are reasonable definitions of $a*b*c$. It would be nice if they agreed, and that's why we'll require associativity.

Example 2.1.6. The following are examples of associative (and nonassociative) binary operations.

1. $+$ and $\cdot$ on $\mathbb{Z}$ (and $\mathbb{Z}_n$) are associative.
2. Matrix multiplication is associative.
3. Subtraction on $\mathbb{Z}$ is a binary operation, but it is not associative. For example,

   $$(3 - 5) - 1 = -2 - 1 = -3,$$

   while

   $$3 - (5 - 1) = 3 - 4 = -1.$$

4. The **cross product** on $\mathbb{R}^3$ is a binary operation, since it combines two vectors to produce a new vector in $\mathbb{R}^3$. However, it is not associative, since[1]

   $$a \times (b \times c) = (a \times b) \times c - b \times (c \times a).$$

5. (Composition of functions) Let S be a set, and define

   $$\mathcal{F}(S) = \{\text{functions } f : S \to S\}.$$

   Define a binary operation on $\mathcal{F}(S)$ by composition: given $f, g \in \mathcal{F}(S)$, define

   $$f * g = f \circ g.$$

[1] This formula is just about the next best thing to associativity. It says that the cross product satisfies something called the *Jacobi identity*, which is important in the study of algebraic objects known as *Lie algebras*.

This is indeed a binary operation, since the composition of two functions is again a function $S \to S$, i.e., it is in $\mathcal{F}(S)$. Is it associative? We need to show that if $f, g, h \in \mathcal{F}(S)$, then $(f \circ g) \circ h = f \circ (g \circ h)$. To show that two functions are equal, we need to show that their values at every element of S are equal. For any $x \in S$, we have

$$(f \circ g) \circ h(x) = (f \circ g)(h(x)) = f(g(h(x)))$$

and

$$f \circ (g \circ h)(x) = f((g \circ h)(x)) = f(g(h(x))).$$

In other words, $(f \circ g) \circ h(x) = f \circ (g \circ h)(x)$ for all $x \in S$, so $(f \circ g) \circ h = f \circ (g \circ h)$, and composition of functions is associative.

2.1.2 Groups

At this point we've surely driven the concept of associativity home, so let's get on with defining groups in a precise way. As we've mentioned before, the benefit of working in such generality is the fact that we will be able to unify all of our examples under one umbrella. We can prove results about many different examples at once, rather than having to consider many different cases.[2]

Definition 2.1.7. A **group** is a set G equipped with a binary operation $* : G \times G \to G$ satisfying the following axioms.

1. **Associativity:** For all $a, b, c \in G$, we have
$$a * (b * c) = (a * b) * c.$$
2. **Identity:** There exists an element $e \in G$ with the property that
$$e * a = a * e = a$$
for all $a \in G$.
3. **Inverses:** For every $a \in G$, there is an element $b \in G$ with the property that
$$a * b = b * a = e.$$

To be concise, we'll often write $\langle G, * \rangle$ to distinguish the operation on G. If the operation is understood, we'll just write G for the group.

Given an element $a \in G$, is not difficult to show the element b in the third axiom is *unique*. Therefore, we will usually write it as a^{-1}, and call it the **inverse** of a.

[2] In Saracino's words, we have an "economy of effort."

Exercise 2.2. Show that inverses in a group are unique. That is, for each $a \in G$, prove that there is *exactly one* element $b \in G$ satisfying $a * b = b * a = e$.

Remark 2.1.8. This seems like a good place to make some remarks about how one should read math. Reading a math book is an active process—you need to stop and think about the material frequently, and it helps to have a pen and paper handy to check things on your own. In particular, when you come across a definition, it is *extremely* helpful to think about two things:

- Immediately try to think of examples that satisfy that definition.
- Think about why the definition is useful. Why are we bothering to make this definition, and why does it say what it says?

We'll address both of these points presently.

Once we start investigating more complex ideas involving groups, it will be nice to have some examples of groups to fall back on. We've already seen some in our motivating discussion, and we'll add in some others that may also be familiar (or perhaps less familiar).

Example 2.1.9. Here are some examples of groups.

1. $\langle \mathbb{Z}, + \rangle$ is a group, as we have already seen.
2. $\langle \mathrm{M_n}(\mathbb{R}), + \rangle$ is a group.
3. $\langle \mathbb{Z}_n, +_n \rangle$ is a group.

Here are some nonexamples.

4. $\langle \mathbb{Z}, \cdot \rangle$ is *not* a group, since multiplicative inverses do not always exist. However, we could try to restrict our attention to the set of elements that do have multiplicative inverses. Indeed, $\langle \{1, -1\}, \cdot \rangle$ is a group. (We do need to be careful here—the restriction of a binary operation to a smaller set need not be a binary operation, since the set may not be closed under the operation. However, $\{1, -1\}$ is definitely closed under multiplication, so we indeed have a group.)
5. $\langle \mathrm{M_n}(\mathbb{R}), \cdot \rangle$ is not a group, since inverses may fail to exist. However, $\langle \mathrm{GL_n}(\mathbb{R}), \cdot \rangle$ is a group. We already saw that it is closed under matrix multiplication, and that the other axioms hold as well.
6. $\langle \mathbb{Z}_n, \cdot_n \rangle$ is not a group, again because inverses fail. However, $\langle \mathbb{Z}_n^\times, \cdot_n \rangle$ will be a group. Again, we just need to verify closure: if $a, b \in \mathbb{Z}_n^\times$, then a and b are both relatively prime to n. But then neither a nor b shares any prime divisors with n, so ab is also relatively prime to n. Thus $ab \in \mathbb{Z}_n^\times$.

Here are some other examples that we have not discussed yet, but should be familiar to you nonetheless.

Example 2.1.10. 1. Just as $\mathbb{Z}$ is a group under addition, so are

$$\mathbb{Q} = \{\text{rational numbers}\}$$
$$\mathbb{R} = \{\text{real numbers}\}$$
$$\mathbb{C} = \{\text{complex numbers}\}.$$

These are all groups under addition, but not multiplication. Since $\mathbb{Z}$ is only a group under addition, we'll usually write $\mathbb{Z}$ for the additive group $\langle \mathbb{Z}, + \rangle$ (and likewise for $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$), with the understanding that the operation is addition.

2. Let $\mathbb{Q}^\times = \mathbb{Q} - \{0\} = \{a/b \in \mathbb{Q} : a \neq 0\}$, and consider $\langle \mathbb{Q}^\times, \cdot \rangle$. We claim that this is a group. Multiplication is certainly associative, and the identity is 1. Given a rational number $a/b \in \mathbb{Q}^\times$, the inverse is just b/a:

$$\frac{a}{b} \cdot \frac{b}{a} = 1.$$

Similarly, we could define $\mathbb{R}^\times = \mathbb{R} - \{0\}$ and $\mathbb{C}^\times = \mathbb{C} - \{0\}$, and these would be groups under multiplication. (The inverse of a nonzero element a is simply $a^{-1} = 1/a$ in both cases.)

3. Let $\mathbb{R}^n = \{(a_1, \ldots, a_n) : a_1, \ldots, a_n \in \mathbb{R}\}$, equipped with the operation of coordinatewise addition:

$$(a_1, \ldots, a_n) + (b_1, \ldots, b_n) = (a_1 + b_1, \ldots, a_n + b_n).$$

Then $\langle \mathbb{R}^n, + \rangle$ is a group. Associativity follows from that for addition of real numbers, the identity is the zero vector, and the inverse of a vector is just its negative. More generally, any vector space is a group—we simply "forget" about scalar multiplication and view it as a group under addition.

We will introduce many more examples of groups as we move through the course. However, now that we've seen some basic examples, let's address the second point that we made above. Why does the definition of a group look the way it does? We want to come up with a general sort of object that subsumes the objects we're used to (such as number systems, vector spaces, and matrices), so the "set with a binary operation" part makes sense. We already talked about why associativity is important. But why do we require that there be an identity and inverses? Well, it all comes down to solving equations! (After all, that is what "algebra" originally meant.) Suppose we are given a group G, and we write down the equation

$$a * x = b,$$

for $a, b, x \in G$. How could we solve for x? Multiply on the left by a^{-1}:

$$\begin{aligned} a^{-1} * (a * x) = a^{-1} * b &\implies (a^{-1} * a) * x = a^{-1} * b \quad \text{(associativity)} \\ &\implies e * x = a^{-1} * b \quad \text{(inverses)} \\ &\implies x = a^{-1} * b \quad \text{(identity)} \end{aligned}$$

In short, groups are algebraic structures, and it is good to be able to solve equations within them.

We'll often break our study of groups down into different collections of groups that satisfy certain properties. Therefore, let's introduce a couple of adjectives for describing groups. We mentioned before that commutativity is "optional" when it comes to groups. Groups that have commutative operations are special, and therefore we have a special name for them.

Definition 2.1.11. A group $\langle G, * \rangle$ is said to be **abelian** if $*$ is commutative, i.e.

$$a * b = b * a$$

for all $a, b \in G$. If a group is not commutative, we'll say that it is **nonabelian**.

Abelian groups are named in honor of Niels Henrik Abel, who is mentioned in the historical discussion at the beginning of these notes. Except for $\mathrm{GL_n}(\mathbb{R})$, all the groups we've seen so far are abelian. Soon we'll see two interesting examples of nonabelian groups, the *symmetric* and *dihedral* groups.

Another useful way to describe groups is via their size. Rather than saying "size" or "cardinality", we speak of the *order* of a group.

Definition 2.1.12. The **order** of a group G, denoted by $|G|$, is the number of elements in G.

If a group G has infinitely many elements, we will write $|G| = \infty$[3] and say that G has **infinite order**.

Example 2.1.13. Most of the examples we have seen so far are infinite groups. In particular, $|\mathbb{Z}| = \infty$.

We already know lots of examples of infinite groups, but such groups are actually quite hard to understand from a structural standpoint (with the exception of $\mathbb{Z}$). We'll be more interested in **finite groups**.

[3]Here we are not worrying about the true cardinality of the group as a set. We only care whether the group has finitely many elements or not, and if it does, we define $|G|$ to be the number of elements in G.

Definition 2.1.14. A group G is said to be **finite** if $|G| < \infty$.

One of the reasons that we study finite groups is that it is much easier to analyze their structure, and we'll be able to classify many types of finite groups (meaning we can organize them into families with similar properties). Therefore, we'd like to have lots of examples at our disposal. Of course, we've already seen two examples of finite groups, namely the groups that arise from the study of modular arithmetic.

Example 2.1.15.

1. For any n, the additive group $\langle \mathbb{Z}_n, +_n \rangle$ is a finite (abelian) group, with $|\mathbb{Z}_n| = n$. We will usually suppress the $+_n$ and simply write $\mathbb{Z}_n$ for this group.

2. Similarly, $\langle \mathbb{Z}_n^\times, \cdot \rangle$ is a finite (abelian) group. Its order is a little harder to determine. We know that $|\mathbb{Z}_n^\times|$ will be the number of elements $a \in \mathbb{Z}_n$ that are relatively prime to n. If you've taken a class in number theory, then you know that this number is given by the *Euler phi function*:

$$|\mathbb{Z}_n^\times| = \varphi(n).$$

   If you haven't seen φ, don't worry about it. We will discuss it later when we need it.

We will produce two more families of finite groups very soon, namely the symmetric and dihedral groups to which we have already alluded.

2.1.3 Group Tables

Now we'll make a slight detour to talk about a tool for working with finite groups. One of the things that makes finite groups easier to handle is that we can write down a table that completely describes the group—we simply list the elements out and multiply "row by column."

Example 2.1.16. Let's look at $\mathbb{Z}_3$, for example. We'll write down a "multiplication table" that tells us how the group operation works for any pair of elements. As we mentioned, each entry is computed as "row times column":

$+$	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

(Of course we have to remember that "times" really means "plus" in this example.) This is called a **group table** (or a **Cayley table**).

Group tables can get pretty tedious when the order of the group is large, but they are handy for analyzing fairly small groups. In particular, let's look at groups of order 1, 2, and 3. If we have a group of order one, there is really only one thing to do. The lone element must be the identity e, and $e * e = e$, so the group table must be:

$*$	e
e	e

Suppose next that we have a group of order two, so that the group contains the identity e and some other element a with $a \neq e$. Then $e * a = a * e = a$, and since a must have an inverse (and the only option is a), we must have $a * a = e$. Therefore, the group table must look like:

$*$	e	a
e	e	a
a	a	e

Finally, suppose we have a group of order three. Then there are three distinct elements: the identity e, and two other elements a and b. The first row and first column are easy to fill in, since e is the identity. We are then left to determine $a * a$, $a * b$, $b * a$, and $b * b$. We claim first that it is impossible to have $a * a = e$. If $a * a = e$, then $b * b = e$ as well, since b must have an inverse. (We can't have $b = a^{-1}$, since inverses are unique.) But then we must have either $a * b = a$ or $a * b = b$. In the first case,

$$b = e * b = (a * a) * b = a * (a * b) = a * a = e,$$

which contradicts the fact that group has three elements. Similar reasoning shows that $a * b \neq b$. The hypothesis that got us into all this trouble was that $a * a = e$, so we can't have $a = a^{-1}$. Also, $a * a \neq a$; if it were, then

$$e = a^{-1} * a = a^{-1} * (a * a) = (a^{-1} * a) * a = e * a = a,$$

which is not true. Therefore, we must have $a * a = b$ instead. Since a must have an inverse (and it must be b),

$$a * b = b * a = e.$$

Finally, $b * b = a$, since

$$b * b = b * (a * a) = (b * a) * a = e * a = a.$$

Consequently, the group table must be:

$*$	e	a	b
e	e	a	b
a	a	b	e
b	b	e	a

In particular, we have shown that for groups of order 3 or less, there is really only one group of each order, since there was only one way that we could fill in each group table. (In language that we'll develop later, there is only one group of each of these orders *up to isomorphism.*) Also, any group of order 1, 2, or 3 is abelian. To see this, we can simply note that the group table is symmetric about the diagonal. It is left as an exercise to show that groups of order 4 and 5 are abelian. The result does not extend to order 6—we will see two groups of order 6 which are not abelian. As you can see, group tables can be quite useful, and we will look at them again once we've defined the symmetric and dihedral groups.

Exercise 2.3. Show that any group of order 4 is abelian. (**Hint:** Compute all possible Cayley tables. Up to a reordering of the elements, there are two possible tables.)

Exercise 2.4. Show that any group of order 5 is abelian. (**Hint:** There is only one possible Cayley table, up to relabeling.)

2.1.4 Remarks on Notation

Now we'll make one last note on working with general groups. We will get really sick of writing $*$ all the time, so we will oftentimes suppress the operation $*$, and simply write

$$ab = a * b.$$

when working with arbitrary groups. If we are dealing with a specific example (such as the integers under addition), we'll still write the operation to make things clear. In keeping with this convention, we'll also have a shorthand way of denoting a product of an element a with itself. We'll write

$$a^2 = a * a,$$

and in general

$$a^n = \underbrace{a * a * \cdots * a}_{n \text{ times}}.$$

We can also define negative powers by

$$a^{-n} = \left(a^{-1}\right)^n.$$

These notational conventions will make it much easier to write down computations when we are working with groups. The point of all this is that computations in an arbitrary group will behave very much like usual multiplication, except that

things may not always commute. (In fact, matrix multiplication might be the best operation to keep in mind.) Also, you should remember that if we are dealing with the additive group $\mathbb{Z}$, a^n really means

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

This goes to show that one has to be careful when writing down computations in a group, since we usually write things multiplicatively.

2.2 The Symmetric and Dihedral Groups

We are now ready to introduce our first really interesting examples of (finite) nonabelian groups, which are called the **symmetric** and **dihedral groups**. (These are actually *families* of groups, one for each positive integer.) Not only will they give examples of finite nonabelian groups, but we will see that their elements are quite different from the examples that we have been considering so far.

2.2.1 The Symmetric Group

The symmetric group is one of the most important examples of a finite group, and we will spend quite a bit of time investigating its properties.[4] It will arise as a special case of the set S_X of bijections from a set X back to itself, so let's begin there.

Before we can proceed, we need some preliminaries on functions. We'll need some of these facts later on when we discuss homomorphisms, so we will work a little more generally than is absolutely necessary right now.

Let X and Y be nonempty sets. Recall that a function $f : X \to Y$ is, at least informally, a rule which assigns to each element $x \in X$ a *unique* element $f(x) \in T$:

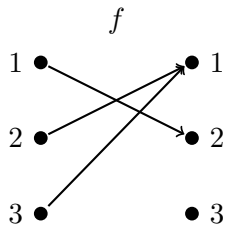
$$x \mapsto f(x).$$

We're eventually going to look at a particular class of functions from a set to itself. To do this, we'll need to know what it means for a function $f : X \to Y$ to be **one-to-one** (or **injective**) and **onto** (or **surjective**).

Definition 2.2.1. A function $f : X \to Y$ is **one-to-one** or **injective** if whenever $x_1, x_2 \in X$ with $f(x_1) = f(x_2)$, then $x_1 = x_2$. Equivalently, if $x_1 \neq x_2$, then $f(x_1) \neq f(x_2)$, or different inputs always yield different outputs.

[4] One author (Herstein) even devotes an entire chapter to it in his book.

Example 2.2.2. 1. Let $X = \{1, 2, 3\}$ (or any set with three elements), and define $f : X \to X$ by $f(1) = 2$, $f(2) = 1$, and $f(3) = 1$. We can represent this mapping pictorially as



This function is not injective, since 2 and 3 both map to 1.

2. Define $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = x^2$. This function is *not* injective, since $f(1) = f(-1) = 1$, for example.

3. On the other hand, define $g : \mathbb{R} \to \mathbb{R}$ by $g(x) = x$. This function is injective.

Definition 2.2.3. A function $f : X \to Y$ is **onto** or **surjective** if for any $y \in Y$, there exists an $x \in X$ such that $f(x) = y$.

Intuitively, f is surjective if every element of Y is "hit" by f, or lies in the image of f.

Example 2.2.4. 1. The function $f : \{1, 2, 3\} \to \{1, 2, 3\}$ described in the previous example is not onto, since no element is mapped to 3 by f.

2. The function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^2$ is not onto.

3. If we define $g : \mathbb{R} \to \mathbb{R}$ by $g(x) = x$, then this is a surjective function.

Any function that is both one-to-one and onto is of particular importance, so we have a special name for such objects.

Definition 2.2.5. A function $f : X \to Y$ is **bijective** (or simply a **bijection**) if f is both one-to-one and onto.

The proofs of the following statements would all be good exercises to try on your own, but we will reproduce them here anyway.

Proposition 2.2.6. *Let X, Y, and Z be sets, and let $f : X \to Y$ and $g : Y \to Z$ be functions.*

1. *If f and g are both one-to-one, then so is $g \circ f$.*
2. *If f and g are both onto, then so is $g \circ f$.*
3. *If f and g are both bijections, then so is $g \circ f$.*

Proof. Let's start with the first one. Assume f and g are one-to-one, and suppose $x_1, x_2 \in X$ with $g \circ f(x_1) = g \circ f(x_2)$. This means that

$$g(f(x_1)) = g(f(x_2)),$$

and since g is one-to-one, $f(x_1) = f(x_2)$. But f is also one-to-one, so $x_1 = x_2$. Thus $g \circ f$ is one-to-one.

Now let's handle the question of surjectivity. Suppose f and g are both onto, and let $z \in Z$. We need to produce an $x \in X$ such that $g \circ f(x) = z$, i.e., $g(f(x)) = z$. Since g is onto, there exists a $y \in Y$ such that $g(y) = z$. Also, f is onto, so there exists an $x \in X$ such that $f(x) = y$. Putting this together, we get

$$g \circ f(x) = g(f(x)) = g(y) = z,$$

so $g \circ f$ is onto.

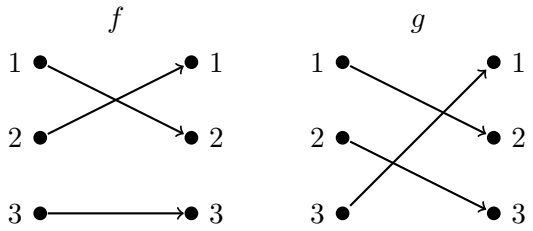
The third statement follows from the first two. If f and g are bijective, then they are both one-to-one and onto. But then $g \circ f$ is one-to-one by (a), and $g \circ f$ is onto by (b), so $g \circ f$ is bijective. □

Now let's formally define S_X, the set of bijections from a set X to itself. We will then proceed prove that S_X is a group under composition.

Definition 2.2.7. Let X be a set. We define

$$S_X = \{f : X \to X : f \text{ is a bijection}\}.$$

Note that Proposition 2.2.6(c) tells us that S_X is closed under composition of functions. In other words, composition defines a binary operation on S_X. But what kind of operation is it? Why, it's associative! We already verified this in Example 2.1.6(5). Unfortunately, it is not a commutative operation. To see this, let $X = \{1, 2, 3\}$, and define f and g by the following diagrams:



Then

$$g \circ f(1) = g(f(1)) = g(2) = 3,$$

but

$$f \circ g(1) = f(g(1)) = f(2) = 1,$$

so $g \circ f \neq f \circ g$. Thus S_X will provide a new example of a *nonabelian* group.

To finish checking that S_X is a group, we need to verify the existence of an identity and inverses. For the first one, recall that any set X has a special bijection from X to X, namely the identity function id_X:

$$\mathrm{id}_X(x) = x$$

for all $x \in X$. Note that for any $f \in S_X$, we have

$$f \circ \mathrm{id}_X(x) = f(\mathrm{id}_X(x)) = f(x)$$

and

$$\mathrm{id}_X \circ f(x) = \mathrm{id}_X(f(x)) = f(x)$$

for all $x \in S$. Thus $\mathrm{id}_X \circ f = f \circ \mathrm{id}_X = f$ for all $f \in S_X$, so id_X serves as an identity for S_X under composition. Finally, what do we know about any bijection? Why, it has an inverse, in the sense that there is another function that "undoes" it. More specifically, if $f \in S_X$ and $y \in X$, there is an $x \in X$ such that $f(x) = y$ (since f is onto). But f is also one-to-one, so this x is unique. Therefore, we can define $f^{-1}(y) = x$. You can then check that

$$f \circ f^{-1}(y) = f(f^{-1}(y)) = f(x) = y = \mathrm{id}_X(y)$$

and

$$f^{-1} \circ f(x) = f^{-1}(f(x)) = f^{-1}(y) = x = \mathrm{id}_X(x),$$

so f^{-1} really is an inverse for f under composition. Therefore, by making all of these observations, we have established the following result:

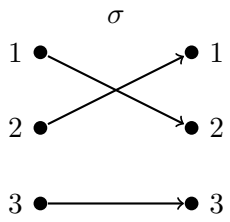
Proposition 2.2.8. *For any set X, S_X forms a group under composition of functions.*

If X is an infinite set, then S_X is fairly hard to understand. One would have to be quite brave to try to work with it. Things are much more tractable (and interesting) when X is finite. Suppose then that X is finite, say with n elements. It doesn't really matter what X is; it only matters that X has n elements. That is, we can label the elements of X to be whatever we want (say, numbers, people, flavors of ice cream, etc.), so we could simply assume that X is the set

$$X = \{1, 2, \ldots, n\}.$$

In this case we call S_X the **symmetric group on n letters**, and we usually denote it by S_n.

An element of S_n is a bijection from $\{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$. In other words, it rearranges the numbers $1, \ldots, n$. Pictorially, we could represent the bijection σ of $\{1, 2, 3\}$ defined by $\sigma(1) = 2$, $\sigma(2) = 1$, and $\sigma(3) = 3$ with the diagram[5]



What we're really trying to get at here is that an element of S_n can be thought of as a **permutation** of the numbers 1 through n.

Definition 2.2.9. S_n is the set of all permutations of the set $\{1, 2, \ldots, n\}$.

How many permutations of $\{1, \ldots, n\}$ are there? In order to define a permutation σ of $\{1, \ldots, n\}$, we need to determine where to send each number. There are n choices for $\sigma(1)$, and there are $n-1$ choices for $\sigma(2)$. There are $n-2$ choices for $\sigma(3)$, and so on, until we reach $\sigma(n)$, for which we only have one choice. In other words, we have observed that the total number of permutations of $\{1, \ldots, n\}$ is

$$n(n-1)(n-2)\cdots 2 \cdot 1 = n!.$$

Phrased in the language of group theory, we have shown that

$$|S_n| = n!.$$

One of the nice things about working with permutations on finite sets is that we can completely describe what they do. However, it would be annoying if we had

[5] Algebraists usually represent elements of the symmetric group by Greek letters, such as σ and τ, so we'll slip into that usage now as well. Note that the permutation that we are calling σ here is the same as the bijection that we called f earlier.

to specify $\sigma(1), \sigma(2), \ldots, \sigma(n)$ separately each time, so we'll introduce a concise notation for representing a permutations. In particular, we can write down a table that completely describes a given permutation, called **two-line notation**:

$$\sigma = \begin{pmatrix} 1 & 2 & \cdots & n \\ \sigma(1) & \sigma(2) & \cdots & \sigma(n) \end{pmatrix}$$

For example, suppose $\sigma \in S_3$ is given by the picture that we considered earlier, i.e., $\sigma(1) = 2$, $\sigma(2) = 1$, and $\sigma(3) = 3$. Then we have

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

Of course if we are going to represent permutations in this way, it would help to know how multiplication works in this notation. As an example, let

$$\tau = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}$$

Then remember that multiplication is really just composition of functions:

$$\begin{aligned} \sigma\tau &= \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}\begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 2 & 3 \\ \sigma(\tau(1)) & \sigma(\tau(2)) & \sigma(\tau(3)) \end{pmatrix} \\ &= \begin{pmatrix} 1 & 2 & 3 \\ \sigma(2) & \sigma(3) & \sigma(1) \end{pmatrix} \\ &= \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix} \end{aligned}$$

On the other hand, what is $\tau\sigma$?

$$\tau\sigma = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}$$

In other words, one moves right to left when computing the product of two permutations. First one needs to find the number below 1 in the rightmost permutation, then find this number in the top row of the left permutation, and write down the number directly below it. Then repeat this process for the rest of the integers 2 through n.

In the example above, note that $\sigma\tau \neq \tau\sigma$, so what have we shown? We have actually verified that S_3 is nonabelian. Indeed, this holds for all larger symmetric groups as well.

Proposition 2.2.10. *For $n \geq 3$, S_n is a nonabelian group.*

Proof. Let $\sigma, \tau \in S_3$ be defined as in the example, and suppose that $n > 3$. Define $\tilde{\sigma}, \tilde{\tau} \in S_n$ by

$$\tilde{\sigma}(i) = \begin{cases} \sigma(i) & \text{if } 1 \leq i \leq 3 \\ i & \text{if } i > 3, \end{cases}$$

and similarly for $\tilde{\tau}$. Then the computation that we performed in S_3 shows that $\tilde{\sigma}\tilde{\tau} \neq \tilde{\tau}\tilde{\sigma}$, so S_n is nonabelian. □

It would be entirely feasible to compute the group table for S_n, at least for relatively small n. We'll display the group table for S_3 here, though we won't perform any of the computations explicitly. (There are $6 \cdot 6 = 36$ different products to compute, and these are left to the interested reader.) Label the elements of S_3 as follows:

$$\iota = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix} \qquad \mu_1 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}$$

$$\rho_1 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix} \qquad \mu_2 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}$$

$$\rho_2 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} \qquad \mu_3 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}$$

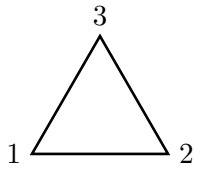
Then the group table for S_3 is:

$*$	ι	ρ_1	ρ_2	μ_1	μ_2	μ_3
ι	ι	ρ_1	ρ_2	μ_1	μ_2	μ_3
ρ_1	ρ_1	ρ_2	ι	μ_3	μ_1	μ_2
ρ_2	ρ_2	ι	ρ_1	μ_2	μ_3	μ_1
μ_1	μ_1	μ_2	μ_3	ι	ρ_1	ρ_2
μ_2	μ_2	μ_3	μ_1	ρ_2	ι	ρ_1
μ_3	μ_3	μ_1	μ_2	ρ_1	ρ_2	ι

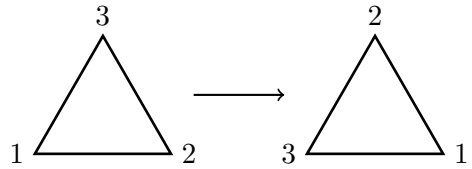
The labeling of the elements may look odd at this point, but we will see a good reason for it quite soon.

2.2.2 The Dihedral Group

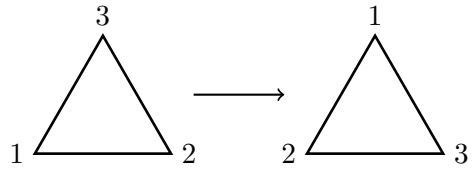
Now it's time to talk about another interesting nonabelian group, the **dihedral group**. Again, we're actually going to be dealing with a family of groups, one for each positive integer. Suppose we have a triangle, and we label the vertices 1, 2, and 3:



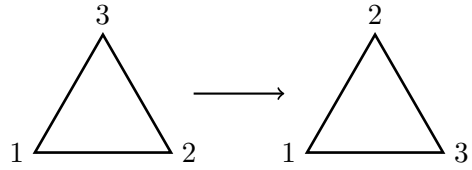
We can rotate the triangle counterclockwise by 120° and obtain a triangle with the same orientation, albeit with the vertices relabeled:



We could also rotate by 240°:



Call these two transformations r_1 and r_2, respectively. We could also reflect the triangle across any of the angle bisectors to obtain a relabeling of the vertices:



Label the reflection depicted above as m_1, and let m_2 and m_3 denote the reflections across the line bisecting angles 2 and 3 (in the original labeling of the triangle), respectively. Define the **identity transformation** i to be the one that simply leaves the triangle unmoved. The set

$$\{i, r_1, r_2, m_1, m_2, m_3\}$$

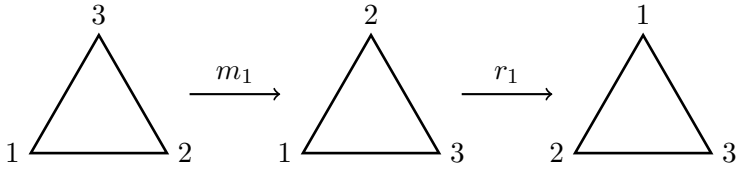
is called the set of **symmetries** of the triangle. Its elements are exactly the transformations of the triangle that reorder the vertices in a way that respects adjacency. (That is, two adjacent vertices are still adjacent after applying any of these transformations.) It is not hard to check (via brute force, if necessary) that this set forms a group under composition, which we denote by D_3 and call the **third dihedral group**. More generally, we can define the n^{th} **dihedral group** D_n to be the set of symmetries of a regular n-gon equipped with the operation of composition.

Definition 2.2.11. The group D_n is the set of all symmetries of the regular n-gon under composition of transformations.

Naturally, we could ask ourselves what the order of D_n should be. In general, there will be n rotations (including the identity transformation) and n reflections, so

$$|D_n| = 2n.$$

Of course we should talk about how to multiply two symmetries of an n-gon. Again, we need to think of multiplication as composition of functions. For example, suppose we wanted to compute $r_1 m_1$ in D_3. We would first have to apply m_1 to the triangle, and then apply r_1:



We see from this picture that $r_1 m_1 = r_2$. By doing this for all possible pairs of elements, we could write down the group table for D_3:

$*$	i	r_1	r_2	m_1	m_2	m_3
i	i	r_1	r_2	m_1	m_2	m_3
r_1	r_1	r_2	i	m_3	m_1	m_2
r_2	r_2	i	r_1	m_2	m_3	m_1
m_1	m_1	m_2	m_3	i	r_1	r_2
m_2	m_2	m_3	m_1	r_2	i	r_1
m_3	m_3	m_1	m_2	r_1	r_2	i

After switching Roman letters to Greek letters, you might notice that this table is the same as the one we constructed for S_3. (This is why we labeled the elements of S_3 in the way that we did.) We have essentially shown that S_3 and D_3 are *isomorphic groups*: they are fundamentally the same group dressed up in different disguises. In this case, it is not too hard to see how one could identify D_3 with S_3, since the elements of D_3 can be viewed as permutations of the vertices. In general, we can view elements of D_n as permutations of the vertices of the regular n-gon, but we cannot realize *all* permutations in this way when $n \geq 4$. In other words, we will see that D_n and S_n are not the same in general. Of course one way to see this at this point is to notice that $|D_n| = 2n$ but $|S_n| = n!$, and these numbers are only equal when $n = 3$.

2.3 Basic Properties of Groups

Now that we've defined groups and we have some interesting examples in our toolkit, it's time to start investigating properties of groups. We'll start of with some simple properties before we really get into the bigger questions regarding the structure and classification of groups. At the very least, this will simplify some of the routine calculations that we need to make when working with groups.

Proposition 2.3.1. *Let G be a group. The identity element $e \in G$ is* unique*, i.e., there is only one element e of G with the property that*

$$ae = ea = a$$

for all $a \in G$.

Proof. For this proof, we need to use the standard mathematical trick for proving uniqueness: we assume that there is another gadget that behaves like the one in which we're interested, and we prove that the two actually have to be the same. Suppose there is another $f \in G$ with the property that

$$af = fa = a$$

for all $a \in G$. Then in particular,

$$ef = fe = e.$$

But since e is an identity,

$$ef = fe = f.$$

Therefore,

$$e = ef = f,$$

so e is unique. □

The next result has to do with *solving equations*, which was our original motivation for requiring that inverses exist. It's called **cancellation**.

Proposition 2.3.2 (Cancellation laws)**.** *Let G be a group, and let $a, b, c \in G$. Then:*

(a) *If $ab = bc$, then $b = c$.*

(b) *If $ba = ca$, then $b = c$.*

Proof. (a) Suppose $ab = ac$. Multiply both sides on the left by a^{-1}:

$$a^{-1}(ab) = a^{-1}(ac).$$

By associativity, this is the same as

$$(a^{-1}a)b = (a^{-1}a)c,$$

and since $a^{-1}a = e$, we have

$$eb = ec.$$

Since e is the identity, $b = c$. The same sort of argument works for (b), except we multiply the equation on the right by a^{-1}. □

The cancellation laws actually give us a very useful corollary. You may have already guessed that this result holds (particularly in light of Exercise 2.2, but we will prove here that inverses in a group are unique.

Corollary 2.3.3. *Let G be a group. Every $a \in G$ has a* unique *inverse, i.e. for each $a \in G$ there is exactly one element a^{-1} with the property that*

$$aa^{-1} = a^{-1}a = e.$$

Proof. Let $a \in G$, and suppose that $b \in G$ has the property that $ab = ba = e$. Then in particular,

$$ab = e = aa^{-1},$$

and by cancellation, $b = a^{-1}$. Thus a^{-1} is unique. □

While we are on the topic of inverses, we will prove two more results about them. The first tells us what happens when we "take the inverse twice," and the second gives us a rule for determining the inverse of a product of two group elements.

Proposition 2.3.4. *If $a \in G$, then $(a^{-1})^{-1} = a$.*

Proof. By definition, $a^{-1}(a^{-1})^{-1} = e$. But $a^{-1}a = aa^{-1} = e$ as well, so by uniqueness of inverses, $(a^{-1})^{-1} = a$. □

Recall from linear algebra that if $A, B \in \mathrm{GL_n}(\mathbb{R})$, we have a formula for $(AB)^{-1}$; it is simply $B^{-1}A^{-1}$. In general there is a rule for determining the inverse of a product of two group elements, which looks just like the rule for matrices—it is the product of the inverses, but in the reverse order.

Proposition 2.3.5. *For any $a, b \in G$, $(ab)^{-1} = b^{-1}a^{-1}$.*

Proof. We'll explicitly show that $b^{-1}a^{-1}$ is the inverse of ab by computing:

$$\begin{aligned}(ab)(b^{-1}a^{-1}) &= ((ab)b^{-1})a^{-1} \\ &= (a(bb^{-1}))a^{-1} \\ &= (ae)a^{-1} \\ &= aa^{-1} \\ &= e.\end{aligned}$$

Of course we also need to check that $(b^{-1}a^{-1})(ab) = e$, which works pretty much the same way:

$$\begin{aligned}(b^{-1}a^{-1})(ab) &= b^{-1}(a^{-1}(ab)) \\ &= b^{-1}((a^{-1}a)b) \\ &= b^{-1}(eb) \\ &= b^{-1}b \\ &= e.\end{aligned}$$

Thus $(ab)^{-1} = b^{-1}a^{-1}$, since inverses are unique. □

In order to be thorough and rigorous, we needed to check that $b^{-1}a^{-1}$ was a *two-sided* inverse in the previous proof. It would be quite annoying if we had to do this all the time, and you may be wondering if there is a shortcut. There is one, which allows us to check that two group elements are inverses of each other simply by multiplying them in only one of the two possible orders. To prove it, we'll use the following proposition.

Proposition 2.3.6. *For any $a, b \in G$, the equations $ax = b$ and $xa = b$ have unique solutions in G.*

Proof. The solution to $ax = b$ is $x = a^{-1}b$, and for $xa = b$ it is $x = ba^{-1}$. These are unique since inverses are unique. □

Proposition 2.3.7. *Let G be a group, and let $a, b \in G$. If either $ab = e$ or $ba = e$, then $b = a^{-1}$.*

Proof. This really amounts to solving the equation $ax = e$ (or $xa = e$). We know from Proposition 2.3.6 that there is a unique solution, namely $x = a^{-1}e = a^{-1}$ (in either case). Therefore, if $ab = e$ or $ba = e$, then b is a solution to either $ax = e$ or $xa = e$, so $b = a^{-1}$. □

Exercise 2.5. Prove that if G is a group and $a, b \in G$ with $ab = a$, then $b = e$.

2.3.1 The Order of an Element

Recall that we've developed some shorthand notation for writing out computations in groups. For one, we have been suppressing the binary operation $*$, and we simply write ab in place of $a * b$. We also developed simpler notation for expressing powers of an element—we write

$$a^n = \underbrace{a * a * \cdots * a}_{n \text{ times}}$$

for $n \in \mathbb{Z}^+$, and negative powers are defined in terms of the inverse:

$$a^{-n} = (a^{-1})^n.$$

In addition, we define $a^0 = e$. Note that this we have written these equations for an arbitrary group—if we're dealing with an *additive* group (like $\mathbb{Z}$ or $\mathbb{Z}_n$), then we would really write a^n as

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

We will generally write things multiplicative, but we will keep this additive convention in mind when we are working with certain specific examples.

Now let G be a group, and let $a \in G$. We're going to investigate the things that we can do with powers of a. Therefore, it will be helpful to assign some notation to the set of all powers of a.

Definition 2.3.8. Given a group G and an element $a \in G$, we define

$$\langle a \rangle = \{a^n : n \in \mathbb{Z}\} = \{\ldots, a^{-2}, a^{-1}, e, a, a^2, a^3, \ldots\}.$$

The first observation that we will make is that the familiar "exponent rules" hold for an arbitrary group element.

Proposition 2.3.9. *Let G be a group, and let $a \in G$.*

(a) $a^n a^m = a^{n+m}$ *for all* $n, m \in \mathbb{Z}$.

(b) $(a^n)^{-1} = a^{-n}$ *for all* $n \in \mathbb{Z}$.

(c) $(a^n)^m = a^{nm}$ *for all* $n, m \in \mathbb{Z}$.

The proof is straightforward and makes a good exercise, so we will leave it to the reader as such.

Exercise 2.6. Prove Proposition 2.3.9.

Now we'll investigate what happens when we keep taking powers of an element. Let's use a specific example to get started.

Example 2.3.10. Consider the group $G = \mathbb{Z}_{12}$. Let's compute the "powers" of some elements. (Recall that in this group "power" really means "multiple".) We'll calculate the powers of 2 first:

$$\begin{aligned}
1 \cdot 2 &= 2 \\
2 \cdot 2 &= 2 +_{12} 2 = 4 \\
3 \cdot 2 &= 2 +_{12} 2 +_{12} 2 = 6 \\
4 \cdot 2 &= 8 \\
5 \cdot 2 &= 10 \\
6 \cdot 2 &= [12]_{12} = 0 \\
7 \cdot 2 &= [14]_{12} = 2 \\
8 \cdot 2 &= [16]_{12} = 4
\end{aligned}$$

and so on. What about powers of 3?

$$1 \cdot 3 = 3$$

$$2 \cdot 3 = 3 +_{12} 3 = 6$$
$$3 \cdot 3 = 3 +_{12} 3 +_{12} 3 = 9$$
$$4 \cdot 3 = [12]_{12} = 0$$
$$5 \cdot 3 = [15]_{12} = 3$$
$$6 \cdot 3 = [18]_{12} = 6$$

and so on. Notice that both lists repeat after a while. In particular, we reach 0 (i.e., the identity) after a certain point. We quantify this phenomenon by saying that these elements have **finite order**.

Definition 2.3.11. Let G be a group. We say that $a \in G$ has **finite order** if there exists $n \in \mathbb{Z}^+$ such that $a^n = e$. The smallest such integer is called the **order** of a, denoted by $|a|$ (or in some books as $o(a)$). If no such integer exists, we say that a has **infinite order**.

Example 2.3.12. 1. The identity element in any group has order 1.

2. In $\mathbb{Z}_{12}$, we have seen that $|2| = 6$ and $|3| = 4$.

3. In D_3, the order of the reflection m_1 is 2. (This is true of all reflections in D_3.) Also, $|r_1| = |r_2| = 3$.

4. In $\mathbb{Z}$, 5 has infinite order, as do all other elements.

We started this discussion by observing that the elements $2, 3 \in \mathbb{Z}_{12}$ have finite order. Do all the elements of $\mathbb{Z}_{12}$ have finite order? Yes—in fact, one can directly check that for all $a \in \mathbb{Z}_{12}$,

$$12 \cdot a = 0,$$

for example. How could we show that all elements have finite order without checking every single element? We need to think about the real reason that an element of $\mathbb{Z}_{12}$ must have finite order—the group is finite, so there are only so many places to put the powers of a given element. (Remember the Pigeonhole Principle?) We can quantify this observation with the following proposition.

Proposition 2.3.13. *Let G be a finite group. Then every element $a \in G$ has finite order.*

Proof. Consider the set

$$\{a^n : n \geq 0\} = \{e, a, a^2, \ldots\} \subseteq G.$$

Since G is finite, this list of powers can't be infinite. (This follows from the Pigeonhole Principle, for instance. We have an infinite list of group elements that must fit into only finitely many slots.) Therefore, two different powers of a must coincide, say $a^i = a^j$, with $j \neq i$. We can assume that $j > i$. Then

$$a^{j-i} = a^j a^{-i} = a^i a^{-i} = e,$$

so a has finite order. (In particular, $|a| \leq j - i$.) Since $a \in G$ was arbitrary, the result follows. □

Remark 2.3.14. There are two questions that we could ask here. First, you may be wondering if the converse of Proposition 2.3.13 holds. That is, if G is a group in which every element has finite order, is G necessarily finite? The answer to this question is no—there are examples of infinite groups in which every element has finite order, but we do not have the tools yet to describe them.

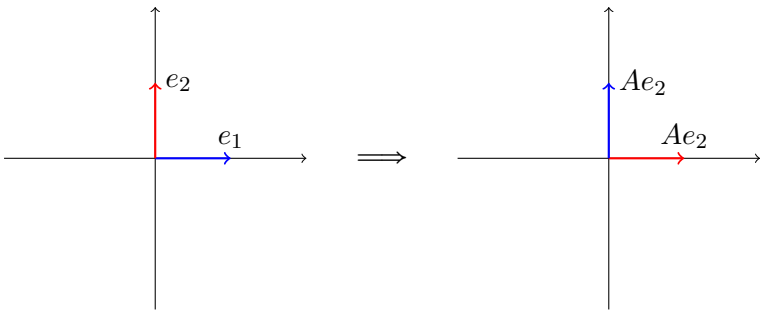
On the other hand, if G is an infinite group, can we have elements of finite order? Yes—for example, 1 and -1 have finite order in $\mathbb{Z}$. (The first one is a bit of a cheat—the identity element of a group always has order 1.) For a more interesting example, we can look to $\mathrm{GL_n}(\mathbb{R})$: if we let

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

then

$$A^2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}^2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

Thus the matrix A has order 2. It's also easy to see why this should be true if we think about the linear transformation that A represents—it reflects the plane $\mathbb{R}^2$ across the line $y = x$. That is, it interchanges the two standard basis vectors:



Let's get on with proving some more basic facts about order. First, we'll relate the order of an element to that of its inverse.

Proposition 2.3.15. *Let G be a group and let $a \in G$. Then $|a| = |a^{-1}|$.*

Proof. Suppose first that a has finite order, with $|a| = n$. Then

$$(a^{-1})^n = a^{-n} = (a^n)^{-1} = e^{-1} = e,$$

so $|a^{-1}| \leq n = |a|$. On the other hand, if we let $m = |a^{-1}|$, then

$$a^m = ((a^{-1})^{-1})^m = (a^{-1})^{-m} = ((a^{-1})^m)^{-1} = e,$$

so $n \leq m$. Thus $n = m$, or $|a| = |a^{-1}|$.

Now suppose that a has infinite order. Then for all $n \in \mathbb{Z}^+$, we have $a^n \neq e$. But then

$$(a^{-1})^n = a^{-n} = (a^n)^{-1} \neq e$$

for all $n \in \mathbb{Z}^+$, so a^{-1} must have infinite order as well. □

Our next proposition completely describes the integers m for which $a^m = e$—they are simply the multiples of $|a|$.

Proposition 2.3.16. *Let G be a group, and let $a \in G$ be an element of finite order. If $m \in \mathbb{Z}$, then $a^m = e$ if and only if $|a|$ divides m.*

Proof. Let $n = |a|$. If $n \mid m$, it is easy. Write $m = nd$ for some $d \in \mathbb{Z}$. Then

$$a^m = a^{nd} = (a^n)^d = e^d = e.$$

On the other hand, if $m \geq n$, we can use the Division Algorithm to write $m = qn+r$ with $0 \leq r < n$. Then

$$e = a^m = a^{qn+r} = a^{qn}a^r = (a^n)^q a^r = ea^r = a^r,$$

so $a^r = e$. But $r < n$, and n is the smallest positive power of a which yields the identity. Therefore r must be 0, and n divides m. □

Note that Proposition 2.3.16 tells us something more general about powers of a: when we proved that elements of finite groups have finite order, we saw that $a^i = a^j$ implied that $a^{j-i} = e$. But this means that $n = |a|$ divides $j - i$ by Proposition 2.3.16. In other words, i and j must be congruent mod n.

Proposition 2.3.17. *Let G be a group, and suppose $a \in G$ is an element of finite order n. Then $a^i = a^j$ if and only if $i \equiv j \bmod n$.*

Along the same lines, we observed that if $a^i = a^j$ with $j > i$, then $a^{j-i} = e$, so a must have finite order. Taking the contrapositive of this statement, we get the following result.

Proposition 2.3.18. *Let G be a group, and suppose $a \in G$ is an element of infinite order. Then the powers of a are all distinct. (That is, $a^i = a^j$ if and only if $i = j$.)*

Finally, if we know that a has order n, then how can we find the orders of other powers of a? It turns out that there is a nice formula in terms of the greatest common divisor.

Theorem 2.3.19. *Let G be a group and suppose $a \in G$ an element of finite order n. Then for all $m \in \mathbb{Z}$,*

$$|a^m| = \frac{n}{\gcd(m, n)}.$$

Proof. Let $d = \gcd(m, n)$. Then

$$(a^m)^{n/d} = a^{mn/d} = (a^n)^{m/d} = e^{m/d} = e.$$

However, we need to check that n/d is the smallest positive integer which gives the identity. Suppose that $k \in \mathbb{Z}^+$ and $(a^m)^k = e$. Then

$$a^{mk} = e,$$

so $n \mid mk$ by the previous proposition. It follows that n/d divides $(m/d)k$. Now we claim that $\gcd(n/d, m/d) = 1$: first use Bézout's lemma to write

$$d = nx + my$$

for some $x, y \in \mathbb{Z}$. Dividing by d, we get

$$1 = (n/d)x + (m/d)y,$$

so n/d and m/d have to be relatively prime. (Any common divisor would have to divide this linear combination, hence it would divide 1.) This means that n/d divides k.[6] In particular, $n/d \leq k$, so $|a^m| = n/d$. □

[6] This requires a general fact from number theory. If $a, b, c \in \mathbb{Z}$ with $a \mid bc$ and $\gcd(a, b) = 1$, then a must divide c.

2.4 Subgroups

We spent the last section studying groups from a very close-up perspective. In particularly, we were almost exclusively focused on facts involving individual elements of groups. We will now begin to study groups with a much wider lens, and analyze both the internal structure of a group and the relationships between different groups. We will start to consider more interesting questions, like how groups are connected to each other via functions—the concept of *homomorphism* and *isomorphism*—and how they are built out of smaller groups. Thus we are interested in questions of *classification* and *structure*:

- **Classification:** When are two groups the same or different? If we are dealing with a mysterious new example of a group, perhaps it is just a more familiar group in disguise.
- **Structure:** How is a group built out of smaller pieces? When working with a big, complicated group, we might be able to break the group up into smaller pieces, which are then easier to analyze.

We've already seen some primitive examples that foreshadow the classification problem. At this point we will start laying the foundation for the structure problem. Eventually we'll want to tear down groups into smaller bits, which will make the larger group easier to understand. To do this, we need to know what these "bits" really are.

Obviously we want to look at subsets of a group G, but we don't just want any old subset of G. We should only consider subsets that carry information about the group structure on G. These subsets we have in mind are called *subgroups* of G.

Definition 2.4.1. Let $\langle G, * \rangle$ be a group. A **subgroup** of G is a nonempty subset $H \subseteq G$ with the property that $\langle H, * \rangle$ is a group.

Note that in order for H to be a subgroup of G, H needs to be a group with respect to the operation that it inherits from G. In other words, H and G *always* carry the same binary operation. Also, we'll write

$$H \leq G$$

to denote that H is a subgroup of G. Finally, if we want to emphasize that $H \leq G$ but $H \neq G$, we will say that H is a **proper** subgroup of G.

In order to think about how to show that a subset of a group is actually a subgroup, let's work with an example.

Example 2.4.2. Let's look at the group $\mathbb{Z}$ (under addition, of course). Define

$$2\mathbb{Z} = \{\text{even integers}\} = \{2n : n \in \mathbb{Z}\}.$$

Is $2\mathbb{Z}$ a subgroup of $\mathbb{Z}$? We need to check that $2\mathbb{Z}$ itself forms a group under addition:

- **Closure:** If $a, b \in 2\mathbb{Z}$, then $a = 2n$ and $b = 2m$ for some $n, m \in \mathbb{Z}$. Then

$$a + b = 2n + 2m = 2(n + m) \in 2\mathbb{Z},$$

  so $2\mathbb{Z}$ is indeed closed under addition.

- **Associativity:** Is there even anything to check here? No—the operation on $\mathbb{Z}$ is already associative, so nothing changes when we pass to a subset of $\mathbb{Z}$.

- **Identity:** The identity for addition on $\mathbb{Z}$ is 0, which is even: $0 = 2 \cdot 0 \in 2\mathbb{Z}$.

- **Inverses:** If $a \in 2\mathbb{Z}$, then $a = 2n$ for some $n \in \mathbb{Z}$, and $-a = -2n = 2(-n) \in 2\mathbb{Z}$.

Therefore, $\langle 2\mathbb{Z}, + \rangle$ is a group, hence it is a subgroup of $\mathbb{Z}$.

In general, we want to be able to check whether a subset of a group is actually a subgroup. Fortunately, this example tells us exactly how to do it.

> **To check that $H \leq G$,** one needs to verify the following:
>
> 1. H is **closed** under the operation on G.
> 2. The **identity** element $e \in G$ is in H.
> 3. For every $a \in H$, its **inverse** a^{-1} is in H.

Example 2.4.3. Naturally, it would be helpful to look at some examples of subgroups.

1. Every group G has two special subgroups, namely

$$\{e\} \text{ and } G.$$

   These are called the **trivial subgroups** of G.[7]

2. We saw earlier that $2\mathbb{Z}$ is a subgroup of $\mathbb{Z}$. There is nothing special about 2 in this example: for any $n \in \mathbb{Z}^+$,

$$n\mathbb{Z} = \{na : a \in \mathbb{Z}\}$$

   is a subgroup of $\mathbb{Z}$. The exact same computations that we performed for $2\mathbb{Z}$ will show that $n\mathbb{Z} \leq \mathbb{Z}$.

3. The rational numbers $\mathbb{Q}$ form an additive subgroup of $\mathbb{R}$.

[7] Many books will reserve the phrase "trivial subroup" only for the identity subgroup $\{e\}$. The group G is sometimes referred to as the **improper subgroup**.

4. Here is an example from linear algebra. Consider the n-dimensional vector space $\mathbb{R}^n$. Then $\mathbb{R}^n$ is, in particular, an abelian group under addition, and any vector subspace of $\mathbb{R}^n$ is a subgroup of $\mathbb{R}^n$.[8] If H is a subspace of $\mathbb{R}^n$, then it is closed under addition, and closure under scalar multiplication guarantees that $0 \in H$ and for $v \in H$, $-v = -1 \cdot v \in H$.

5. Let D_n be the nth dihedral group, and let

$$H = \{i, r_1, r_2, \ldots, r_{n-1}\}$$

be the set of rotations in D_n. Then $H \leq D_n$. It is closed, since the composition of two rotations is another rotation, the identity $i \in H$, and for any $1 \leq j \leq n-1$,

$$r_j^{-1} = r_{n-j},$$

which is again in H.

6. Let $\mathrm{GL_n}(\mathbb{R})$ be our usual group of invertible $n \times n$ matrices under matrix multiplication, and define

$$\mathrm{SL_n}(\mathbb{R}) = \{A \in \mathrm{GL_n}(\mathbb{R}) : \det(A) = 1\}.$$

Then $\mathrm{SL_n}(\mathbb{R}) \leq \mathrm{GL_n}(\mathbb{R})$. To see that it is closed, we recall that $\det(AB) = \det(A)\det(B)$, so if $A, B \in \mathrm{SL_n}(\mathbb{R})$,

$$\det(AB) = \det(A)\det(B) = 1 \cdot 1 = 1,$$

and $AB \in \mathrm{SL_n}(\mathbb{R})$. The identity matrix surely has determinant 1, and if $A \in \mathrm{SL_n}(\mathbb{R})$, then

$$\det(A^{-1}) = \frac{1}{\det(A)} = \frac{1}{1} = 1,$$

so $A^{-1} \in \mathrm{SL_n}(\mathbb{R})$. Therefore, $\mathrm{SL_n}(\mathbb{R})$ is a subgroup of $\mathrm{GL_n}(\mathbb{R})$, called the **special linear group**.[9]

[8] A word of caution about this statement: any subspace of $\mathbb{R}^n$ is necessarily a subgroup, but there are plenty of subgroups that are not vector subspaces. For example, the *rational* vector space $\mathbb{Q}^n$ is an additive subgroup of $\mathbb{R}^n$, but it is not a *real* subspace. It is a $\mathbb{Q}$-subspace, however, which goes to show that the field of scalars is critical when talking about subspaces of vector spaces. Even worse, the set

$$\mathbb{Z}^n = \{(a_1, a_2, \ldots, a_n) : a_i \in \mathbb{Z}\}$$

is a subgroup of $\mathbb{R}^n$, called a **lattice**, but it is not a vector space of any kind. (If you're feeling extra brave/curious, it is an example of a more general object, called a **module**.)

[9] You might wonder what makes the special linear group "special." If we view the matrices in $\mathrm{GL_n}(\mathbb{R})$ as invertible linear transformations from $\mathbb{R}^n$ to itself, then the elements of $\mathrm{SL_n}(\mathbb{R})$ are precisely the transformations that preserve volume and orientation.

7. Here's our last example, and a more interesting one at that. Exercise 2.3 leads you to the observation that there are two distinct groups of order 4, and both are abelian. One of them is of course $\mathbb{Z}_4$, and the other is a new example—it is called the **Klein 4-group** and denoted by V_4.[10] If we write $V_4 = \{e, a, b, c\}$, then the elements satisfy the relations

$$a^2 = b^2 = c^2 = e$$

and

$$ab = c, bc = a, ca = b.$$

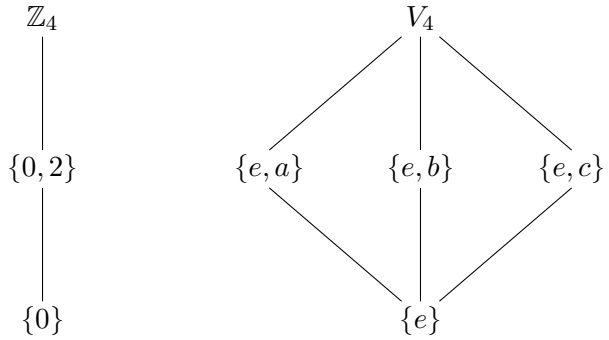
Note that V_4 is abelian by Exercise 2.3[11], and that exercise also shows that $\mathbb{Z}_4$ and V_4 really are different groups, since they have completely different Cayley tables. We will now observe this in a different way, by checking that they have different subgroup structures. First, we claim that the only nontrivial subgroup of $\mathbb{Z}_4$ is $H = \{0, 2\}$. It's easy to check that this is a subgroup, but why is it the only one? We will prove a result to this effect quite soon, or we could observe that the other possible proper subgroups are

$$\{0, 1\}, \{0, 3\}, \{0, 1, 2\}, \{0, 1, 3\}, \{0, 2, 3\},$$

and it is easy to check that none of these are closed under addition. On the other hand, V_4 has three subgroups (in addition to $\{e\}$ and V_4 itself):

$$\{e, a\}, \{e, b\}, \{e, c\}.$$

We therefore know all the subgroups of these two groups, and we can represent this pictorially with something called a **subgroup lattice** diagram:



[10] The notation V_4 comes from Felix Klein's mother tongue, German. The German word for *four* is *vier*, so his group is known as *Viergruppe*.

[11] You could also check that if G is a group with the property that $x^2 = e$ for all $x \in G$, then G must be abelian.

As we have already insinuated, we will sometimes be able to tell when two groups are different by studying these sorts of lattice diagrams. For example, $\mathbb{Z}_4$ has only one subgroup of order 2, while V_4 has three such subgroups. This is one indication that these are indeed different groups. (If you go on to study the field known as *Galois theory* at all, you will see that subgroup diagrams are quite important in that context.)

2.4.1 Subgroup Criteria

Aside from directly verifying the appropriate axioms, there are a couple of other criteria one can use to show that a subset of a group is actually a subgroup. When dealing with specific examples, it is often easiest to simply verify the axioms. However, when proving things about subgroups it can be useful to use one of the following characterizations. The first one shows that we can collapse the usual subgroup axioms into a single condition.

Theorem 2.4.4. *Let G be a group. A nonempty subset $H \subseteq G$ is a subgroup if and only if whenever $a, b \in H$, $ab^{-1} \in H$.*

Proof. Suppose that $H \leq G$, and let $a, b \in H$. Then $b^{-1} \in H$, so $ab^{-1} \in H$ since H is closed.

Conversely, suppose that $ab^{-1} \in H$ for all $a, b \in H$. Then for any $a \in H$, we can take $a = b$ and conclude that

$$e = aa^{-1} \in H,$$

so H contains the identity. Since $e \in H$, for any $a \in H$ we have

$$a^{-1} = ea^{-1} \in H,$$

so H is closed under taking inverses. Finally, we claim that H is closed under the group operation. If $a, b \in H$, then $b^{-1} \in H$, so $b^{-1}a^{-1} \in H$, and therefore

$$ab = \left((ab)^{-1}\right)^{-1} = (b^{-1}a^{-1})^{-1} \in H.$$

Thus H is closed, hence a subgroup of G. □

The next criterion is quite interesting. It obviously reduces the number of things that one needs to check, but it only works for a *finite* subset of a group G.

Theorem 2.4.5. *Let G be a group and H a nonempty **finite** subset of G. Then H is a subgroup if and only if H is closed under the operation on G.*

Proof. If H is a subgroup, then it is obviously closed by hypothesis.

On the other hand, we are assuming that H is closed, so we need to verify that $e \in H$ and that for every $a \in H$, $a^{-1} \in H$ as well. Since $\{e\} \leq G$, we may assume that H is nontrivial, i.e. that H contains an element a distinct from the identity. Since H is closed, the elements

$$a, a^2, a^3, \ldots$$

are all in H, and since H is finite, this list cannot go on forever. That is, we must eventually have duplicates on this list, so

$$a^i = a^j$$

for some $1 \leq i < j \leq |H|$. Since $i < j$, $j - i \geq 0$ and we have

$$a^i = a^j = a^{j-i}a^i,$$

and using cancellation, we get

$$a^{j-i} = e.$$

Therefore, $e \in H$. Now observe that $j - i - 1 \geq 0$, so $a^{j-i-1} \in H$, and

$$aa^{j-i-1} = a^{j-i} = e,$$

so $a^{-1} = a^{j-i-1} \in H$. Therefore, H is a subgroup of G. □

This theorem has an easy corollary, which is useful when the group is finite.

Corollary 2.4.6. *If G is a **finite** group, a nonempty subset $H \subseteq G$ is a subgroup of G if and only if it is closed under the operation on G.*

2.4.2 Cyclic Subgroups

Some of the examples that we have mentioned are actually cases of very special kinds of subgroups, called **cyclic subgroups**. Suppose G is a group, and let $a \in G$. Earlier, we defined shorthand notation for the set of all powers of a:

$$\langle a \rangle = \left\{ a^j : j \in \mathbb{Z} \right\}.$$

In fact, the results of Proposition 2.3.9 show that $\langle a \rangle$ is a subgroup of G:

- **Closure:** $a^i a^j = a^{i+j} \in \langle a \rangle$ for all $i, j \in \mathbb{Z}$.
- **Identity:** $e = a^0 \in \langle a \rangle$.
- **Inverses:** Since $a^j a^{-j} = a^{j-j} = a^0 = e$, we have $(a^j)^{-1} = a^{-j} \in \langle a \rangle$ for all $j \in \mathbb{Z}$.

In other words, every element of a group G "generates" a whole subgroup of G. We attach a special name to such subgroups.

Definition 2.4.7. Let G be a group and $a \in G$. The subgroup $\langle a \rangle$ is called the **cyclic subgroup** generated by a.

When we say that a "generates" $\langle a \rangle$, we mean that that $\langle a \rangle$ is created entirely out of the element a. In a certain sense, $\langle a \rangle$ is the *smallest* possible subgroup of G which contains a. Let's try to make this more precise. If $H \leq G$ and $a \in H$, then H must contain the elements

$$a, a^2, a^3, \ldots,$$

since H is closed. It also must contain e and a^{-1}, hence all of the elements

$$\ldots, a^{-2}, a^{-1}, e, a, a^2, \ldots,$$

i.e. all powers of a. That is, $\langle a \rangle \subseteq H$, and we have proven the following fact:

Theorem 2.4.8. *Let G be a group and let $a \in G$. Then $\langle a \rangle$ is the smallest subgroup of G containing a, in the sense that if $H \leq G$ and $a \in H$, then $\langle a \rangle \subseteq H$.*

Of course we've already encountered several examples of cyclic subgroups in our studies thus far.

Example 2.4.9.

1. Our first example of a subgroup, $2\mathbb{Z} \leq \mathbb{Z}$, is a cyclic subgroup, namely $\langle 2 \rangle$. Similarly, $n\mathbb{Z}$ is cyclic for any $n \in \mathbb{Z}$.
2. The subgroup consisting of all the rotations in D_n,

   $$H = \{i, r_1, r_2, \ldots, r_{n-1}\} \leq D_n,$$

   is cyclic since $H = \langle r_1 \rangle$.
3. All the proper subgroups of $\mathbb{Z}_4$ and V_4 that we listed are cyclic. In addition, $\mathbb{Z}_4$ is a cyclic subgroup of itself, but V_4 is not.
4. The trivial subgroup $\{e\}$ is always a cyclic subgroup, namely $\langle e \rangle$.

Cyclic subgroups are useful in part because they allow us to gather information about an element of G by studying the subgroup that it generates. Properties of the subgroup are reflected in those of the element, and vice versa. One example is the relationship between the order of an element and the order of the associated cyclic subgroup:

Theorem 2.4.10. *Let G be a group. For each $a \in G$, we have $|\langle a \rangle| = |a|$.*

Proof. Suppose first that a has infinite order. Then by Proposition 2.3.18, the elements of $\langle a \rangle$ are all distinct. It follows that $|\langle a \rangle| = \infty$.

Now suppose a has finite order, say $n = |a|$. Recall that $a^i = a^j$ if and only if $i \equiv j \bmod n$, so the elements of $\langle a \rangle$ are

$$e, a, a^2, a^3, \ldots, a^{n-1},$$

of which there are n. □

2.5 Cyclic Groups

We saw in the last section that every element of a group G generates a cyclic subgroup of G. Is it possible that G could be a cyclic subgroup of itself? That is, can a single element a "generate" the whole group G? Of course the answer is yes—we already remarked that $\mathbb{Z}_4$ is a cyclic subgroup of itself, and the same is true of any cyclic subgroup (viewed as a group in its own right). There are many examples of this special family of groups.

Definition 2.5.1. A group G is called **cyclic** if $G = \langle a \rangle$ for some $a \in G$. The element a is called a **generator** for G.

Cyclic groups are extremely well-understood. We'll see that they have very nice properties, and we can completely classify them almost immediately. First, let's do some examples, many of which we have already seen.

Example 2.5.2.

1. One of our first examples of a group is actually a cyclic one: $\mathbb{Z}$ forms a cyclic group under addition. What is a generator for $\mathbb{Z}$? Both 1 and -1 generate it, since every integer $n \in \mathbb{Z}$ can be written as a "power" of 1 (or -1):
$$n = n \cdot 1 = \underbrace{1 + 1 + \cdots + 1}_{n \text{ times}}.$$
Thus,
$$\mathbb{Z} = \langle 1 \rangle = \langle -1 \rangle.$$
These are actually the only two generators.

2. How about a finite cyclic group? For any n, $\mathbb{Z}_n$ is cyclic, and 1 is a generator in much the same way that 1 generates $\mathbb{Z}$. There are actually plenty of other generators, and we can characterize them by using our knowledge of greatest common divisors. We'll postpone this until we've made a couple of statements regarding cyclic groups.

3. The Klein 4-group is not cyclic, since every non-identity element has order 2. If it were cyclic, it would contain an element of order 4.

4. The dihedral group D_3 is not cyclic. The rotations all have order 3, so

$$\langle r_1 \rangle = \langle r_2 \rangle = \{i, r_1, r_2\}.$$

   On the other hand, all of the reflections have order 2, so

$$\langle m_1 \rangle = \{i, m_1\}, \quad \langle m_2 \rangle = \{i, m_2\}, \quad \langle m_3 \rangle = \{i, m_3\}.$$

5. The group $\langle \mathbb{Q}, + \rangle$ is not cyclic. Suppose to the contrary that it is cyclic, and let $\frac{p}{q}$ be a generator. Let $\frac{r}{s} \in \mathbb{Q}$, and assume it is written in lowest terms (i.e., $\gcd(r, s) = 1$) and that $\gcd(q, s) = 1$. Since $\frac{p}{q}$ generates $\mathbb{Q}$, there exists $n \in \mathbb{Z}$ such that

$$n \cdot \frac{p}{q} = \frac{r}{s},$$

   or $nps = qr$. In particular, s divides qr. However, this is impossible, since we have assumed s is relatively prime to both q and r.

Now let's start making some fundamental observations regarding cyclic groups. First, if $G = \langle a \rangle$ is cyclic, how big is it? We already saw the answer in Theorem 2.4.10, which showed that our overloading of the word "order" was appropriate after all.

Theorem 2.5.3. *If $G = \langle a \rangle$ is cyclic, then $|G| = |a|$.*

If we pair this result with Theorem 2.3.19, we can characterize the generators of any finite cyclic group.

Proposition 2.5.4. *The generators of a finite cyclic group $G = \langle a \rangle$ of order n are precisely the elements a^r for which $\gcd(r, n) = 1$.*

Proof. By Theorem 2.3.19,

$$|a^r| = \frac{n}{\gcd(r, n)}.$$

Note that a^r generates G if and only if $|a^r| = n$. If a^r has order n, then $|\langle a^r \rangle| = n$ and $\langle a^r \rangle \subseteq \langle a \rangle$. Since both sets have the same (finite) number of elements, they must be the same. On the other hand, if $\langle a^r \rangle = \langle a \rangle$, then

$$|a^r| = |\langle a^r \rangle| = |\langle a \rangle| = n.$$

Now note that $|a^r| = n$ precisely when $\gcd(r, n) = 1$. □

Remark 2.5.5. As a special case, this proposition allows us to characterize the generators of $\mathbb{Z}_n$. Recall that $\mathbb{Z}_n = \langle 1 \rangle$. Remember that we have to be careful here—when we write something like a^m in an arbitrary group, we mean "multiply a by itself m times," where "multiply" should be interpreted as whatever the group operation happens to be. In particular, the operation in $\mathbb{Z}_n$ is addition, so a^m really means

$$\underbrace{a + a + \cdots + a}_{m \text{ times}} = m \cdot a.$$

Therefore, the generators of $\mathbb{Z}_n$ are precisely the elements of the form $r \cdot 1$, where $\gcd(r, n) = 1$. In other words, an element $a \in \mathbb{Z}_n$ is a generator for the cyclic group $\mathbb{Z}_n$ if and only if $\gcd(a, n) = 1$.

On a related note, let $G = \langle a \rangle$ be a finite cyclic group with $n = |G| = |a|$. If we take any element $b \in G$ and compute b^n, what do we get? Well, we know that $b = a^i$ for some $i \in \mathbb{Z}$, so

$$b^n = (a^i)^n = a^{in} = (a^n)^i = (a^{|a|})^i = e^i = e.$$

Therefore:

Theorem 2.5.6. *Let G be a finite cyclic group. Then for any element $b \in G$, we have $b^{|G|} = e$.*

If we combine this result with Proposition 2.3.16, then what can we say about $|b|$ in relation to $|G|$? We must have that $|b|$ divides $|G|$.

Proposition 2.5.7. *Let G be a finite cyclic group. For any $b \in G$, $|b|$ divides $|G|$.*

These last two results are not as serendipitous as they may seem at first glance. These phenomena for cyclic groups actually hold for *any* finite group. Once we have established Lagrange's theorem in the next section, we'll see why this is true.

Finally, the astute reader will notice a key similarity between all of the examples of cyclic groups that we've listed above: they are all *abelian*. It is not hard to see that this is always the case.

Theorem 2.5.8. *Every cyclic group is abelian.*

Proof. Let G be a cyclic group and let a be a generator for G, i.e. $G = \langle a \rangle$. Then given two elements $x, y \in G$, we must have $x = a^i$ and $y = a^j$ for some $i, j \in \mathbb{Z}$. Then

$$xy = a^i a^j = a^{i+j} = a^{j+i} = a^j a^i = yx,$$

and it follows that G is abelian. □

The converse to Theorem 2.5.8 is far from true. In particular, the Klein 4-group is an abelian group of order 4, but we have already observed that it is not cyclic. In fact, V_4 is the smallest example of a non-cyclic group (abelian or not).

2.5.1 Subgroups of Cylic Groups

It is quite easy to analyze the structure of cyclic groups in great detail. As a start, we can completely describe the subgroups of any cyclic group.

Theorem 2.5.9. *Any subgroup of a cyclic group is cyclic.*

Proof. Let $G = \langle a \rangle$ be a cyclic group and let H be a subgroup of G. We may assume that $H \neq \{e\}$, since $\{e\}$ is already known to be cyclic. Then H contains an element other than e, which must have the form a^m for some $m \in \mathbb{Z}$, since G is cyclic. Assume that m is the *smallest* positive integer for which $a^m \in H$. We claim that $H = \langle a^m \rangle$. To do this, we need to show that if $a^n \in H$, then a^n is a power of a^m.

Suppose that $a^n \in H$, and use the Division Algorithm to write $n = qm + r$, where $0 \leq r < m$. Then

$$a^n = a^{qm+r} = a^{qm}a^r = (a^m)^q a^r.$$

Since H is a subgroup, $(a^m)^{-q} \in H$, hence $(a^m)^{-q}a^n \in H$, and it follows that

$$a^r = (a^m)^{-q}a^n$$

is in H. But $r < m$ and we have assumed that m is the smallest positive integer such that $a^m \in H$, so we must have $r = 0$. In other words, $a^n = (a^m)^q$, so $a^n \in \langle a^m \rangle$. Since a^n was an arbitrary element of H, we have shown that $H \subseteq \langle a^m \rangle$. Since $a^m \in H$, we also have $\langle a^m \rangle \subseteq H$, so $H = \langle a^m \rangle$, and H is cyclic. □

This theorem has a particularly nice corollary, which tells us a lot about the structure of $\mathbb{Z}$ as an additive group.

Corollary 2.5.10. *The only subgroups of $\mathbb{Z}$ are the cyclic subgroups $n\mathbb{Z}$, where $n \in \mathbb{Z}$.*

Proof. The cyclic subgroups of $\mathbb{Z}$ are simply $n\mathbb{Z} = \langle n \rangle$ for any $n \in \mathbb{Z}$. By the theorem, the only subgroups of $\mathbb{Z}$ are the cyclic ones, so we are done. □

The next corollary is quite interesting. Recall that if $G = \langle a \rangle$ is a finite cyclic group of order n and $a^j \in G$, we have a formula for the order of a^j:

$$|a^j| = \frac{n}{\gcd(j, n)}.$$

This tells us something in particular, which we actually proved earlier: if $a^j \in \langle a \rangle$, then $|a^j|$ divides $|\langle a \rangle|$. In particular, the order of the cyclic subgroup generated by a^j must divide the order of $\langle a \rangle$. This holds more generally—in fact, it is true for any finite group, as we will see when we prove Lagrange's theorem. However, a partial converse holds for cyclic groups (but not all groups): if m divides the order of a group G, then G has a subgroup of order m. Moreover, if a cyclic group has a subgroup of a given order, then that subgroup is unique.[12]

Corollary 2.5.11. *Let $G = \langle a \rangle$ be a cyclic group of order n. If m is a positive divisor of m, then G has exactly one subgroup of order m.*

Proof. First we need to show that G even has a subgroup of order m whenever $m \mid n$. Well, suppose that $m \mid n$, and put $b = a^{n/m}$. Then by Theorem 2.3.19,

$$|b| = \frac{n}{\gcd(n/m, n)} = \frac{n}{n/m} = m.$$

Therefore, $\langle b \rangle$ has m elements, so G has a subgroup of order m.

Now we need to show that G only possesses one subgroup of order m. Suppose that a^j is another element of order m, so that $\langle a^j \rangle$ has m elements. Then

$$|a^j| = \frac{n}{\gcd(j, n)} = m = \frac{n}{\gcd(n/m, n)} = |b|.$$

Therefore, $\gcd(j, n) = \gcd(n/m, n) = n/m$. In particular, n/m divides j, so we can write

$$j = r(n/m)$$

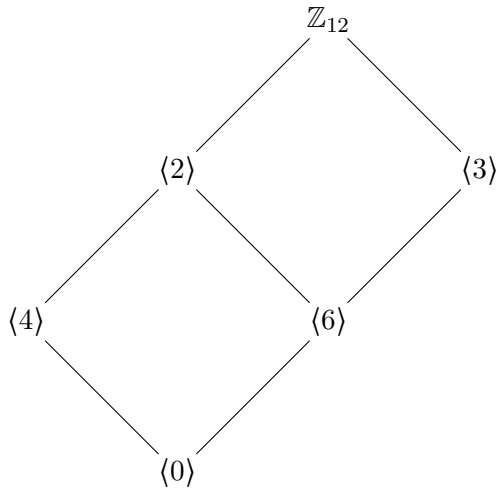
for some $r \in \mathbb{Z}$. Then

$$a^j = a^{rn/m} = (a^{n/m})^r = b^r,$$

so $a^j \in \langle b \rangle$. This forces $\langle a^j \rangle \subseteq \langle b \rangle$, and since both sets have the same (finite) number of elements, $\langle a^j \rangle = \langle b \rangle$. □

Example 2.5.12. Let's use this corollary to write down all of the subgroups of $\mathbb{Z}_{12}$. We know that there will be one of each order that divides 12, and these divisors are 1, 2, 3, 4, 6, and 12. The order 1 subgroup is just $\{0\}$, and for order 12 we have the whole group $\mathbb{Z}_{12}$. For the others, we need to find an element of that order. We can see that 6 has order 2, so $\{0, 6\}$ is our order 2 subgroup. For order 3, we can use 4 as the generator, so $\{0, 4, 8\}$ is the subgroup. For order 4, we have $3 \in \mathbb{Z}_{12}$, and the subgroup is $\{0, 3, 6, 9\}$. Finally, 2 has order 6, and the subgroup is $\{0, 2, 4, 6, 8, 10\}$. Therefore, the subgroup lattice of $\mathbb{Z}_{12}$ looks like:

[12] These latter two statements fail miserably for arbitrary finite groups. If m divides the order of the group, there need not be a subgroup of order m. (It is true for abelian groups, however.) Also, if a subgroup of order m exists, it need not be unique. (Look at the Klein 4-group, for example.)



2.5.2 Classification of Cyclic Groups

As one final word on cyclic groups, we should mention that these groups are very easy to classify. They have a very rigid structure, as evidenced by some of the results we have proven regarding their elements and subgroups. Indeed, we can go even further—there are really only two flavors of cyclic group: finite (where we really have one cyclic group for each positive integer) and infinite.

Let's think about what happens in the infinite case first. If we have an infinite cyclic group $G = \langle a \rangle$, then a has infinite order, and the group elements are simply

$$G = \left\{ a^j : j \in \mathbb{Z} \right\}.$$

How does the group operation work? We have

$$a^j a^k = a^{j+k},$$

so we can multiply two elements by simply adding the exponents as integers. We haven't formally introduced the concept of *isomorphism* yet, but the map $a^j \mapsto j$ will allow us to set up an isomorphism between the infinite cyclic group G and $\mathbb{Z}$. That is, there is only one infinite cyclic group up to isomorphism, namely $\mathbb{Z}$.

Things are almost as easy in the finite case as well. Suppose that $G = \langle a \rangle$ is a finite cyclic group of order n. Then to add two elements, we have

$$a^j a^k = a^{j+k}.$$

However, we have to remember that in doing this, $j+k$ may have exceeded the order of the group, and we can reduce it. That is, we can write

$$j + k = nq + r,$$

so

$$a^j a^k = a^{j+k} = a^{nq+r} = a^{nq} a^r = e a^r = a^r.$$

Therefore, multiplication corresponds to addition of the exponents modulo n. We'll see that this allows us to identify G with the elementary cyclic group $\mathbb{Z}_n$. In other words, there is only one finite cyclic group of each order $n \in \mathbb{Z}^+$.

2.6 Lagrange's Theorem

In our investigation of cyclic groups, we noticed one thing—that if $G = \langle a \rangle$ is finite, the order of any element divides the order of the group. This implies something else: if $H \leq \langle a \rangle$, then we know that $H = \langle a^m \rangle$ for some $m \in \mathbb{Z}$, and

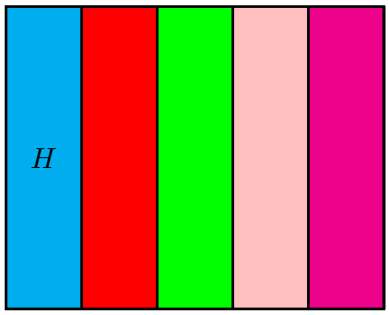
$$|H| = |\langle a^m \rangle| = |a^m| \text{ divides } |a| = |G|.$$

Put more plainly, the order of any subgroup divides the order of the group. This is no accident—it holds much more generally. In fact, it is true not just for cyclic groups, but for any finite group and any subgroup.

Theorem 2.6.1 (Lagrange)**.** *Let G be a finite group and H be a subgroup of G. Then $|H|$ divides $|G|$.*

We're not going to prove Lagrange's theorem yet. The proof isn't hard, but we don't quite have the right tools for writing it down yet. We need to develop some new language in order to properly prove it. The concepts that we'll talk about will be really important for the rest of the course, but the first thing they will buy us is a proof of Lagrange's theorem.

The idea of the proof will be the following: given a finite group G and $H \leq G$, we want to "carve up" G into a collection of subsets, all of which are determined by the subgroup H. We will first need to figure out what these subsets should be.



The proper way to say this is that we want to *partition* G. In order to do this, we need to talk about **equivalence relations**.

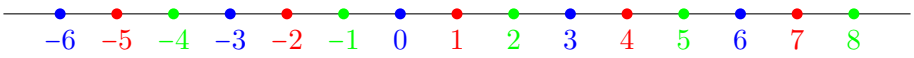
Remark 2.6.2. Lagrange's theorem is a statement about *finite* groups, but we are going to look at arbitrary groups for this part. A lot of the techniques (specifically the idea of cosets) will be important in a more general setting.

Before we formally introduce equivalence relations, let's do an example for motivation. It should be eerily familiar, or it will be once we're done with it.

Example 2.6.3. Let's look at our favorite group—$\mathbb{Z}$. We just saw that the only subgroups of $\mathbb{Z}$ are those of the form $n\mathbb{Z}$, for $n \in \mathbb{Z}^+$. Let's take $3\mathbb{Z}$, for example. What if we start "translating" this subgroup by integers? That is, we add a fixed integer to each element of $3\mathbb{Z}$:

$$3\mathbb{Z} + 0 = \{\ldots, -6, -3, 0, 3, 6, \ldots\}$$
$$3\mathbb{Z} + 1 = \{\ldots, -5, -2, 1, 4, 7, \ldots\}$$
$$3\mathbb{Z} + 2 = \{\ldots, -4, -1, 2, 5, 8, \ldots\}$$

This is depicted on the color-coded number line below:



What if we keep translating? We just get the same sets all over again:

$$3\mathbb{Z} + 3 = 3\mathbb{Z} + 0$$
$$3\mathbb{Z} + 4 = 3\mathbb{Z} + 1$$

and so on. Also, do these "translates" of $3\mathbb{Z}$ have any overlap? No—if two of the sets are distinct, then they are actually completely disjoint. Finally, they fill up all of $\mathbb{Z}$, in the sense that their union is $\mathbb{Z}$. In other words, every integer belongs to one of the sets on this list.

What do you notice about the elements of each of these sets? They are all congruent mod 3:

$$3\mathbb{Z} + 0 = \{a \in \mathbb{Z} : a \equiv 0 \bmod 3\}$$
$$3\mathbb{Z} + 1 = \{a \in \mathbb{Z} : a \equiv 1 \bmod 3\}$$
$$3\mathbb{Z} + 2 = \{a \in \mathbb{Z} : a \equiv 2 \bmod 3\}$$

In fact, we could say that these sets are determined by the relationship of "congruence mod 3"—two integers a and b fall into the same "translate" if and only if they are congruent mod 3, i.e., if and only if 3 divides $a - b$.

Before moving on, let's recap the key observations that we've made in this example. By shifting around the subgroup $3\mathbb{Z}$, we obtain translates of the subgroup such that:

1. Two integers $a, b \in \mathbb{Z}$ belong to the same translate if and only if $a \equiv b \bmod 3$.
2. Distinct translates are disjoint.
3. Their union is all of $\mathbb{Z}$.

These three conditions are the ones we would like to be able to encode in more general groups. Note that the latter two conditions say that we have **partitioned** $\mathbb{Z}$, which you may know corresponds to defining an *equivalence relation* on $\mathbb{Z}$.

2.6.1 Equivalence Relations

In mathematics, when one wants to partition a set, the natural way to do it is with something called an *equivalence relation*. These objects are important throughout mathematics, and not just in the field of algebra. This means that we are about to step away from algebra for a moment and talk about some even more abstract ideas (if you can believe such a thing). The primary examples will be from algebra, of course, but these sorts of concepts are ubiquitous throughout mathematics. They are used in algebra, topology, analysis, combinatorics, and other fields. Therefore, this will be a good place for you to encounter them.

Even though equivalence relations are fairly abstract, we'll use congruence mod n as a guiding example. In that case, we have some very nice things going on. There are three things in particular that we can single out:

1. If $a \in \mathbb{Z}$, then $a \equiv a \bmod n$.
2. If $a, b \in \mathbb{Z}$ and $a \equiv b \bmod n$, then $b \equiv a \bmod n$.
3. If $a, b, c \in \mathbb{Z}$ with $a \equiv b \bmod n$ and $b \equiv c \bmod n$, then $a \equiv c \bmod n$.

These are the properties that we will eventually want equivalence relations to possess. Before we can properly define them, we need to make the word "relation" precise.

Definition 2.6.4. Let S be a set. A **relation** on S is a set $R \subseteq S \times S$.

Remark 2.6.5. The definition above is the formal, precise way of describing a relation. It is not a particularly helpful way in which to think about relations. You should simply think of a relation as a way to "pair off" elements of S. To this end, we will usually write $x \sim y$ to mean $(x, y) \in R$, and we will say that "x is related to y." We will usually refer to $\sim$ as the relation, and not worry too much about the proper definition.

We will not concern ourselves with general relations. Our main interests will lie in relations that behave like "congruence mod n."

Definition 2.6.6. Let S be a set. An **equivalence relation** on S is a relation $\sim$ on S satisfying the following three properties:

1. **Reflexivity:** $a \sim a$ for all $a \in S$.
2. **Symmetry:** If $a, b \in S$ and $a \sim b$, then $b \sim a$.
3. **Transitivity:** If $a, b, c \in S$ with $a \sim b$ and $b \sim c$, then $a \sim c$.

We will usually think of an equivalence relation as a way of pairing off elements of S by equivalence. If $a \sim b$, we will usually say that "a is equivalent to b."

The whole point of introducing equivalence relations is to obtain a method of partitioning a group. Therefore, we will of course be interested in subsets of S that are somehow "determined" by an equivalence relation.

Definition 2.6.7. Given $a \in S$, define

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

We call $[a]$ the **equivalence class** of a, and a is called a **representative**[13] of the equivalence class.

Example 2.6.8. Here are some examples of equivalence relations.

1. This is one of the most important examples, and it's the one we've already seen. Fix $n \in \mathbb{Z}$, and define $\sim$ on $\mathbb{Z}$ by $a \sim b$ if and only if $a \equiv b \bmod n$. We have already checked that this is an equivalence relation. What are the equivalence classes? For $a \in \mathbb{Z}$, we have

$$[a] = \{b \in \mathbb{Z} : b \equiv a \bmod n\} = n\mathbb{Z} + a.$$

2. Define $\sim$ on $\mathbb{R}$ by $a \sim b$ if $a = b$. Then $\sim$ is an equivalence relation:

   - **Reflexive:** For any $a \in \mathbb{R}$, $a = a$, so $a \sim a$.
   - **Symmetric:** If $a, b \in \mathbb{R}$ and $a = b$, then $b = a$, so $a \sim b$ implies that $b \sim a$.
   - **Transitive:** If $a, b, c \in \mathbb{R}$ with $a = b$ and $b = c$, then $a = c$.

   What are the equivalence classes? Each class consists of a single element:

$$[a] = \{a\}.$$

[13] Note that an equivalence class may have many representatives. In fact, a representative is just a chosen element of the equivalence class, so any element of the class can be taken as a representative.

3. Define $\sim$ on $\mathrm{GL_n}(\mathbb{R})$ by $A \sim B$ if $\det(A) = \det(B)$. It's easy to see that this is an equivalence relation, and that
$$[A] = \{B \in \mathrm{GL_n}(\mathbb{R}) : \det(B) = \det(A)\}.$$

4. This example is really the prototypical equivalence relation, in the sense that any equivalence relation can be viewed in this way. Let S be any set, and let $\{S_i\}$ be a **partition** of S. That is, $S_i \subseteq S$ for all i, $S = \bigcup S_i$, and the S_i are pairwise disjoint:
$$S_i \cap S_j = \varnothing$$
if $i \neq j$. We call the S_i the **cells** of the partition. We can define an equivalence relation $\sim$ on S by $a \sim b$ if and only if a and b belong to the same cell S_i. It is easy to check that $\sim$ is reflexive, symmetric, and transitive, hence an equivalence relation.

Exercise 2.7. Verify that the relation $\sim$ defined in Example 2.6.8(4) is an equivalence relation.

Aside 2.6.9. While we're on the topic of specific examples, let's make a couple of comments regarding $\mathbb{Z}$ with the relation of congruence mod n. Note that the equivalence classes are simply congruence classes mod n: every element of the class $[a]$ is congruent to a mod n. Thus all elements of $[a] = n\mathbb{Z} + a$ yield the same element of $\mathbb{Z}_n$ when reduced mod n. Therefore, the equivalence classes $n\mathbb{Z} + a$ can be identified with elements of $\mathbb{Z}_n$. We will see soon that this is really the correct way to view $\mathbb{Z}_n$. The elements of $\mathbb{Z}_n$ are actually the *equivalence classes* mod n, and we define modular addition and multiplication by
$$[a] + [b] = [a + b]$$
and
$$[a][b] = [ab].$$
Of course, we would need to know that these operations are *well-defined*, meaning they don't depend on the choice of representatives for the equivalence classes. We'll take care of that when we revisit $\mathbb{Z}_n$ later on. We will discuss these ideas in more detail when we encounter quotient groups, of which $\mathbb{Z}_n$ will be the foremost example.

Now let's return to general equivalence relations. We said that we were introducing them to allow us to partition sets, so we should check that they actually do this. We've already seen that a partition of a set S actually imposes an equivalence relation on S, so we are really checking that equivalence relations and partitions go hand in hand.

Theorem 2.6.10. *Let $\sim$ be an equivalence relation on a set S. The equivalence classes of $\sim$ **partition** S, in the sense that:*

1. *Given $a, b \in S$, either $[a] \cap [b] = \varnothing$, or $[a] = [b]$.*
2. *S is the union of all the equivalence classes of $\sim$. That is, every element of S belongs to some equivalence class (and only one, by condition 1).*

Proof. Let $a, b \in S$. Then either $[a] \cap [b] = \varnothing$, or $[a] \cap [b]$ is nonempty. In the second case, there is at least one element $c \in [a] \cap [b]$. Then $c \in [a]$, so $a \sim c$, and $c \in [b]$, so $c \sim b$. By transitivity, $a \sim b$. Thus $a \in [b]$, and if $x \in [a]$, then $x \sim b$ by transitivity, so $x \in [b]$ as well. Thus $[a] \subseteq [b]$. By symmetry, we also have $b \in [a]$, and a similar transitivity argument shows that $[b] \subseteq [a]$. Thus $[a] = [b]$.

To see that S is the union of the equivalence classes, we just need to notice that every $a \in S$ belongs to one of the equivalence classes. Specifically, $a \in [a]$ since $\sim$ is reflexive. Thus S is contained in the union of the equivalence classes, so in fact S equals the union of the equivalence classes. Therefore, the equivalence classes of $\sim$ partition S. □

2.6.2 Cosets

Let's step back to reality now. (Or, as close to reality as abstract algebra can be.) We are especially interested in the case where the set in question is actually a group, and the equivalence relation has something to do with a given subgroup. That is, we want to partition a group G into subsets, each of which is determined by some fixed subgroup H. Once we have done this, we will able to write down a proof of Lagrange's theorem in a nice way. Our present goal then is to find an equivalence relation on a group G which is somehow related to a subgroup H. There are two very similar ones that we can define, and either one will work.

Let's return for a moment to our key example. The relation $\sim$ that we defined on $\mathbb{Z}$ by

$$a \sim b \iff a \equiv b \bmod n$$

is really equivalent to specifying that $a \sim b$ if and only if $n \mid (a - b)$. This in turn is equivalent to saying that $a - b \in n\mathbb{Z}$. Let's try to generalize this idea to subgroups in general.

Example 2.6.11. Let G be a group and $H \leq G$. Define a relation $\sim_H$ on G by: $a \sim_H b$ if and only if $ab^{-1} \in H$. Is this an equivalence relation?

- **Reflexive:** If $a \in G$, then $aa^{-1} = e \in H$, so $a \sim_H a$.
- **Symmetric:** Suppose that $a, b \in G$ and $a \sim_H b$, so $ab^{-1} \in H$. To show that $b \sim_H a$, we need to know that $ba^{-1} \in H$. But

$$ba^{-1} = (ab^{-1})^{-1} \in H,$$

since H is a subgroup of G. Thus $b \sim_H a$, and the relation is symmetric.

- **Transitive:** Suppose that $a \sim_H b$ and $b \sim_H c$. Is $a \sim_H c$? We need to know that $ac^{-1} \in H$. Well,

$$ac^{-1} = a(b^{-1}b)c^{-1} = (ab^{-1})(bc^{-1}),$$

and $ab^{-1} \in H$ (since $a \sim_H b$) and $bc^{-1} \in H$ (since $b \sim_H c$), so $ac^{-1} \in H$. It follows that $a \sim_H c$.

Here's the next logical question: what are the equivalence classes of $\sim_H$? Well,

$$[a]_H = \{b \in G : b \sim_H a\} = \left\{b \in G : ba^{-1} \in H\right\}.$$

If $a \sim_H b$, then $ab^{-1} \in H$, i.e., there exists $h \in H$ such that

$$ba^{-1} = h,$$

or $b = ha$. If we define

$$Ha = \{ha : h \in H\},$$

then we have just shown that $[a]_H \subseteq Ha$. On the other hand, given $b = ha \in Ha$,

$$ba^{-1} = (ha)a^{-1} = h(aa^{-1}) = h \in H,$$

so $b \sim_H a$. Thus $Ha \subseteq [a]_H$. Therefore, we have shown that the equivalence classes are exactly

$$[a]_H = Ha,$$

and we'll generally use Ha instead of the $[\cdot]_H$ notation when working in the group case. These equivalence classes actually have a special name, which we will use exclusively from now on.

Definition 2.6.12. Let G be a group and $H \leq G$. Sets of the form Ha for $a \in G$ are called **(right) cosets** of H in G.

We have already established several facts regarding cosets. Let's summarize what we proved above.

Theorem 2.6.13. *Let G be a group, H a subgroup of G, and let $\sim_H$ be the relation on G given by*

$$a \sim_H b \iff ab^{-1} \in H.$$

Then $\sim_H$ is an equivalence relation, and the equivalence classes are precisely the right cosets of H:

$$[a]_H = Ha.$$

Furthermore, if $a, b \in G$, then either $Ha \cap Hb = \varnothing$ or $Ha = Hb$, and

$$Ha = Hb \iff ab^{-1} \in H.$$

Finally, G is the union of all the right cosets of H, so the cosets partition G.

Two of the examples of equivalence relations that we mentioned last time are actually the relation $\sim_H$ in disguise. In fact, this relation is meant to be a generalization of congruence mod n.

Example 2.6.14. 1. Consider the group $\mathbb{Z}$ and the subgroup $H = n\mathbb{Z}$. Given $a, b \in \mathbb{Z}$, we have seen that $a \sim_H b$ if and only if $a - b \in H$, which is just the additive way of writing $ab^{-1} \in H$. Thus $\sim_H$ is really just congruence mod n, and the right cosets of H are

$$[a]_H = n\mathbb{Z} + a = \{nc + a : c \in \mathbb{Z}\},$$

and there are n of them, namely $n\mathbb{Z} + 0, n\mathbb{Z} + 1, \ldots, n\mathbb{Z} + (n-1)$.

2. Consider $H = \mathrm{SL_n}(\mathbb{R}) \leq \mathrm{GL_n}(\mathbb{R})$. If $A, B \in \mathrm{GL_n}(\mathbb{R})$, $A \sim_H B$ precisely when $AB^{-1} \in \mathrm{SL_n}(\mathbb{R})$, or

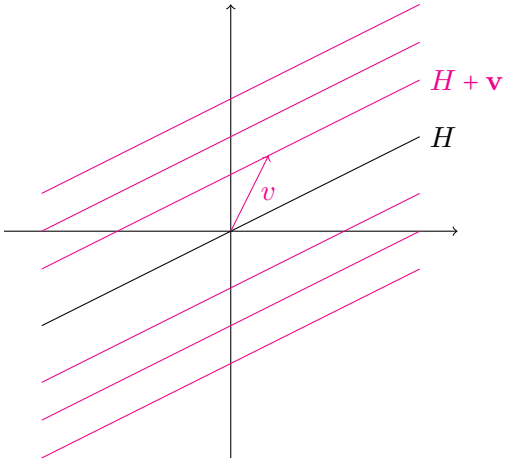
$$\det(AB^{-1}) = 1.$$

But

$$\det(AB^{-1}) = \det(A)\det(B^{-1}) = \frac{\det(A)}{\det(B)},$$

so $A \sim_H B$ if and only if $\det(A) = \det(B)$.

3. Recall that any vector space is an abelian group under addition, and any vector subspace is a subgroup. In particular, let $G = \mathbb{R}^2$ and let H be a 1-dimensional subspace, i.e., a line through the origin. If $\mathbf{v} \in \mathbb{R}^2$ is any vector, the coset $H + \mathbf{v}$ is just a "parallel translate" of the line H.



4. Since Lagrange's theorem deals with finite groups, it would probably be helpful to look at one of those. Let's look at the dihedral group D_3, and let

$$H = \{i, r_1, r_2\},$$

the rotation subgroup. What are the cosets? There are only 2: H itself, and

$$Hm_1 = \{m_1, m_2, m_3\}.$$

(Note that we could also take m_2 or m_3 as a representative of the coset.)

Now that we've built up the appropriate machinery, let's go ahead and use it to formally state and prove Lagrange's theorem.

Theorem 2.6.15 (Lagrange)**.** *Let G be a finite group. For any subgroup $H \leq G$, $|H|$ divides $|G|$.*

Proof. Let $Ha_1, \ldots, Ha_k$ denote the distinct cosets of H in G. That is, $a_1, \ldots, a_k$ all represent different cosets of H, and these are all the cosets. We know that the cosets of H partition G, so

$$|G| = \#(Ha_1) + \cdots + \#(Ha_k). \tag{2.1}$$

(Here $\#$ means the *cardinality* of the set, or simply the number of elements in that set.) Therefore, it will be enough to show that each coset has the same number of elements as H.

We need to exhibit a bijection between H and Ha_i for each i. Fix i and define a function $f : H \to Ha_i$ by

$$f(h) = ha_i.$$

If we can prove that f is a bijection, then we will have

$$|H| = \#(Ha_i)$$

for all i. This is fairly straightforward: if $h_1, h_2 \in H$ with $f(h_1) = f(h_2)$, then

$$h_1 a_i = h_2 a_i,$$

which implies that $h_1 = h_2$ by right cancellation. Thus f is one-to-one. To see that it is onto, let ha_i be an arbitrary element of Ha_i; then $f(h) = ha_i$.

Thus all the cosets have the same number of elements, namely $|H|$, and (2.1) really says that

$$|G| = \underbrace{|H| + \cdots + |H|}_{k \text{ times}} = k|H|.$$

Therefore, $|H|$ does indeed divide $|G|$. □

The number that we called k in the proof is actually quite useful, and we will therefore give it a special name.

Definition 2.6.16. The number of distinct (right) cosets of H in G is called the **index** of H in G, denoted by

$$[G : H].$$

The set of all right cosets of H in G is denoted by G/H, so

$$\#(G/H) = [G : H].$$

Note that we can actually rephrase Lagrange's theorem in terms of the index: if G is a finite group and $H \leq G$, then

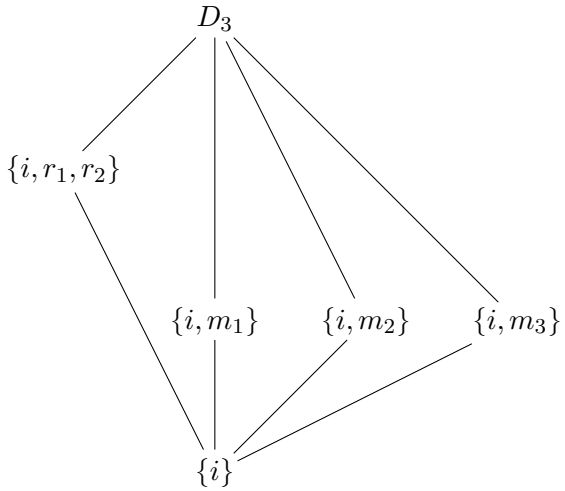
$$|G| = |H|[G : H].$$

We'll now begin to see that Lagrange's theorem has many very useful consequences. For one, it greatly simplifies the search for subgroups of a given finite group.

Example 2.6.17. Let's try to find all the subgroups of D_3. Since $|D_3| = 6$, we know that the only possible orders are the divisors of 6: $1, 2, 3$, and 6. We can then seek out subgroups by their order:

$$\begin{aligned} 1: &\quad \{i\} \\ 2: &\quad \{i, m_1\}, \{i, m_2\}, \{i, m_3\} \\ 3: &\quad \{i, r_1, r_2\} \\ 6: &\quad D_3 \end{aligned}$$

We can even draw a lattice diagram to illustrate the subgroup structure of D_3:



Lagrange's theorem also has a couple of easy yet powerful corollaries. One simply states that the order of any element divides the order of the group, which we already know for cyclic groups. The other tells us that groups of prime order are particularly special, and they behave in a very rigid way.

Corollary 2.6.18. *Let G be a finite group with $|G| = n$, and let $a \in G$. Then $|a|$ divides $n = |G|$, and*

$$a^n = e.$$

Proof. We have already seen that $|a| = |\langle a \rangle|$, and $\langle a \rangle$ is a subgroup of G, so its order divides $|G|$ by Lagrange's theorem. Therefore, $|a| \mid n$, and we can write $n = |a|m$ for some $m \in \mathbb{Z}$. Then

$$a^n = a^{|a|m} = \left(a^{|a|}\right)^m = e. \qquad \square$$

Corollary 2.6.19. *If G is a finite group of prime order p, then G is cyclic.*

Proof. Since p is prime, $p \geq 2$, and G contains at least one element a with $a \neq e$. By the previous corollary, $|a|$ divides p, and since $a \neq e$, $|a| \neq 1$. Since p is prime, we must have $|a| = p$, so a generates G. Thus G is cyclic. $\square$

In this proof, we showed that any nonidentity element of a group G with prime order is actually a generator for the group. This implies the following fact regarding subgroups of such groups.

Corollary 2.6.20. *If G is a finite group of prime order p, then G has no subgroups other than $\{e\}$ and G itself.*

For those of you who have taken a class in number theory, there are some interesting theorems from that field which can be rephrased in terms of group theory, and they then follow from Lagrange's theorem. In particular, one can consider Euler's theorem, Fermat's little theorem, and Wilson's theorem from this perspective.

Remark 2.6.21. We could have defined an alternative equivalence relation on G by $a \ {}_H\!\sim\ b$ if and only if $b^{-1}a \in H$. Then similar computations show that ${}_H\!\sim$ is an equivalence relation, and that the equivalence classes are

$${}_H[a] = aH = \{ah : h \in H\}.$$

These are called **left cosets**. We have used right cosets to prove Lagrange's theorem, but many texts use left cosets instead. Everything works perfectly well either way, but we'll stick to right cosets for now. However, we will soon need to consider

both left and right cosets—we will study a special type of subgroup called *normal subgroups*, and these will be defined[14] by the relationship between left and right cosets.

2.7 Homomorphisms

We've spent some time now analyzing the internal structure of groups. This was the beginning of our study of the relationships that various groups have to one other. In particular, we talked about what it means for one group to sit inside another as a subgroup. Now we'll return to the parallel problem that we've mentioned previously—that of classification.

We've talked quite a bit about the problem of determining when two groups are the same or different. We've even seen some examples of groups that appeared to be quite different, but they were fundamentally the same. Now we want to formalize this, and to finally talk about isomorphisms. Therefore, let's think about what it would take to formally show that two groups are isomorphic.

To see that two groups are isomorphic, we could use a group table argument, as we have done before. However, that could be difficult for large groups, or simply impossible for infinite groups. Therefore, there has to be a better way. Well, we would need to know first that the elements of one group correspond exactly to the elements of the other. The proper way to do this is to exhibit a bijection between them. This certainly isn't enough—we also need to know that the two groups have the same group structure. That is, the bijection in question needs to take into account the fact that the underlying sets are groups. Therefore, we need to discuss functions between groups that respect the corresponding binary operations. This leads to the idea of a **homomorphism**.

Definition 2.7.1. Let G_1 and G_2 be groups. A function $\varphi : G_1 \to G_2$ is called a **homomorphism** if

$$\varphi(ab) = \varphi(a)\varphi(b)$$

for all $a, b \in G_1$.

Example 2.7.2. 1. You've already seen an example of a homomorphism in linear algebra. If n and m are positive integers, we've seen that the vector spaces $\mathbb{R}^n$ and $\mathbb{R}^m$ are abelian groups under addition. Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation. Then T is a group homomorphism, since

$$T(v + w) = T(v) + T(w)$$

for all $v, w \in \mathbb{R}^n$.

[14] We will actually see that there are many ways to characterize normal subgroups, and this is one of them.

2. Define $\varphi : \mathbb{Z} \to \mathbb{Z}_n$ by
$$\varphi(a) = a \bmod n = [a]_n.$$
That is, $\varphi(a)$ is the remainder left when a is divided by n. Then φ is a homomorphism, since
$$\varphi(a+b) = [a+b]_n = [a]_n +_n [b]_n = \varphi(a) +_n \varphi(b),$$
since we defined $[a]_n +_n [b]_n$ to be $[a+b]_n$. This homomorphism is called **reduction mod** n.

3. If G_1 and G_2 are any groups, there is always a homomorphism $\varphi : G_1 \to G_2$ given by
$$\varphi(a) = e$$
for all $a \in G_1$, where e denotes the identity element of G_2. This is a homomorphism, since
$$\varphi(ab) = e = e \cdot e = \varphi(a)\varphi(b)$$
for all $a, b \in G_1$. This map is called the **trivial homomorphism**.

4. Let G be a group, and consider the identity map $\mathrm{id} : G \to G$. This is a homomorphism, since
$$\mathrm{id}(a+b) = a+b = \mathrm{id}(a) + \mathrm{id}(b)$$
for any $a, b \in G$. Thus there is always a homomorphism $\mathrm{id} : G \to G$, appropriately called the **identity homomorphism**.

5. Define $\varphi : \mathrm{GL_n}(\mathbb{R}) \to \mathbb{R}^\times$ by $\varphi(A) = \det(A)$. This is a homomorphism: if $A, B \in \mathrm{GL_n}(\mathbb{R})$, then
$$\varphi(AB) = \det(AB) = \det(A)\det(B) = \varphi(A)\varphi(B).$$

Often we'll be interested in knowing whether a homomorphism is one-to-one or onto, or both. There are special names for these sorts of things, and some are more commonly used than others.

Definition 2.7.3. Let $\varphi : G_1 \to G_2$ be a homomorphism.

- If φ is one-to-one, it is called a **monomorphism**.
- If φ is onto, it is called an **epimorphism**.
- If φ is bijective, it is called an **isomorphism**.
- An isomorphism from a group to itself is called an **automorphism**.

The first two of these are used sparingly, and we will usually say that the homomorphism is one-to-one (respectively, onto) rather than calling it a monomorphism (respectively, epimorphism). However, the last two are important, particularly the concept of an isomorphism. The main reason for this is that we can now define what it means for two groups to be isomorphic.

Definition 2.7.4. Two groups G_1 and G_2 are **isomorphic**, written

$$G_1 \cong G_2,$$

if there is an isomorphism $\varphi : G_1 \to G_2$.

Let's talk about some examples. In a few of these, we will formalize some of the "isomorphisms" that we've seen already.

Example 2.7.5. Consider the group $\mathbb{R}$ (under addition, of course), and let

$$\mathbb{R}^+ = \{a \in \mathbb{R} : a > 0\}.$$

Then it is not hard to check that $\mathbb{R}^+$ is a group under multiplication.[15] Define a homomorphism $\varphi : \mathbb{R} \to \mathbb{R}^+$ by

$$\varphi(a) = e^a = \exp(a).$$

This is in fact a homomorphism, since

$$\varphi(a+b) = e^{a+b} = e^a e^b = \varphi(a)\varphi(b).$$

Is it one-to-one? We know from calculus that the function e^x passes the horizontal line test, so it is one-to-one. What about surjectivity? Given a positive real number c, we could define $a = \log(c)$.[16] Then

$$\varphi(a) = e^a = e^{\log(c)} = c,$$

so φ is indeed surjective. Therefore, it is an isomorphism, and $\mathbb{R} \cong \mathbb{R}^+$. Note that we could have also defined an isomorphism in the other direction using the natural logarithm.

Example 2.7.6. The next example is an important one. Recall that we mentioned that cyclic groups are really boring, in that they are all either $\mathbb{Z}$ or $\mathbb{Z}_n$ in disguise. Let's prove this once and for all.

[15] In fact, it is a subgroup of the multiplicative group $\mathbb{R}^\times$ of nonzero real numbers.

[16] We should note here that the notation log denotes the natural logarithm. Mathematicians usually do not consider logarithms to any other bases, so it is generally assumed that the base is e.

- Suppose that $G = \langle a \rangle$ is an infinite cyclic group. Define $\varphi : \mathbb{Z} \to G$ by

$$\varphi(j) = a^j.$$

We claim that this is an isomorphism. First, it is a homomorphism, since

$$\varphi(i+j) = a^{i+j} = a^i a^j = \varphi(i)\varphi(j).$$

Is it one-to-one? If $\varphi(i) = \varphi(j)$, then

$$a^i = \varphi(i) = \varphi(j) = a^j.$$

Since a has infinite order, $a^i = a^j$ if and only if $i = j$. (We proved this in Proposition 2.3.18.) To see that it is onto, observe that if $b \in G$, then $b = a^j$ for some $j \in \mathbb{Z}$, and $\varphi(j) = a^j$. Thus φ is an isomorphism, and $G \cong \mathbb{Z}$.

- Now suppose that $G = \langle a \rangle$ is a finite cyclic group of order n. Define $\varphi : \mathbb{Z}_n \to G$ by

$$\varphi(j) = a^j.$$

This is again a homomorphism, since if $i, j \in \mathbb{Z}_n$, then

$$\varphi(i +_n j) = a^{[i+j]_n} = a^{i+j},$$

where we have used the fact that $[i+j]_n \equiv i + j \mod n$ (by definition, more or less). It follows that

$$\varphi(i +_n j) = a^i a^j = \varphi(i)\varphi(j).$$

Furthermore, φ is one-to-one: if $\varphi(i) = \varphi(j)$, then $a^i = a^j$ in G. But we proved in Proposition 2.3.17 that $a^i = a^j$ in a finite cyclic group if and only if $i \equiv j \mod n$, and $i, j \in \{0, 1, \ldots, n-1\}$, so this forces $i = j$. It is easy to see that φ is onto, since if $a^i \in G$, then $\varphi(i) = a^i$. Therefore, φ is an isomorphism, and $G \cong \mathbb{Z}_n$.

Therefore, we have shown that there is only one cyclic group of each order, up to isomorphism. In summary:

Theorem 2.7.7. *Let G be a cyclic group.*

1. *If G is infinite, then $G \cong \mathbb{Z}$.*
2. *If G is finite with $|G| = n$, then $G \cong \mathbb{Z}_n$.*

Since any group of prime order is cyclic by Corollary 2.6.19, we have also shown the following.

Theorem 2.7.8. *Let p be a prime number. If G is a group of order p, then $G \cong \mathbb{Z}_p$.*

2.7.1 Basic Properties of Homomorphisms

Before going further, we should mention some simple properties that are satisfied by homomorphisms. These will be very useful to have around, so we will get them out of the way now.

Proposition 2.7.9. *Let $\varphi : G_1 \to G_2$ be a homomorphism.*

(a) *If e and e' denote the identity elements of G_1 and G_2, respectively, then*

$$\varphi(e) = e'.$$

(b) *For all $a \in G_1$,*

$$\varphi(a^{-1}) = \varphi(a)^{-1}.$$

(c) *More generally, if $a \in G_1$, then*

$$\varphi(a^n) = \varphi(a)^n$$

for all $n \in \mathbb{Z}$.

Proof. (a) Observe that

$$\varphi(e) = \varphi(e^2) = \varphi(e)^2,$$

so $\varphi(e)$ is an idempotent. But you proved in an earlier exercise that a group only has one idempotent, namely the identity. Therefore, $\varphi(e) = e'$. We could also argue as follows: for any $a \in G_1$, we have

$$\varphi(a)\varphi(e) = \varphi(ae) = \varphi(a) = \varphi(a)e'.$$

Left cancellation then implies that $\varphi(e) = e'$.

(b) If $a \in G_1$, then we have

$$\varphi(a)\varphi(a^{-1}) = \varphi(aa^{-1}) = \varphi(e) = e'$$

by part (a). Therefore, $\varphi(a^{-1}) = \varphi(a)^{-1}$.

(c) We handled the case $n = 0$ in part (a). Suppose then that $n \in \mathbb{Z}^+$. We will proceed via mathematical induction. For the base case of $n = 1$, we certainly have

$$\varphi(a^1) = \varphi(a) = \varphi(a)^1.$$

Suppose then that the result holds for $n-1$. Then

$$\varphi(a^n) = \varphi(aa^{n-1}) = \varphi(a)\varphi(a^{n-1}) = \varphi(a)\varphi(a)^{n-1} = \varphi(a)^n.$$

by the inductive hypothesis. It then follows that if $n < 0$, we have

$$\varphi(a^{-n}) = \varphi((a^{-1})^n) = \varphi(a^{-1})^n = \varphi(a)^{-n}$$

so the result holds for all $n \in \mathbb{Z}$. □

This last proposition might make you wonder how the order of an element behaves when a homomorphism is applied. In particular, you may be curious as to whether the order of $\varphi(a)$ relates to the order of a somehow. Indeed it does.

Proposition 2.7.10. *Let G_1 and G_2 be groups, and let $\varphi : G_1 \to G_2$ be a homomorphism. If $a \in G_1$ has finite order, then $|\varphi(a)|$ divides $|a|$.*

Proof. Let $n = |a|$. Then $a^n = e$, so

$$\varphi(a)^n = \varphi(a^n) = \varphi(e) = e.$$

Thus $\varphi(a)^n = e$, so we must have that $|\varphi(a)|$ divides $|a|$. □

The last result is a little more high-brow, but it will be good to have around (and it isn't too hard to handle).

Theorem 2.7.11. *Let G_1, G_2, and G_3 be groups, and let $\varphi : G_1 \to G_2$ and $\psi : G_2 \to G_3$ be homomorphisms.*

(a) *The composition $\psi \circ \varphi$ is a homomorphism.*

(b) *If ψ and φ are both isomorphisms, then $\psi \circ \varphi$ is an isomorphism.*

(c) *If φ is an isomorphism, then φ^{-1} is an isomorphism.*

Proof. (a) Let $a, b \in G_1$. Then

$$\psi \circ \varphi(ab) = \psi(\varphi(ab)) = \psi(\varphi(a)\varphi(b)) = \psi(\varphi(a))\psi(\varphi(b)),$$

since φ and ψ are both homomorphisms.

(b) If φ and ψ are isomorphisms, then $\psi \circ \varphi$ is a homomorphism by (a), and it is a bijection since φ and ψ both are, so $\psi \circ \varphi$ is an isomorphism.

(c) Suppose that $x, y \in G_2$. Then $x = \varphi(a)$ and $y = \varphi(b)$ for some $a, b \in G_1$, so

$$\varphi^{-1}(xy) = \varphi^{-1}(\varphi(a)\varphi(b)) = \varphi^{-1}(\varphi(ab)) = ab = \varphi^{-1}(x)\varphi^{-1}(y).$$

Thus φ^{-1} is an isomorphism as well. □

Note that this last part implies that if φ is an isomorphism, then $|\varphi(a)| = |a|$ for all $a \in G_1$.

Proposition 2.7.12. *If $\varphi : G_1 \to G_2$ is an isomorphism, then $|\varphi(a)| = |a|$ for all $a \in G_1$.*

Proof. Suppose first that $a \in G_1$ has finite order. Then we showed in Proposition 2.7.10 that $|\varphi(a)|$ divides $|a|$. But φ^{-1} is also a homomorphism, so $|a| = |\varphi^{-1}(\varphi(a))|$ must divide $|\varphi(a)|$. Since both integers divide each other, they must be the same, and $|a| = |\varphi(a)|$.

Now note that Proposition 2.7.10 also implies that a has finite order if and only if $\varphi(a)$ does. It follows that a has infinite order if and only if $\varphi(a)$ does, so we are done. □

2.8 The Symmetric Group Redux

We're now going to take a short detour into the symmetric group, which we introduced quite some time ago. We've put off the details in order to develop some more general tools for working with groups. Never fear; we'll be going back into the abstraction very soon, but right now we'll try to have a little fun with S_n.

There is a good reason for shifting gears so abruptly—S_n is really the most fundamental finite nonabelian group, so it would be very useful to understand its structure. We will see via Cayley's theorem that every group is really a permutation group, in a certain sense. In fact, groups were originally thought of only as permutation groups. It was Cayley who first gave the abstract definition of a group and then proved his eponymous theorem. Therefore, if we could somehow understand all of the subgroups of S_n, then we would understand all finite groups. This is an ambitious goal, and it is far too much to ask for. However, understanding the symmetric group will give us information on it and other groups that are easily realized as permutations (like D_n, for example).

2.8.1 Cycle Decomposition

When we originally talked about S_n, we introduced some compact notation for writing down a permutation, called **two-line notation**: if $\sigma \in S_n$, then

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & \cdots & n \\ \sigma(1) & \sigma(2) & \sigma(3) & \cdots & \sigma(n) \end{pmatrix}.$$

This is obviously a nice, short way to write down an element of S_n, and multiplying two permutations in two-line notation is relatively straightforward. However, because of its simplicity it can obscure the structure of σ, and it's not good enough

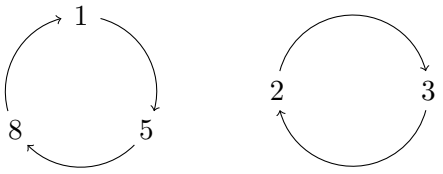
from the standpoint of algebra. For example, it we look at the permutation $\sigma \in S_8$ given by

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 5 & 3 & 2 & 4 & 8 & 6 & 7 & 1 \end{pmatrix},$$

what is really going on?

- σ fixes 4, 6, and 7.
- 2 and 3 only interact with each other, in the sense that σ sends 2 to 3 and 3 to 2.
- 1, 5, and 8 only interact with each other.

Thus we've written down a bunch of extra information that isn't really necessary, and we've obscured some of the structure that σ possesses in the process. The relevant data is the fact that σ "cycles through" 1, 5, and 8 and it "cycles through" 2 and 3, while fixing all other numbers:



With this in mind, perhaps there is another way to write down σ that truly captures what σ is doing. There is, and it is called the **cycle decomposition** of σ.

Let's be a little more precise about what we've done above: if we apply σ repeatedly to 1, we have

$$\begin{aligned} \sigma(1) &= 5 \\ \sigma^2(1) &= \sigma(5) = 8 \\ \sigma^3(1) &= \sigma(8) = 1 \\ \sigma^4(1) &= \sigma(1) = 5 \end{aligned}$$

and so on. We can use the shorthand notation

$$(1\ 5\ 8)$$

to indicate that σ maps 1 to 5, 5 to 8, and 8 back to 1. A permutation written this way is called a *cycle*. Similarly, we have

$$\begin{aligned} \sigma(2) &= 3 \\ \sigma^2(2) &= \sigma(3) = 2 \end{aligned}$$

and so on, so we write the cycle

$$(2\ 3)$$

to describe this "part" of σ. All the other numbers in the set $\{1, 2, \ldots, 8\}$ are fixed by σ, which we will denote by writing a cycle of length 1 for each such number. Thus in the notation we are developing, we can decompose σ and write it as a product of cycles:

$$\sigma = (1\ 5\ 8)(2\ 3)(4)(6)(7).$$

It is actually customary to suppress the numbers that are fixed by σ, since they information that they encode is extraneous. Therefore, we will simply write

$$\sigma = (1\ 5\ 8)(2\ 3),$$

with it understood that any numbers not appearing are fixed.

We've been throwing around the word "cycle" without properly defining it. You probably have the idea already, but we should still be rigorous.

Definition 2.8.1. A permutation of the form

$$\sigma = (i_1\ i_2\ \cdots\ i_k),$$

where $\sigma(i_m) = i_{m+1}$ for $1 \leq m < k$ and $\sigma(i_k) = i_1$ is called a k**-cycle**.

Note that there may be many different ways of expressing the same permutation in cycle notation. For example, the cycle $(5\ 8\ 1)$ is the same as $(1\ 5\ 8)$, since both cycles indicate that 1 maps to 5, 5 maps to 8, and 8 maps to 1. However, the cycles $(1\ 5\ 8)$ and $(1\ 8\ 5)$ are different—the first says that 1 maps to 5, while the second maps 1 to 8. It is important to remember that a cycle records the order in which the numbers are "hit" as we iterate σ.

Cycles are important in that they form the building blocks of elements of S_n. When finding the cycle decomposition of a permutation σ, we are really factoring σ into cycles, in much the same way that one factors integers into primes.

Example 2.8.2. Let's consider the permutation $\tau \in S_9$ given by

$$\tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 4 & 9 & 6 & 3 & 7 & 1 & 8 & 2 & 5 \end{pmatrix}.$$

How do we write τ as a product of cycles? We start with 1, and we find the *cycle determined by it*:

$$\tau = \left(\begin{array}{ccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ \downarrow & & \searrow & \downarrow & & \searrow & & & \\ 4 & 9 & 6 & 3 & 7 & 1 & 8 & 2 & 5 \end{array} \right)$$

Thus we have the cycle

$$(1\ 4\ 3\ 6).$$

Now we go to the next smallest integer, which is 2, and we find its cycle:

$$\tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 4 & 9 & 6 & 3 & 7 & 1 & 8 & 2 & 5 \end{pmatrix}$$

so the cycle is

$$(2\ 9\ 5\ 7\ 8).$$

Thus

$$\tau = (1\ 4\ 3\ 6)(2\ 9\ 5\ 7\ 8).$$

Example 2.8.3. Let's find the cycle decomposition of $\sigma \in S_5$, where

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 1 & 5 & 4 \end{pmatrix}.$$

Well, the cycle determined by 1 is

$$(1\ 2\ 3),$$

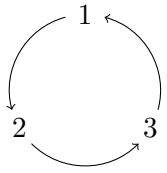
and the smallest integer left is 4, whose cycle is

$$(4\ 5).$$

Therefore,

$$\sigma = (1\ 2\ 3)(4\ 5).$$

Remark 2.8.4. Note that we have written our permutations in such a way that the cycles always start with the smallest possible integer. This is the standard way to write permutations in cycle notation, but it is not necessary to do so. For example, the cycles (1 2 3) and (3 1 2) are the same. You could think of them both as encodings of the picture



As long as you write down the elements in the correct order (i.e., when going around the cycle depicted above), it doesn't matter where you choose to start.

What do you notice about the cycles in the two examples above? They have no integers in common—the precise way to say this is that they are **disjoint**. It's not much of a stretch to see that what we have done in our examples thus far would work for any $\sigma \in S_n$. In other words, we have essentially observed the following result:

Theorem 2.8.5. *Every element of S_n can be written as a product of disjoint cycles.*

It's easy enough to see that this result holds when looking at specific examples, but you might wonder how we would prove such a statement in general. Well, there's actually a very orderly way of doing it that involves working with a certain equivalence relation. In particular, if we define a relation $\sim$ on the set $\{1, 2, \ldots, n\}$ by

$$i \sim j \quad \iff \quad \sigma^m(i) = j$$

for some $m \in \mathbb{Z}^+$, then we claim that $\sim$ is an equivalence relation. Since $\sigma^{n!} = \mathrm{id}$ as a consequence of Lagrange's theorem, we have $\sigma^{n!}(i) = i$ for all i. Thus $i \sim i$, and $\sim$ is reflexive. If $i \sim j$ with $\sigma^m(i) = j$, then we have

$$\sigma^{n!-m}(j) = \sigma^{n!-m}(\sigma^m(i)) = \sigma^{n!}(i) = i,$$

so $j \sim i$. Hence $\sim$ is symmetric. Finally, suppose $i \sim j$ and $j \sim k$, with $\sigma^m(i) = j$ and $\sigma^l(j) = k$. Then

$$\sigma^{l+m}(i) = \sigma^l(\sigma^m(i)) = \sigma^l(j) = k,$$

so $i \sim k$, and it follows that $\sim$ is transitive. Thus $\sim$ is an equivalence relation. It is then clear that the equivalence classes would all have the following form:

$$[i] = \{i, \sigma(i), \sigma^2(i), \sigma^3(i), \ldots\}$$

for each $i \in \{1, 2, \ldots, n\}$. But the elements of this equivalence class are precisely the numbers that appear in the cycle that contains i! Therefore, we can use these equivalence classes to break up σ into cycles, which shows that each element of S_n has a cycle decomposition. Moreover, these equivalence classes partition the set $\{1, 2, \ldots, n\}$, so these cycles are necessarily disjoint.

Now we know that the cycles really are the elementary pieces out of which the elements of S_n are built. A fair question to ask would be: what's the big deal? The fact of the matter is that the cycle decomposition gives us lots of information about S_n. For one, it simplifies calculations greatly. The next theorem tells us that multiplication of disjoint cycles is very nice, though we won't prove it.

Theorem 2.8.6. *Disjoint cycles commute. That is, if $\sigma, \tau \in S_n$ are disjoint cycles, then $\sigma\tau = \tau\sigma$.*

As is often the case when proving things about elements of S_n, the proofs are fairly technical and not very enlightening. Therefore, we'll trust our intuition and see how the proof would work via some examples.

Example 2.8.7. Let $\sigma, \tau \in S_7$ be given by

$$\sigma = (1\ 4\ 3)$$

and

$$\tau = (2\ 5\ 6).$$

To see that σ and τ commute, we would need to check that $\sigma\tau(j) = \tau\sigma(j)$ for all $j \in \{1, 2, 3, 4, 5, 6, 7\}$. We can ignore 7, since it is fixed by both σ and τ. For 1, we have

$$\sigma\tau(1) = \sigma(1) = 4$$

and

$$\tau\sigma(1) = \tau(4) = 4,$$

since τ fixes both 1 and 4. Similarly, for 2 we would have

$$\sigma\tau(2) = \sigma(5) = 5$$

and

$$\tau\sigma(2) = \tau(2) = 5.$$

The calculation for the other integers is similar. Hopefully this is enough to convince you that disjoint cycles commute in general. The proof is really nothing more than the observations that we have made here, but it is simply harder to write down in general.

The fact that disjoint cycles commute buys us a lot in terms of computations. For one, when we multiply permutations together, we can switch the order of two disjoint cycles, which can greatly simplify the procedure. Also, it lets us gain information about permutations much more quickly that we could if we were using two-line notation. For example, what if we wanted to find the inverse of a permutation? If we write $\sigma = \sigma_1\sigma_2\cdots\sigma_m$, where $\sigma_1, \ldots, \sigma_m$ are disjoint cycles, then the inverse is

$$\sigma^{-1} = (\sigma_m)^{-1}\cdots(\sigma_1)^{-1},$$

so we simply need to know how to find the inverse of a cycle. This is easy: if we have the cycle

$$\tau = (i_1\ i_2\ \cdots\ i_k),$$

the inverse is simply

$$\tau^{-1} = (i_1\ i_k\ i_{k-1}\ \cdots\ i_2).$$

Example 2.8.8. Let's find the inverse of $\sigma = (1\ 4\ 7\ 12)(3\ 9\ 2\ 5)$ in S_{12}. Doing so amounts to finding the inverses of the two cycles that make up σ. The inverse of the first is

$$(1\ 12\ 7\ 4),$$

and the second is

$$(3\ 5\ 2\ 9).$$

Therefore,

$$\sigma^{-1} = (3\ 5\ 2\ 9)(1\ 12\ 7\ 4) = (1\ 12\ 7\ 4)(3\ 5\ 2\ 9),$$

since the cycles are disjoint.

Here's another place where cycle decomposition is useful. Given an element of S_n, how do we find its order? We could raise it to higher and higher powers until we get the identity, but this could take an annoyingly long time. The answer becomes much easier when we write the permutation as a product of disjoint cycles. First we need to know what the order of a cycle is.

Proposition 2.8.9. *The order of a k-cycle is k. That is, the order of a cycle is simply the length of the cycle.*

It is not too hard to see how to prove this, but it is unnecessarily cumbersome to write down the proof.

Now that we know the order of a cycle, we claim that we can compute the order of any element of S_n. Suppose that

$$\sigma = \sigma_1\sigma_2\cdots\sigma_m$$

is a product of cycles, where σ_j is a cycle of length k_j. If we raise σ to any power d, what do we get? Since the disjoint cycles commute, we have

$$\sigma^d = \sigma_1^d\sigma_2^d\cdots\sigma_k^d.$$

For σ^d to equal the identity, we need to have σ_j^d to equal the identity for all j. What's the smallest d that works? We would need that $|\sigma_j| \mid d$ for all j, so d is a multiple of k_j for all j. Thus the smallest number that works is the *least common multiple* of the orders of the cycles. That is,

$$|\sigma| = \operatorname{lcm}(k_1, \ldots, k_m).$$

Example 2.8.10. Define $\sigma \in S_{15}$ by

$$\sigma = (1\ 4\ 7\ 3\ 9)(5\ 8)(6\ 12\ 10).$$

What is the order of σ? The orders of the disjoint cycles that make up σ are 5, 2, and 3, so

$$|\sigma| = \operatorname{lcm}(5, 2, 3) = 30.$$

Obviously the order would be very hard to find otherwise, since we would have to raise σ to the 30th power.

2.8.2 Application to Dihedral Groups

We will see very soon that, at least in theory, a good understanding of the symmetric group will help us understand other groups as well. (This is a consequence of a result called Cayley's theorem.) Of course it would be difficult to understand all groups in this way, but the ones that can readily be realized as permutation groups will be easier to understand. A natural class of groups to which this applies is the family of *dihedral groups*.

Example 2.8.11. We've seen in the past that we can identify any dihedral group D_n with a subgroup of S_n. To do this, we simply write down the permutation that corresponds to the action of an element of D_n on the vertices of a regular n-gon. For example, if we are looking at the rotation r_2 in D_3, we have the corresponding element

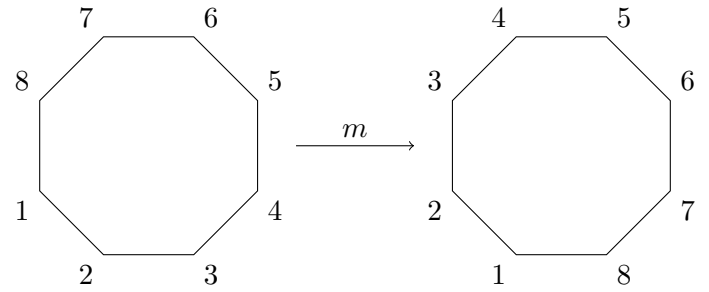
$$r_2 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}.$$

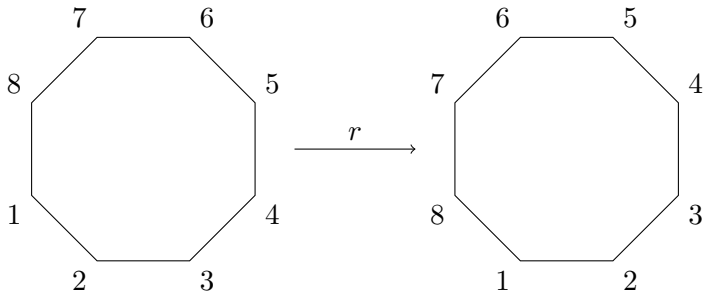
In cycle notation, this is

$$r_2 = (1\ 3\ 2).$$

Therefore, cycle notation gives us a much easier way of computing products of elements in D_n. We simply convert the symmetries to permutations, write them as products of disjoint cycles, and then multiply.

As another example, let's go up to a larger dihedral group, say D_8. Then if we wanted to compute the product of the reflection m across the line bisecting the side $(1, 2)$ and the counterclockwise rotation r through 45 degrees.





We can write

$$m = (1\ 2)(3\ 8)(4\ 7)(5\ 6)$$

and

$$r = (1\ 2\ 3\ 4\ 5\ 6\ 7\ 8),$$

so

$$mr = (1\ 2)(3\ 8)(4\ 7)(5\ 6)(1\ 2\ 3\ 4\ 5\ 6\ 7\ 8) = (1)(2\ 8)(3\ 7)(4\ 6)(5).$$

If we drop the one-cycles, we get

$$mr = (2\ 8)(3\ 7)(4\ 6),$$

and we see that mr is the reflection across the line connecting vertices 1 and 5. Thus we can see that cycle notation has very practical uses, too, since it can simplify computations in permutation groups.

2.8.3 Cayley's Theorem

Before we began attacking the symmetric group again, we were looking at homomorphisms and isomorphisms. In particular, we formalized a couple of the isomorphisms that we had mentioned in the past. There was another example that we ran into earlier, which we will now examine in a more proper way. It will lead to a very important theorem in group theory.

Example 2.8.12. We've just been discussing the idea of representing elements of the dihedral group as permutations. This is something we did a long time ago, when we observed via group tables that $D_3 \cong S_3$. Without going into too much detail, let's define an actual isomorphism. We'll do it element by element, though this is a terrible method in general. Define $\varphi : D_3 \to S_3$ by

$$\begin{aligned} \varphi(i) &= \iota & \varphi(m_1) &= (2\ 3) \\ \varphi(r_1) &= (1\ 2\ 3) & \varphi(m_2) &= (1\ 3) \\ \varphi(r_2) &= (1\ 3\ 2) & \varphi(m_3) &= (1\ 2) \end{aligned}$$

Then φ is clearly one-to-one and onto, but is it a homomorphism? Yes, though it would be tedious to check all of the products. For example, we have

$$\varphi(r_1 m_1) = \varphi(m_3) = (1\ 2),$$

while

$$\varphi(r_1)\varphi(m_1) = (1\ 2\ 3)(2\ 3) = (1\ 2).$$

You can verify the other products by comparing the two group tables. This is essentially what we did when we originally encountered this phenomenon, and we simply gave each element and its image under φ the same label. We have thus formally shown that $D_3 \cong S_3$.

Remark 2.8.13. We mentioned that in general, $D_n \not\cong S_n$. The main obstacle is that these two groups have different orders when $n \geq 4$. However, we can always define a *monomorphism* $\varphi : D_n \to S_n$. For example, one of the exercises requires you to convert two elements of D_4 to permutations, and then check that this conversion plays well with the group operations. This can be done in general, and instead of having $D_n \cong S_n$, we can say that D_n is isomorphic to a *subgroup* of S_n.

The statement in the previous remark holds even more generally—we'll see that *any* group is isomorphic to a subgroup of S_X for some X. In particular, if $|G| = n$, then G is isomorphic to a subgroup of S_n. This result is called **Cayley's theorem**, and we'll prove it shortly.

Before we can actually prove Cayley's theorem, there's one small result about homomorphisms that we'll need. If $\varphi : G_1 \to G_2$ is a group homomorphism, the **image** of φ is defined to be

$$\varphi(G_1) = \{y \in G_2 : y = \varphi(a) \text{ for some } a \in G_1\}.$$

Then we will prove that the image of any homomorphism is always a subgroup of the codomain.

Proposition 2.8.14. *If $\varphi : G_1 \to G_2$ is a homomorphism, then $\varphi(G_1)$ is a subgroup of G_2.*

Proof. We've already seen that if e_1 and e_2 denote the identity elements of G_1 and G_2, respectively, then $e_2 = \varphi(e_1)$, so the identity of G_2 is in $\varphi(G_1)$. Also, if $a \in G_1$, then $\varphi(a)^{-1} = \varphi(a^{-1}) \in \varphi(G_1)$, so $\varphi(G_1)$ is closed under taking inverses. Therefore, we just need to check that $\varphi(G_1)$ is closed. Let $x, y \in \varphi(G_1)$. Then $x = \varphi(a)$ and $y = \varphi(b)$ for some $a, b \in G_1$, so

$$xy = \varphi(a)\varphi(b) = \varphi(ab) \in \varphi(G_1).$$

Thus $\varphi(G_1) \leq G_2$. □

Theorem 2.8.15 (Cayley)**.** *Let G be a group. There is a one-to-one homomorphism $\varphi : G \to S_G$, where S_G denotes the set of all bijections from G to itself. Consequently, G is isomorphic to a subgroup of S_G. In particular, if $|G| = n$, then G is isomorphic to a subgroup of the symmetric group S_n.*

Proof. We need to figure out some way of defining a function $\varphi : G \to S_G$. That is, given an element $a \in G$, we need to define a bijection $L_a : G \to G$. For $a \in G$, define

$$L_a(g) = ag$$

for all $g \in G$. Is $L_a \in S_G$, i.e., is L_a a bijection? If $g_1, g_2 \in G$ and $L_a(g_1) = L_a(g_2)$, then

$$ag_1 = ag_2,$$

so $g_1 = g_2$ by cancellation. Thus L_a is one-to-one. Now let's show that L_a is onto. If $g \in G$, then

$$L_a(a^{-1}g) = a(a^{-1}g) = g,$$

so L_a is in fact a bijection. Now we define $\varphi : G \to S_G$ by

$$\varphi(a) = L_a.$$

We need to check that φ is a homomorphism and that it is one-to-one. First, if $a, b \in G$, then $\varphi(ab) = L_{ab}$, and for any $g \in G$, we have

$$L_{ab}(g) = (ab)g = a(bg) = aL_b(g) = L_a(L_b(g)) = L_a \circ L_b(g),$$

so $L_{ab} = L_aL_b$. That is, $\varphi(ab) = \varphi(a)\varphi(b)$, so φ is a homomorphism. To see that it is one-to-one, suppose that $a, b \in G$ and that $\varphi(a) = \varphi(b)$. Then $L_a = L_b$, so for all $g \in G$,

$$ag = L_a(g) = L_b(g) = bg.$$

By cancellation, $a = b$, so φ is one-to-one. Thus $G \cong \varphi(G) \leq S_G$, so G is isomorphic to a subgroup of S_G.

If G is finite, with $|G| = n$, then S_G is simply S_n. Thus any finite group is isomorphic to a subgroup of S_n, where n is the order of the group. □

What we have shown in Cayley's theorem is that *every* group can be viewed as a group of permutations. More specifically, G can be identified with a group of permutations of the elements of G itself. This is the primary importance of Cayley's theorem—it tells us that every group is a symmetry group.

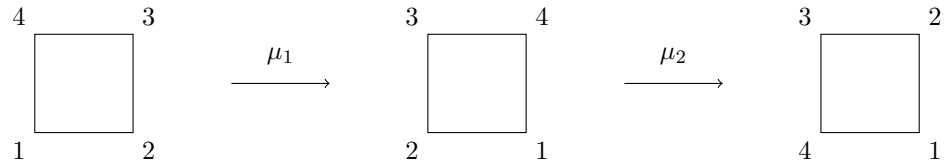
2.8.4 Even and Odd Permutations and the Alternating Group

Now that we've talked about disjoint cycle decompositions, we'll discuss another way of writing permutations. To do so, we'll think of a permutation as a procedure which "switches" integers one pair at a time.

Example 2.8.16. Let's view the dihedral group D_4 as a group of permutations in S_4, and consider the rotation through 180°, which we have denoted by r_2. We know that we can write r_2 as the product of two 2-cycles:

$$r_2 = (1\ 3)(2\ 4).$$

Therefore, r_2 is really given by interchanging two pairs of vertices. The same thing happens if we look at the rotation r_1, which we can view as a composition of two reflections, which we will denote μ_1 and μ_2:



Thus we can write r_1, which is described by the 4-cycle

$$r_1 = (1\ 2\ 3\ 4),$$

instead as a product of 2-cycles:

$$r_1 = (1\ 3)(1\ 2)(3\ 4).$$

This signifies that the rotation can be obtained simply by swapping certain pairs of vertices.

It turns out that we can do what we have done in the example more generally—we can describe a permutation by simply interchanging elements two at a time. For example, given the 3-cycle $(1\ 3\ 2)$, we can write

$$(1\ 3\ 2) = (1\ 2)(1\ 3).$$

These 2-cycles have a special name.

Definition 2.8.17. A 2-cycle is called a **transposition**.

Much as we saw that we could write a permutation as a product of disjoint cycles, we have the following result regarding transpositions.

Theorem 2.8.18. *Every permutation in S_n can be written as a product of transpositions.*

As with other results regarding S_n, we won't prove this, but we'll do some examples that will give us an idea of how the proof would go.

Example 2.8.19. Let $\sigma = (1\ 3\ 2\ 8)(4\ 6\ 5) \in S_8$. We can write the first cycle as

$$(1\ 3\ 2\ 8) = (1\ 8)(1\ 2)(1\ 3),$$

and the second cycle is

$$(4\ 6\ 5) = (4\ 5)(4\ 6).$$

Thus

$$\sigma = (1\ 8)(1\ 2)(1\ 3)(4\ 5)(4\ 6).$$

Remark 2.8.20. Let's make a few observations at this point.

1. The transpositions that make up a permutation **need not be disjoint**.
2. We really only need to know how to write a cycle as a product of transpositions.
3. If $\tau = (i_1\ i_2\ i_3 \cdots i_k)$ is a k-cycle, we can write
$$\tau = (i_1\ i_k)(i_1\ i_{k-1})\cdots(i_1\ i_3)(i_1\ i_2).$$
4. The decomposition into transpositions is not unique—we could toss in a bunch of extra transpositions and get the same permutation. To see this in action, observe that we could write the 3-cycle (1 3 2) as
$$(1\ 2)(1\ 3)(3\ 2)(3\ 2),$$
where we have effectively multiplied by the identity permutation on the right. For a less trivial version of this phenomenon, observe that the 3-cycle (1 3 2) is the same as (3 2 1), which decomposes into transpositions as
$$(3\ 1)(3\ 2).$$
Thus by writing the permutation in two different (but equivalent) ways, we have obtained two different decompositions into transpositions.[17]

The decomposition into transpositions gives us information about the structure of S_n. The relevant information here is the **parity** of a permutation.

Definition 2.8.21. A permutation $\sigma \in S_n$ is called **even** if it can be written as a product of an even number of transpositions, and **odd** if it is a product of an odd number of transpositions.

We're going to see that S_n is partitioned into even and odd permutations. To do this, we'll first have to see that the notions of even and odd are in fact mutually exclusive. Since the decomposition into transpositions is far from unique, there's no reason *a priori* that a permutation couldn't be both even and odd. Fortunately, it is impossible.

Theorem 2.8.22. *A permutation $\sigma \in S_n$ cannot be both even and odd.*

[17] Thanks to Sahil Seekond for pointing out this more interesting argument in class.

Proof. We're going to give a proof that uses linear algebra. Let $\{e_1, e_2, \ldots, e_n\}$ be the standard ordered basis for $\mathbb{R}^n$. We can view σ as a permutation of these basis vectors, via the reordering

$$\{e_1, e_2, \ldots, e_n\} \mapsto \left\{e_{\sigma(1)}, e_{\sigma(2)}, \ldots, e_{\sigma(n)}\right\}.$$

The trick that we will employ is to consider the determinant of the matrix whose columns are given by the reordered basis vectors. Note that the matrix associated to the original basis is the identity matrix:

$$\begin{pmatrix} e_1 & e_2 & \cdots & e_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix} = I.$$

We need to consider the effect that σ has on the determinant of this matrix. Recall that the act of interchanging two columns of a matrix negates the determinant, so a single transposition has the effect of multiplying the determinant by -1. Therefore, if σ is odd, say

$$\sigma = \sigma_1 \sigma_2 \cdots \sigma_k$$

where each σ_i is a transposition and k is odd, then we are performing column interchanges on the identity matrix, so

$$\det\left(e_{\sigma(1)}, e_{\sigma(2)}, \ldots, e_{\sigma(n)}\right) = (-1)^k \det(e_1, e_2, \ldots, e_n) = (-1)^k \det(I) = -1.$$

On the other hand, if σ is even, say

$$\sigma = \tau_1 \tau_2 \cdots \tau_m$$

with each τ_i a transposition and m even, we have

$$\det\left(e_{\sigma(1)}, e_{\sigma(2)}, \ldots, e_{\sigma(n)}\right) = (-1)^m \det(I) = 1.$$

These two determinants have to agree, so σ cannot be both even and odd. □

Remark 2.8.23. This proof uses a couple of powerful ideas from algebra that we won't get to discuss in this class. The first is the idea of a **group action**: we let a group act on a set (i.e., elements of the group are viewed as bijections/permutations of the set), and we can gain information about both the group and the set in the process. The second involves the particular case when a group acts on a vector space. This leads to a very popular branch of algebra called **representation theory**.

Now that we know that nothing fishy happens regarding the parity of permutations, we can proceed.

Definition 2.8.24. The set of all permutations in S_n is denoted by A_n, and it is called the **alternating group** on n letters.

We've referred to A_n as a group, so we had better check it. We will verify it directly by checking the axioms, and then we will begin the next section by exploring a slightly more clever option. This will also lead us nicely into the next topic.

Theorem 2.8.25. $A_n \leq S_n$.

Proof. The identity permutation can be thought of as a product of zero 2-cycles, so it is automatically even. (If you don't like this argument, we can also write $\iota = (1\ 2)(1\ 2)$, so it is even.) Also, the product of two even permutations is clearly even: if $\sigma, \tau \in A_n$ are written as

$$\sigma = \sigma_1\sigma_2\cdots\sigma_k, \quad \tau = \tau_1\tau_2\cdots\tau_m,$$

where σ_i, τ_j are all transpositions and k, m are even, then

$$\sigma\tau = \sigma_1\cdots\sigma_k\tau_1\cdots\tau_m$$

is a product of $k+m$ transpositions. Since k and m are even, so is $k+m$, so $\sigma\tau \in A_n$, and A_n is closed. Finally, if $\sigma \in A_n$ has the form

$$\sigma = (i_1\ j_1)(i_2\ j_2)\cdots(i_k\ j_k)$$

with k even, then

$$\sigma^{-1} = (i_k\ j_k)\cdots(i_1\ j_1),$$

so σ^{-1} is even as well. Therefore, $A_n \leq S_n$. □

We'll do a little more with A_n in the future, since it is a very important subgroup of S_n. Right now, we'll make one small observation: what is the order of A_n? To answer this, it helps to look at cosets. Let $\sigma \in S_n$ be a permutation that is not in A_n; then σ must be odd. The coset

$$A_n\sigma = \{\tau\sigma : \tau \in A_n\}$$

must consist entirely of odd permutations. This is because the product of an even permutation with an odd permutation is odd: if

$$\sigma = \sigma_1\sigma_2\cdots\sigma_k$$

and

$$\tau = \tau_1\tau_2\cdots\tau_m,$$

where σ_i and τ_j are all transpositions, k is odd, and m is even, then

$$\tau\sigma = \tau_1 \cdots \tau_m \sigma_1 \cdots \sigma_k$$

is a product of $k+m$ transpositions, so $\tau\sigma$ is odd. In fact, $A_n\sigma$ contains all the odd permutations: if $\tau \in S_n$ is odd, then $\tau\sigma^{-1} \in A_n$, so $\tau\sigma^{-1}\sigma \in A_n\sigma$.[18] Thus $A_n\sigma$ consists of precisely the odd permutations in S_n. Since every permutation is either even or odd, these are the only two right cosets of A_n in S_n. Therefore,

$$|A_n| = \frac{|S_n|}{2} = \frac{n!}{2}.$$

This shows that exactly half of the permutations in S_n are even, and half are odd.

2.9 Kernels of Homomorphisms

If we had done things in a slightly different order, we could have found a much slicker way to see that A_n is a subgroup of S_n. (Actually, we could have established even more than that, but we'll get there shortly.) Despite this, the associated ideas will still allow us to transition nicely into the next topic.

First, we claim that there is a homomorphism defined on S_n which is intimately connected with A_n. To construct it, we first have to note that the set $\{-1, 1\}$ (viewed as integers, or real numbers, or whatever you like) forms a group under multiplication. With this in hand, we define a function $\operatorname{sgn} : S_n \to \{-1, 1\}$ by

$$\operatorname{sgn}(\sigma) = \begin{cases} 1 & \text{if } \sigma \text{ is even} \\ -1 & \text{if } \sigma \text{ is odd.} \end{cases}$$

We claim that sgn is a homomorphism (called the **sign** of a permutation). Let $\sigma, \tau \in S_n$. If σ, τ are both even, then $\sigma\tau$ is even as well, so

$$\operatorname{sgn}(\sigma\tau) = 1 = 1 \cdot 1 = \operatorname{sgn}(\sigma)\operatorname{sgn}(\tau).$$

Similarly, if both are odd, then $\sigma\tau$ is even, so

$$\operatorname{sgn}(\sigma\tau) = -1 \cdot -1 = \operatorname{sgn}(\sigma)\operatorname{sgn}(\tau).$$

Finally, if one is even and one is odd (suppose without loss of generality that σ is odd), then $\sigma\tau$ is odd, so

$$\operatorname{sgn}(\sigma\tau) = -1 \cdot 1 = \operatorname{sgn}(\sigma)\operatorname{sgn}(\tau).$$

[18] Just as with the other two cases, it is easy to check that the product of two odd permutations is even.

Thus sgn is a homomorphism. That's all well and good, but what role does A_n play here? Everything in A_n is mapped to 1, which is the identity in the group $\{-1, 1\}$:

$$A_n = \{\sigma \in S_n : \operatorname{sgn}(\sigma) = 1\}.$$

In general, the set of all elements that map to the identity under a homomorphism has special significance, as we will see shortly.

Aside 2.9.1. The sign homomorphism might seem a bit contrived—all it really does is tell us whether a permutation is even or odd, which we know how to do anyway. However, it does have uses. We are about to see that homomorphisms and subgroups are tied together very closely, and the sign homomorphism gives us a glimpse of this phenomenon with regard to the alternating group. More generally, the sign of a permutation comes up often in computations in certain branches of mathematics. For example, it is used in combinatorics, where people study *inversions* in permutations. Combinatorialists write their permutations in *one-line notation*: given the permutation

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 1 & 5 & 2 & 4 & 6 \end{pmatrix} \in S_6,$$

they would simply strip off the second line and write

$$\sigma = 315246.$$

They think of permutations as sequences of numbers from the set $\{1, 2, \ldots, n\}$ in which each number appears exactly once. An inversion appears when two numbers occur out of their natural order. For example, 3 appears before 1, so this counts as an inversion. Similarly, 3 comes before 2, and 5 comes before both 2 and 4, so there are four inversions in all. On the other hand, if we write σ as a product of transpositions, we have

$$\sigma = (1\ 2)(1\ 4)(1\ 5)(1\ 8).$$

There are four transpositions, which is conveniently the number $N(\sigma)$ of inversions in σ. Since the decomposition into transpositions is far from unique, this need not hold in general.[19] However, we do have

$$\operatorname{sgn}(\sigma) = (-1)^{N(\sigma)},$$

so the number of inversions is tied to the parity of σ.

The sign of a permutation also appears in the definition of the determinant. In all likelihood you learned to compute the determinant of an $n \times n$ matrix via cofactor

[19] It is not merely a coincidence, though. There is always a way to write a permutation σ as a product of a *minimal* number of *simple* transpositions, which are transpositions of the form $(i\ i{+}1)$. This number is equal to $N(\sigma)$.

expansions. This definition is very practical for hand calculations, but there is an alternative formulation that can be useful in theoretical computations: if $A = (a_{ij})$ is an $n \times n$ matrix, then

$$\det(A) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) a_{\sigma(1)1} a_{\sigma(2)2} \cdots a_{\sigma(n)n}.$$

This definition is quite messy, but it can be handy in proofs. (It really says that the determinant is something called an n-multilinear alternating form.)

There is a structure related to any group homomorphism which allows us to gather information about the homomorphism. In particular, it will give us an easy way of checking if a homomorphism is one-to-one. This structure is called the **kernel** of the homomorphism.

Definition 2.9.2. Let $\varphi : G_1 \to G_2$ be a group homomorphism. The **kernel** of φ is defined to be

$$\ker \varphi = \{x \in G_1 : \varphi(x) = e_2\}.$$

Example 2.9.3. You've already seen several examples of kernels, albeit in disguise.

1. Let $\operatorname{sgn} : S_n \to \{\pm 1\}$ denote the sign homomorphism from above. Then $\ker(\operatorname{sgn}) = A_n$.

2. Recall that if $T : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation, then T is a group homomorphism. Its kernel is

   $$\ker T = \{x \in \mathbb{R}^n : T(v) = 0\},$$

   which you know better as the **null space** of T.

3. Let $\det : \mathrm{GL_n}(\mathbb{R}) \to \mathbb{R}^\times$ denote the determinant map. Then

   $$\ker(\det) = \{A \in \mathrm{GL_n}(\mathbb{R}) : \det(A) = 1\},$$

   which we know as the special linear group $\mathrm{SL_n}(\mathbb{R})$.

4. Let $\varphi : \mathbb{Z} \to \mathbb{Z}_n$ denote the homomorphism given by reduction mod n. What is the kernel of φ? Well,

   $$\ker \varphi = \{a \in \mathbb{Z} : \varphi(a) = 0\},$$

   and $\varphi(a) = 0$ exactly when $a \equiv 0 \bmod n$. That is, $\varphi(a) = 0$ if and only if $n \mid a$, so

   $$\ker \varphi = n\mathbb{Z}.$$

5. Let G_1 and G_2 be groups, and let $\varphi : G_1 \to G_2$ be the trivial homomorphism, so $\varphi(a) = e$ for all $a \in G_1$. Then

$$\ker \varphi = G_1,$$

   or the whole group G_1.

6. Let G_1 and G_2 be groups, and suppose that $\varphi : G_1 \to G_2$ is a one-to-one homomorphism. Then

$$\ker \varphi = \{e_1\},$$

   since $\varphi(x) = e_2$ implies that $x = e_1$ by the injectivity of φ.

What do you notice about the kernel in all of these examples? These kernels are all subgroups, and it is not too far-fetched to think that this might hold in general. It's also not too hard to prove.

Proposition 2.9.4. *Let $\varphi : G_1 \to G_2$ be a homomorphism. Then*

$$\ker \varphi \leq G_1.$$

Proof. Certainly $e_1 \in \ker \varphi$, since $\varphi(e_1) = e_2$, and if $a \in \ker \varphi$, then

$$\varphi(a^{-1}) = \varphi(a)^{-1} = e_2^{-1} = e_2,$$

so $a^{-1} \in \ker \varphi$. Now suppose that $a, b \in \ker \varphi$. Then

$$\varphi(ab) = \varphi(a)\varphi(b) = e_2 e_2 = e_2,$$

so $ab \in \ker \varphi$. Thus $\ker \varphi$ is a subgroup of G_1. □

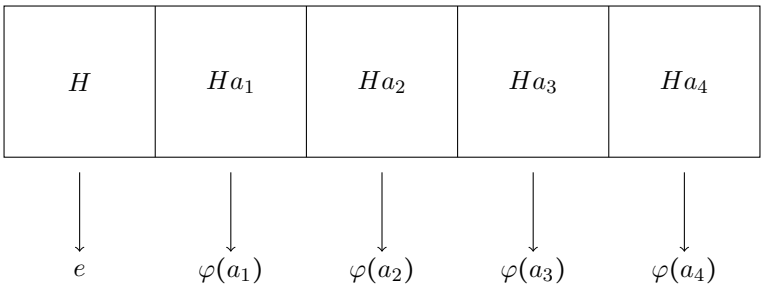
We mentioned that the kernel encodes information about whether a homomorphism is one-to-one or not. In fact, it gives us a measure of how badly φ fails to be one-to-one, in a certain sense. To be more precise, suppose that $a \in G_1$ and $x \in \ker \varphi$. Then

$$\varphi(xa) = \varphi(x)\varphi(a) = e\varphi(a) = \varphi(a),$$

so a and xa have the same image under φ. That is, φ is *constant on the right coset* $(\ker \varphi)a$. On the other hand, suppose that $a, b \in G_1$ with $\varphi(a) = \varphi(b)$. Then

$$e = \varphi(a)\varphi(b)^{-1} = \varphi(a)\varphi(b^{-1}) = \varphi(ab^{-1}),$$

so $ab^{-1} \in \ker \varphi$. This means that $(\ker \varphi)a = (\ker \varphi)b$, so $\varphi(a) = \varphi(b)$ if and only if a and b belong to the same coset of $\ker \varphi$. This tells us that if φ fails to be one-to-one, it does so in a very uniform way, in the sense that the preimage of any point $y \in \varphi(G_1)$ consists of an entire coset of $\ker \varphi$. In other words, φ crushes each coset down to a single point: if $H = \ker \varphi$, then



This phenomenon also makes it fairly easy to see a nice criterion for checking the injectivity of a homomorphism.

Theorem 2.9.5. *A homomorphism $\varphi : G_1 \to G_2$ is one-to-one if and only if* $\ker \varphi = \{e_1\}$.

Proof. If φ is one-to-one, then certainly $\ker \varphi = \{e_1\}$. On the other hand, suppose that $\ker \varphi = \{e_1\}$ and $a, b \in G_1$ with $\varphi(a) = \varphi(b)$. Then we have just seen that $ab^{-1} \in \ker \varphi$, so $ab^{-1} = e_1$. But then $a = b$, so φ is one-to-one. □

2.10 Quotient Groups and Normal Subgroups

The next natural thing to study after kernels is the notion of a **normal subgroup**, of which any kernel is an example. Unfortunately, this idea seems somewhat unmotivated, so we will introduce it alongside the concept of a **quotient group** (or **factor group**). These groups will make it clear why normal subgroups are necessary. Quotient groups are kind of strange—they don't make a lot of sense when you first see them, they're not so bad once you've become accustomed to them. Therefore, we'll have to tread lightly and be very careful in digesting these ideas. As usual, we'll try to motivate the conversation with an example that we already know very well.

2.10.1 The Integers mod n

We've mentioned in the past that the way in which we defined $\mathbb{Z}_n$ was somewhat awkward, and that there was a more proper way of doing things. We're now ready to discuss this more precise description of $\mathbb{Z}_n$.

Recall that since $n\mathbb{Z} \leq \mathbb{Z}$, we have an equivalence relation on $\mathbb{Z}$ given by $a \sim b$ if and only if $a - b \in n\mathbb{Z}$. This amounts to saying that

$$a \sim b \iff a \equiv b \bmod n.$$

The equivalence classes are the (right) cosets:

$$n\mathbb{Z} + a = \{b \in \mathbb{Z} : b \equiv a \bmod n\},$$

which are sometimes called **congruence classes**. There are n of these classes, namely

$$n\mathbb{Z}+0,\ n\mathbb{Z}+1,\ldots,\ n\mathbb{Z}+(n-1).$$

Now we claim that we can actually define an binary operation on the set of right cosets—that is, we will combine two cosets together to produce a new coset. We do this by addition:

$$(n\mathbb{Z}+a)+(n\mathbb{Z}+b)=n\mathbb{Z}+(a+b).$$

We simply add two cosets by adding their representatives and then taking the coset that corresponds to the result. What is the potential caveat here? We need to know that this operation is *well-defined*: if we take two *different* representatives for the cosets, do we get the same answer? If $a' \equiv a \bmod n$ and $b' \equiv b \bmod n$, then $a'+b' \equiv a+b \bmod n$, so

$$(n\mathbb{Z}+a')+(n\mathbb{Z}+b')=n\mathbb{Z}+(a'+b')=n\mathbb{Z}+(a+b).$$

Thus addition is well-defined on cosets. We now claim further that this turns the set of left cosets into a group. What is the identity? Why, it's $n\mathbb{Z}+0$, since

$$(n\mathbb{Z}+0)+(n\mathbb{Z}+a)=n\mathbb{Z}+(0+a)=n\mathbb{Z}+a$$

for all $a \in \mathbb{Z}$. How about inverses? If $a \in \mathbb{Z}$, then

$$(n\mathbb{Z}+a)+(n\mathbb{Z}+(-a))=n\mathbb{Z}+(a-a)=n\mathbb{Z}+0,$$

so the inverse of $n\mathbb{Z}+a$ is $n\mathbb{Z}+(-a)$. If we call the set of (right) cosets

$$\mathbb{Z}/n\mathbb{Z} \quad \text{(Read: "}\mathbb{Z} \bmod n\mathbb{Z}\text{")},$$

then we have shown that $\mathbb{Z}/n\mathbb{Z}$ is a group under addition. We will see that this group simply gives another way of viewing $\mathbb{Z}_n$. It also not hard to see that the two "versions" of $\mathbb{Z}_n$ that we have produced are isomorphic, but we will wait to prove it until we have introduced a tool known as the First Homomorphism Theorem.

Note that there is nothing really different here from what we have done in the past. The "add and reduce" rule was essentially addition in $\mathbb{Z}/n\mathbb{Z}$, but with the added step of reducing the result to the smallest positive coset representative.

2.10.2 General Quotient Groups

Now we could ask a more general question: if G is a group and $H \leq G$, and we let G/H denote the set of (right) cosets of H in G, can we turn G/H into a group (in the same way as we did with $\mathbb{Z}/n\mathbb{Z}$)? You may be wondering why we would even want to do so in the first place. It sort of made sense for $\mathbb{Z}/n\mathbb{Z}$, since it just gave us a better way of viewing modular arithmetic. But why should it be useful in the general setting? Well, sometimes we don't care what happens "up to" the subgroup

H, meaning that we won't bother distinguishing elements of G which differ by an element of H. This is particularly true when H is the kernel of some homomorphism.

Suppose that we are trying to classify groups via isomorphism, and we have a group G that we would like to realize in a more recognizable way. We may not be lucky enough to be able to define an isomorphism with a more familiar group, but we may be able to define a homomorphism φ from some group G' *onto* G. If we can identify the kernel of φ, then we will be able to say

$$G \cong G'/\ker\varphi.$$

In forming the quotient, we've basically eliminated the one-to-one defect of φ. Of course we will have to prove that this actually happens, which we will do when we prove the *First Isomorphism Theorem*.

Now that we've talked about the "why," let's talk about how we would define a general quotient group. Say we have a group G and a subgroup $H \leq G$. Let's define an operation on G/H in the same way that we did for $\mathbb{Z}/n\mathbb{Z}$: given two cosets Ha and Hb, define

$$(Ha)(Hb) = H(ab).$$

We saw that in the example, we needed to know that the operation was well-defined. Is the same true here? If we pick different representatives, say $ha \in Ha$ and $gb \in Hb$, then we have

$$(Hha)(Hgb) = H(hagb),$$

and we need this to equal $H(ab)$. For these two cosets to coincide, we would need

$$hagb(ab)^{-1} \in H.$$

But

$$hagb(ab)^{-1} = hagbb^{-1}a^{-1} = haga^{-1},$$

and $h \in H$, so it would be enough to have $aga^{-1} \in H$. In other words, we have shown that:

> The operation on G/H is well-defined if and only if $aga^{-1} \in H$ for all $g \in H$ and $a \in G$, or $aHa^{-1} \subseteq H$ for all $a \in G$.

Therefore, to make this operation well-defined, we will simply *require* that our subgroup H be a special kind of subgroup, one for which the above condition holds.

Definition 2.10.1. A subgroup $N \leq G$ is **normal** if $aNa^{-1} \subseteq N$ for all $a \in G$. We write

$$N \trianglelefteq G$$

to denote that N is a normal subgroup of G.

Now that we've defined normal subgroups, we can see that we've been working toward the following result.

Theorem 2.10.2. *If G is a group and $N \trianglelefteq G$, then G/N is a group with respect to the operation*

$$(Na)(Nb) = N(ab).$$

Proof. We already saw that the normality of N is exactly what is needed to ensure that the operation is well-defined. Also, the operation is associative since the operation on G is. What is the identity? It's Ne, since

$$(Na)(Ne) = N(ae) = Na = N(ea) = (Ne)(Na).$$

The inverse is also exactly what you would expect:

$$(Na)(Na^{-1}) = Ne = (Na^{-1})(Na)$$

for all $a \in G$. □

Here's another question regarding quotient groups. If G is a finite group and $N \trianglelefteq G$, what is the order of G/N? We know that the number of left cosets of N in G is the index of N in G, which is

$$[G : N] = \frac{|G|}{|N|}$$

by Lagrange's theorem. Therefore,

$$|G/N| = \frac{|G|}{|N|}.$$

Now would be the natural place to talk about some examples of quotient groups. However, this would be hard to do without having examples of normal subgroups. Therefore, we'll take a brief foray into normal subgroups first, and then produce some specific quotient groups.

2.10.3 Normal Subgroups

Now that we've talked about why normal subgroups are important, let's investigate some of their properties. We'll start with some examples.

Example 2.10.3.

1. Let G be any group. The trivial subgroups $\{e\}$ and G are normal.

2. Let G be an *abelian* group. Are there more normal subgroups than just the trivial ones? Yes—*any* subgroup of G is normal. If $H \leq G$, then for any $h \in H$, we have
$$aha^{-1} = aa^{-1}h = h$$
for all $a \in G$, so $aHa^{-1} \subseteq H$. (In fact, it is easy to see that the two sets are equal in this case.)

There is a whole plethora of subgroups that are normal—kernels of homomorphisms are always normal subgroups.

Proposition 2.10.4. *Let $\varphi : G_1 \to G_2$ be a homomorphism. Then for any $a \in G_1$, we have $a(\ker\varphi)a^{-1} \subseteq \ker\varphi$.*

Proof. Let $x \in \ker\varphi$. We simply need to verify that $axa^{-1} \in \ker\varphi$. Well,
$$\varphi(axa^{-1}) = \varphi(a)\varphi(x)\varphi(a^{-1}) = \varphi(a)e_2\varphi(a^{-1}) = \varphi(a)\varphi(a)^{-1} = e_2,$$
so $axa^{-1} \in \ker\varphi$. Thus $a(\ker\varphi)a^{-1} \subset \ker\varphi$. □

Example 2.10.5. 1. We can deduce that $\mathrm{SL_n}(\mathbb{R}) \trianglelefteq \mathrm{GL_n}(\mathbb{R})$, since $\mathrm{SL_n}(\mathbb{R}) = \ker(\det)$. After we develop some more tools, we'll be able to identify the quotient group $\mathrm{GL_n}(\mathbb{R})/\mathrm{SL_n}(\mathbb{R})$ in a more recognizable form.

2. The alternating group A_n is a normal subgroup of S_n, since A_n is the kernel of the sign homomorphism. In this case, we can actually compute the quotient group explicitly. Note that S_n/A_n has exactly two elements, and we have seen that any group of order 2 must be isomorphic to $\mathbb{Z}_2$. Alternatively, we defined the sign homomorphism $\mathrm{sgn} : S_n \to \{1, -1\}$, where the latter group is endowed with multiplication. (Since it has two elements, it is isomorphic to $\mathbb{Z}_2$ as well.) We will see soon that there is a very natural way to show that $S_n/A_n \cong \{1, -1\}$ as well.

Now let's prove some things about normal subgroups. The following result tells us that there are multiple ways in which we can describe normal subgroups.

Proposition 2.10.6. *Let G be a group and $N \leq G$. The following statements are equivalent:*

(a) *$N \trianglelefteq G$.*

(b) *For all $a \in G$, $aNa^{-1} = N$.*

(c) *Every left coset of N is also a right coset, in the sense that $Na = aN$ for all $g \in G$.*

Proof. We'll start by showing that (a) implies (b). Let $a \in G$; since $N \trianglelefteq G$, $aNa^{-1} \subseteq N$. On the other hand, if $g \in N$, then

$$g = a(a^{-1}ga)a^{-1} \in aNa^{-1},$$

since $a^{-1}Na \subseteq N$. Therefore, $N \subseteq aNa^{-1}$, and (b) is proven.

To see that (b) implies (c), let $a \in G$. Then

$$Na = \{ga : g \in N\} = \{aha^{-1}a : h \in N\} = \{ah : h \in N\} = aN.$$

Thus the left and right cosets of N coincide.

The proof that (c) implies (a) is similar. If $a \in G$, then $aN = Na$, so for any $g \in G$, $aga^{-1} = haa^{-1} = h$ for some $h \in N$. Thus $N \trianglelefteq G$. □

Condition (c) in the theorem above will give us a new way of seeing that certain subgroups are normal. It will also allow us to talk about a nonexample of a normal subgroup.

Example 2.10.7. Consider $K = \{i, m_1\}$ as a subset of D_3. Then $K \leq D_3$, but is it normal? The left cosets are

$$K = \{i, m_1\},\ r_1K = \{r_1, m_3\},\ r_2K = \{r_2, m_2\},$$

while the right cosets are

$$K = \{i, m_1\},\ Kr_1 = \{r_1, m_2\},\ Kr_2 = \{r_2, m_3\}.$$

These do not coincide, so K is not normal in D_3.

Consider D_3, with $H = \{i, r_1, r_2\}$. Then we claim that $H \trianglelefteq D_3$: the only right cosets are

$$H = \{i, r_1, r_2\},\ Hm_1 = \{m_1, m_3, m_2\},$$

while the left cosets are

$$H = \{i, r_1, r_2\},\ m_1H = \{m_1, m_2, m_3\}.$$

The left and right cosets are the same, so H is normal in D_3. Since it's normal, we can form the quotient group D_3/H. There are only two cosets, so the quotient group is automatically isomorphic to $\mathbb{Z}_2$.

Example 2.10.8. We'll soon learn of a very easy way of computing quotient groups, but let's try to do another one explicitly first. Let $G = \mathbb{Z}_{15}$, and let $H = \langle 5 \rangle = \{0, 5, 10\}$. Then $H \trianglelefteq G$ (since G is abelian), and there are five cosets:

$$H + 0 = \{0, 5, 10\}$$
$$H + 1 = \{1, 6, 11\}$$

$$H + 2 = \{2, 7, 12\}$$
$$H + 3 = \{3, 8, 13\}$$
$$H + 4 = \{4, 9, 14\}.$$

Normally we would need to figure out exactly how the operation works on G/H, but things are easy here. We know that $|G/H| = 5$, so there is only one possibility: $G/H \cong \mathbb{Z}_5$. (Of course we could also compute explicitly with the elements of the quotient group, and we would see that the group behaves exactly as $\mathbb{Z}_5$ should.)

2.10.4 The First Isomorphism Theorem

Now we'll state and prove a result that has been alluded to already, the *First Isomorphism Theorem*. (It's also called the *Fundamental Homomorphism Theorem* by some authors, including Saracino.) This theorem is one of the most important in algebra, and it ties together the concepts of homomorphism, kernel, and quotient group. It basically says that an onto homomorphism gives an isomorphism by factoring out the kernel. Indeed, we've already seen the core idea of the proof, albeit in picture form, and we've also established certain elements of the proof.

$\ker\varphi$	$\longrightarrow$	$\varphi(e)$
$(\ker\varphi)a_1$	$\longrightarrow$	$\varphi(a_1)$
$(\ker\varphi)a_2$	$\longrightarrow$	$\varphi(a_2)$
$(\ker\varphi)a_3$	$\longrightarrow$	$\varphi(a_3)$
$(\ker\varphi)a_4$	$\longrightarrow$	$\varphi(a_4)$

This picture illustrates the key idea that a homomorphism φ is constant on each right coset of its kernel. Therefore, we can build a well-defined map $\tilde{\varphi}$ on the quotient group $G/\ker\varphi$, which turns out to be a homomorphism.

Theorem 2.10.9 (First Isomorphism Theorem)**.** *Let G and G' be groups, and let $\varphi : G \to G'$ be a homomorphism. Suppose that φ is onto and let $N = \ker\varphi$. Then*

$$G/N \cong G',$$

and the isomorphism is given by $\tilde{\varphi} : G/N \to G'$, where

$$\tilde{\varphi}(Na) = \varphi(a).$$

Proof. We first have to check that $\tilde{\varphi}$ is even well-defined. But we saw this already: φ is constant on each coset, so if $Na = Nb$, then $\varphi(a) = \varphi(b)$ and $\tilde{\varphi}$ is well-defined. When we did this, we also saw that $\tilde{\varphi}$ was one-to-one: if $\tilde{\varphi}(Na) = \tilde{\varphi}(Nb)$, then $\varphi(a) = \varphi(b)$, which means that $\varphi(ab^{-1}) = e$. Thus $ab^{-1} \in \ker\varphi = N$, so $Na = Nb$. Also, $\tilde{\varphi}$ is onto because φ is: if $x \in G'$, then $x = \varphi(a)$ for some $a \in G$, and

$$\tilde{\varphi}(Na) = \varphi(a) = x.$$

(This is really the content of the figure above—since φ is onto and it is constant on each coset, it sets up a bijection between the quotient $G/\ker\varphi$ and G'.) Therefore, we really only need to check that φ is a homomorphism. If $Na, Nb \in G/N$, then we have

$$\tilde{\varphi}((Na)(Nb)) = \tilde{\varphi}(Nab) = \varphi(ab),$$

while

$$\tilde{\varphi}(Na)\tilde{\varphi}(Nb) = \varphi(a)\varphi(b) = \varphi(ab).$$

Since these are the same, we see that $\tilde{\varphi}$ is a homomorphism. It now follows that $G/N \cong G'$. □

Now let's put this theorem to use in identifying some quotient groups with more familiar ones.

Example 2.10.10. 1. Let $\varphi : \mathbb{Z} \to \mathbb{Z}_n$ be given by reduction mod n. We already proved that this was a homomorphism, and it is onto: if $a \in \mathbb{Z}_n = \{0, 1, \ldots, n-1\}$, then $\varphi(a) = a$. Also, we saw that $\ker\varphi = n\mathbb{Z}$, so the Fundamental Homomorphism Theorem tells us that

$$\mathbb{Z}/n\mathbb{Z} = \mathbb{Z}/\ker\varphi \cong \mathbb{Z}_n.$$

We already observed this fact, but now it becomes even easier when we hit it with the hammer that is gifted to us by the First Isomorphism Theorem.

2. We mentioned earlier that if $\mathrm{sgn} : S_n \to \{1, -1\}$ denotes the sign homomorphism, then $A_n = \ker(\mathrm{sgn})$. Furthermore, sgn is onto, so the First Isomorphism Theorem tells us that

$$S_n/A_n \cong \{1, -1\}.$$

3. We previously verified that $\det : \mathrm{GL_n}(\mathbb{R}) \to \mathbb{R}^\times$ is a homomorphism, and its kernel is $\mathrm{SL_n}(\mathbb{R})$. In fact, it is an epimorphism: if $\alpha \in \mathbb{R}$ is any real number, then

$$\det\begin{pmatrix} \alpha & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix} = \alpha.$$

Thus the First Isomorphism Theorem applies, and it tells us that

$$\mathrm{GL_n}(\mathbb{R})/\mathrm{SL_n}(\mathbb{R}) \cong \mathbb{R}^\times.$$

4. Let's do one more example from linear algebra. Suppose that $n > m$, and define a linear transformation $T : \mathbb{R}^n \to \mathbb{R}^{n-m}$ by

$$T(x_1, \ldots, x_{n-m}, x_{n-m+1}, \ldots, x_n) = (x_1, \ldots, x_{n-m}).$$

   Then we know that T is a group homomorphism, and its kernel is

$$\ker T = \{\mathbf{x} \in \mathbb{R}^n : x_1 = x_2 = \cdots = x_{n-m} = 0\},$$

   which is an m-dimensional subspace of $\mathbb{R}^n$. We could identify this subspace with $\mathbb{R}^m$, and the First Isomorphism Theorem tells us that

$$\mathbb{R}^n/\mathbb{R}^m = \mathbb{R}^n/\ker T \cong \mathbb{R}^{n-m}.$$

5. Define $\varphi : \mathbb{Z}_{15} \to \mathbb{Z}_5$ by $\varphi(a) = a \bmod 5$. What is the kernel of φ? If $\varphi(a) = 0$, then $a \equiv 0 \bmod 5$, so $5 \mid a$. Thus

$$\ker \varphi = \{0, 5, 10\},$$

   and since this is a group of order 3, it is isomorphic to $\mathbb{Z}_3$. In other words, we can think of $\mathbb{Z}_3$ as a subgroup of $\mathbb{Z}_{15}$. What is the quotient group $\mathbb{Z}_{15}/\mathbb{Z}_3$? Well, the First Isomorphism Theorem says that

$$(\mathbb{Z}_{15})/(\mathbb{Z}_3) \cong \mathbb{Z}_5.$$

   It's not too hard to generalize this: if $d \mid n$, then $\mathbb{Z}_d \leq \mathbb{Z}_n$, and

$$\mathbb{Z}_n/\mathbb{Z}_d \cong \mathbb{Z}_{n/d}.$$

Remark 2.10.11. The astute reader is probably wondering why this theorem was called the *First* Isomorphism Theorem, and you're probably guessing if there are others. There are indeed two more theorems, conveniently called the Second and Third Isomorphism Theorems. They are less commonly used than the First Isomorphism Theorem, so we won't really discuss them. In fact, they are both proven using the first theorem.

There is another theorem that is related to the three homomorphism theorems, and it is often discussed in tandem with them. It is usually called the **Correspondence Theorem** (or sometimes the Fourth Isomorphism Theorem) and it says that the subgroups of a quotient group are related to the subgroups of the original group in a very rigid way. If G is a group and $N \trianglelefteq G$, there is a one-to-one correspondence between the subgroups of G/N and subgroups of G containing N. Let $\psi : G \to G/N$ denote the projection onto the quotient group. If $H \leq G$ and $N \subseteq H$, then we obtain a subgroup H' of G/N from H by

$$H' = \psi(H) = \{y \in G/N : y = \psi(x) \text{ for some } x \in H\}.$$

In other words, we just take the image of H in G/N. To go the other way, suppose that $K \leq G/N$. Define the inverse image of K to be

$$K' = \psi^{-1}(K) = \{x \in G : \psi(x) \in K\}.$$

Then K' is a subgroup of G which contains N. This sets up a bijection between the subgroups of G/N and the subgroups of G containing N. Additionally, this correspondence preserves normality: if $H \trianglelefteq G$, then $\pi(H) \trianglelefteq G/N$, and if $K \trianglelefteq G/N$, then $\psi^{-1}(K) \trianglelefteq G$.

2.10.5 Aside: Applications of Quotient Groups

We've now talked about quotient groups, namely how to build them and to identify them via the Fundamental Homomorphism Theorem. This will really be what we will need them for, but it would be good to know that there is a less abstract application of quotient groups. We'll talk about a couple of these applications, but with the First Isomorphism Theorem still looming large in the background.

The core idea of the First Isomorphism Theorem is the following: we took a homomorphism that was not necessarily one-to-one, and we made it one-to-one by "identifying" the elements that had a common image. This is another way of thinking about quotient groups: when forming a quotient group, one identifies all the elements belonging to a particular coset. That is, we don't care to distinguish the individual elements of the coset. This can be a helpful way of thinking about quotient groups, and it is often how quotient groups arise when working less abstractly than we have been here.

Example 2.10.12. Sometimes groups arise naturally in other branches of math as a means of distinguishing objects. For example, topologists often try to classify surfaces by associating groups[20] to them. The classification problem is then reduced to a question about classification of groups, which is much easier to answer in general. Sometimes the group that arises is nonabelian, which can make it hard to work with. However, it is possible to "approximate" the group with an abelian one. This abelian group is called the **abelianization** of the original group, and it arises as a quotient group.

Let G be a group. What do we need in order to be able to say that G is abelian? We would need that for all $a, b \in G$,

$$ab = ba.$$

If we multiply on the right by a^{-1}, then b^{-1}, we see that this is equivalent to saying that

$$aba^{-1}b^{-1} = e.$$

[20] If you're curious, the constructions that algebraic topologists usually consider are the *fundamental group*, the *homology* groups, and the *cohomology* groups.

Thus in order to make G abelian, we need to identify elements of the form $aba^{-1}b^{-1}$ with e; that is, we need to form the quotient by a subgroup that contains all the $aba^{-1}b^{-1}$. We don't want to lose too much, so we'd like to take the smallest possible subgroup. Therefore, we'll take

$$[G,G] = \langle aba^{-1}b^{-1} \mid a,b \in G \rangle.$$

The elements of the form $aba^{-1}b^{-1}$ are called **commutators**, and the set $[G,G]$ is the smallest subgroup containing all the commutators, called the **commutator subgroup**. One should keep in mind that $[G,G]$ does not consist only of commutators; its elements are products of commutators. It can be checked that $[G,G]$ is a normal subgroup, and that the quotient

$$G_{\text{ab}} = G/[G,G]$$

is an abelian group. It is the "best possible" abelian group associated to G, in some sense, and it is called the **abelianization of** G.

Example 2.10.13. This example is one that is near and dear to my heart, but it is a little less pertinent than the first one. (Therefore, you may feel free to stop reading now if you want to stick to algebra.) One of the branches of math that I study is a field called *harmonic analysis*. This area basically deals with trying to do calculus on groups. (This isn't a great explanation, but it's reasonably correct.) If you've taken a course on differential equations, you may have already seen some of the ideas that we'll discuss here. People often try to solve differential equations using objects called Fourier series, which give a new way of representing **periodic functions**. A function $f : \mathbb{R} \to \mathbb{R}$ is called periodic if there is a real number p such that

$$f(x) = f(x+p)$$

for all $x \in \mathbb{R}$. It's called periodic because the function repeats after every interval of length p. For an example, we could see that sine and cosine are both periodic functions, with period 2π. Let's suppose, for simplicity, that we're dealing with a function of period 1; that is, it repeats after every interval of length one. Then

$$f(x) = f(x+n)$$

for all $x \in \mathbb{R}$ and $n \in \mathbb{Z}$. In other words, if two real numbers differ by an integer, then the function takes the same value at those two points. It would thus make sense to identify two real numbers when they differ by an integer, or when they belong to the same coset of $\mathbb{Z}$ in $\mathbb{R}$. This would lead us to look at the quotient group $\mathbb{R}/\mathbb{Z}$, and to instead consider functions on this new group.

This big question that we are left with now is how to identify $\mathbb{R}/\mathbb{Z}$ in a more recognizable way. Trying to look at cosets would be quite unwieldy, since there are uncountably many of them. Indeed, there is one for every $x \in [0,1]$, and in fact 0

and 1 are identified as well. How could we view $\mathbb{R}/\mathbb{Z}$, at least in a geometric sense? If we identify the endpoints of an interval, we get something that looks like a circle. In fact, $\mathbb{R}/\mathbb{Z}$ is isomorphic to something called the **circle group**. Define

$$\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\},$$

which is the unit circle in the complex plane. It turns out that $\mathbb{T}$ is actually a group under the usual multiplication of complex numbers. If $a, b \in \mathbb{T}$, then

$$|ab| = |a||b| = 1 \cdot 1 = 1,$$

so $\mathbb{T}$ is closed. Also, if $a \in \mathbb{T}$, then $a^{-1} = 1/a$ is also in $\mathbb{T}$, and $1 \in \mathbb{T}$, so $\mathbb{T}$ is a group. We want to use the First Isomorphism Theorem, so we need to define a homomorphism from $\mathbb{R}$ onto $\mathbb{T}$. We can do this by

$$\varphi(x) = e^{2\pi i x} = \cos(2\pi x) + i\sin(2\pi x).$$

For those who are familiar with complex numbers, it is not hard to check that this maps into $\mathbb{T}$ and that it is a homomorphism: if $x, y \in \mathbb{R}$, then

$$\varphi(x + y) = e^{2\pi i(x+y)} = e^{2\pi i x} e^{2\pi i y} = \varphi(x)\varphi(y).$$

It's easier to see that φ is onto by writing a complex number in polar form as

$$z = \cos(2\pi x) + i\sin(2\pi x),$$

and observing that $\varphi(x) = z$. Thus $\varphi : \mathbb{R} \to \mathbb{T}$ is onto. What is its kernel? Well,

$$\ker\varphi = \{x \in \mathbb{R} : \varphi(x) = 1\} = \left\{x \in \mathbb{R} : e^{2\pi i x} = 1\right\},$$

and the only $x \in \mathbb{R}$ for which this holds are the integers. That is, $\ker\varphi = \mathbb{Z}$, and the First Isomorphism Theorem tells us that

$$\mathbb{R}/\mathbb{Z} = \mathbb{T}.$$

Therefore, the study of periodic functions on $\mathbb{R}$ can be reduced to the study of functions on the circle $\mathbb{T}$.

2.11 Direct Products of Groups

Now that we've finished with quotient groups, we'll talk about one last construction that we can make with groups, called the **direct product**. Along with all the other things we've talked about (subgroups, normal subgroups, homomorphisms, quotient groups), direct products are crucial to discussing the structure of finite groups in general. In particular, we will use direct products in the next section to classify all finite abelian groups up to isomorphism.

The direct product gives us a way of sticking two groups together to obtain a new group. Suppose we have two groups G_1 and G_2. We can turn the cartesian product $G_1 \times G_2$ into a group by defining a binary operation as follows: given $(a_1, a_2), (b_1, b_2) \in G_1 \times G_2$, put

$$(a_1, a_2)(b_1, b_2) = (a_1 b_1, a_2 b_2).$$

That is, we simply multiply elements of $G_1 \times G_2$ componentwise. The claim is that this operation turns $G_1 \times G_2$ into a group. Let's check this.

- **Closure:** $G_1 \times G_2$ is closed under this operation, simply because G_1 and G_2 are closed under their respective operations.
- **Associativity:** Yes, the operation is easily seen to be associative since G_1 and G_2 have associative operations.
- **Identity:** If e_1 and e_2 denote the identity elements of G_1 and G_2, respectively, then it is easy to check that (e_1, e_2) is the identity in $G_1 \times G_2$.
- **Inverses:** If $(a_1, a_2) \in G_1 \times G_2$, then the inverse of this element is given by

$$(a_1, a_2)^{-1} = (a_1^{-1}, a_2^{-1}).$$

Therefore, $G_1 \times G_2$ is a group.

Definition 2.11.1. If G_1 and G_2 are groups, the group $G_1 \times G_2$ is called the **direct product** of G_1 and G_2.

Let's do a couple of short examples, which will indicate why direct products are useful in analyzing certain groups in terms of other, more well-understood groups.

Example 2.11.2. Let $G_1 = G_2 = \mathbb{R}$. Then $G_1 \times G_2 = \mathbb{R} \times \mathbb{R}$, under the operation

$$(a_1, a_2) + (b_1, b_2) = (a_1 + b_1, a_2 + b_2).$$

Thus the direct product is simply $\mathbb{R}^2$, viewed as an abelian group. Thus elements of direct products behave like vectors, in some sense.

Example 2.11.3. This example is meant to emphasize the fact that the operation in each component depends on the operation in that group. Let $G_1 = \mathbb{Z}_3$, and let $G_2 = S_3$. Then $G_1 \times G_2$ has order $3 \cdot 6 = 18$, and multiplication in this group looks as follows:

$$(1, (1\ 2))(1, (1\ 2\ 3)) = (1 +_3 1, (1\ 2)(1\ 2\ 3)) = (2, (2\ 3)),$$

for example. On the other hand,

$$(1, (1\ 2\ 3))(1, (1\ 2)) = (2, (1\ 3)),$$

which shows that $G_1 \times G_2$ is not abelian (since S_3 is not abelian).

Example 2.11.4. Let $G_1 = G_2 = \mathbb{Z}_2$. Then the direct product $\mathbb{Z}_2 \times \mathbb{Z}_2$ has four elements:

$$\mathbb{Z}_2 \times \mathbb{Z}_2 = \{(0,0),(0,1),(1,0),(1,1)\}.$$

We also claim that it is an abelian group. (This is not too hard to see, and you'll prove it in one of the exercises.) Clearly $(0,0)$ serves as the identity, while

$$\begin{aligned} (0,1)+(0,1) &= (0,0) \quad (0,1)+(1,0) = (1,1) \\ (1,0)+(1,0) &= (0,0) \quad (1,0)+(1,1) = (0,1) \\ (1,1)+(1,1) &= (0,0) \quad (1,1)+(0,1) = (1,0) \end{aligned}$$

If we label the elements as follows:

$$\begin{aligned} e &\longleftrightarrow (0,0) \\ a &\longleftrightarrow (1,0) \\ b &\longleftrightarrow (0,1) \\ c &\longleftrightarrow (1,1). \end{aligned}$$

we see that $\mathbb{Z}_2 \times \mathbb{Z}_2 \cong V_4$. Thus this direct product is simply another avatar of the Klein 4-group.

The following are some fundamental facts about direct products. Their proofs are left as exercises.

Proposition 2.11.5. *Let G_1 and G_2 be groups.*

(a) $G_1 \times G_2 \cong G_2 \times G_1$.

(b) *If G_1 and G_2 are abelian, then so is $G_1 \times G_2$.*

(c) $|G_1 \times G_2| = |G_1||G_2|$.

(d) *If $(a_1,a_2) \in G_1 \times G_2$, then $o(a_1,a_2) = lcm(o(a_1),o(a_2))$.*

(e) *If G_1 and G_2 are cyclic, and $|G_1|$ and $|G_2|$ are relatively prime, then $G_1 \times G_2$ is cyclic.*

It is worth noting that everything we have said so far regarding direct products works for more than just two groups. In other words, we can take a direct product of any (finite) number of groups by simply extending our previous definition in the natural way.

Definition 2.11.6. Let $G_1, G_2, \ldots, G_n$ be groups. The **direct product** of this family is the set $G_1 \times G_2 \times \cdots \times G_n$ endowed with the binary operation

$$(a_1, a_2, \ldots, a_n)(b_1, b_2, \ldots, b_n) = (a_1 b_1, a_2 b_2, \ldots, a_n b_n).$$

If $e_1, e_2, \ldots, e_n$ denote the identities of $G_1, G_1, \ldots, G_n$, then

$$(e_1, e_2, \ldots, e_n)$$

is the identity element of $G_1 \times G_2 \times \cdots \times G_n$. Finally, if $(a_1, a_2, \ldots, a_n) \in G_1 \times G_2 \times \cdots \times G_n$, then

$$(a_1, a_2, \ldots, a_n)^{-1} = (a_1^{-1}, a_2^{-1}, \ldots, a_n^{-1}).$$

Remark 2.11.7. Let $G_1, G_2, \ldots, G_n$ be groups, and define $\overline{G}_i \subseteq G_1 \times G_2 \times \cdots \times G_n$ by

$$\overline{G}_i = \big\{(e_1, e_2, \ldots, e_{i-1}, a_i, e_{i+1}, \ldots, e_n) : a_i \in G_i\big\}.$$

Then it is straightforward to check that $\overline{G}_i$ is a subgroup of $G_1 \times G_2 \times \cdots \times G_n$, and in fact it is a normal subgroup. Also, $\overline{G}_i$ is easily seen to be isomorphic to G_i in a natural way. Therefore, each "factor" of a direct product can be viewed as a normal subgroup of the direct product in a canonical way.

2.11.1 Internal Direct Products

At the beginning of this section, we expressed a desire to analyze the structure of finite groups by using direct products. Therefore, we'd really like to go in the direction opposite to what we've been doing so far—rather than build the direct product of two (or more) finite groups, we need to *decompose* a given group G and write it as a direct product of some of its subgroups. This task is actually quite difficult (or even impossible in many cases), so we will focus on a much simpler question.

Let G be a group, and suppose H and K are subgroups of G. We'd like to know whether it is possible to write $G \cong H \times K$ or not. Looking back at the properties of the direct product, what would we need to know at the very least? Well, we showed that the factors of a direct product are always *normal* subgroups that intersect trivially, so we would need to require $H, K \trianglelefteq G$ and $H \cap K = \{e\}$. We also need H and K to "fill up" all of G, in the sense that every element of G can be obtained by combining elements of H and K. More precisely, if we define

$$HK = \{hk : h \in H, k \in K\},$$

then we will require $G = HK$. Thus every element of G can be expressed in terms of elements of H and K, and the fact that $H \cap K = \{e\}$ will force this representation to be unique. What we are really getting at is the following theorem.

Theorem 2.11.8. *Let G be a group, and let $H, K \trianglelefteq G$. Suppose that*

1. *$HK = G$, and*
2. *$H \cap K = \{e\}$.*

Then $G \cong H \times K$.

The idea of the proof is as follows: there is an obvious map $\varphi : H \times K \to G$ given by

$$\varphi(h, k) = hk.$$

Our claim is that φ is an isomorphism. To see this, we first need to check that φ is a homomorphism. If $(h_1, k_1), (h_2, k_2) \in H \times K$, then

$$\varphi((h_1, k_1)(h_2, k_2)) = \varphi(h_1h_2, k_1k_2) = h_1h_2k_1k_2,$$

while

$$\varphi(h_1, k_1)\varphi(h_2, k_2) = (h_1k_1)(h_2k_2) = h_1h_1k_1k_2.$$

In order for φ to be a homomorphism, we'll need to know that elements of H and K commute with each other.

Lemma 2.11.9. *If $H, K \trianglelefteq G$ and $H \cap K = \{e\}$, then $hk = kh$ for all $h \in H$ and $k \in K$.*

Proof. Let $h \in H$ and $k \in K$. To show that $hk = kh$, it would suffice to show that

$$hkh^{-1}k^{-1} = e.$$

Let's look at this commutator. If we group things one way, we have

$$(hkh^{-1})k^{-1} \in K,$$

since $hkh^{-1} \in K$ (as K is normal) and $k^{-1} \in K$. On the other hand,

$$h(kh^{-1}k^{-1}) \in H,$$

again since H is normal. Thus $hkh^{-1}k^{-1} \in H \cap K = \{e\}$, so $hkh^{-1}k^{-1} = e$. Therefore, $hk = kh$. □

Now we can prove the theorem.

Proof of Theorem 2.11.8. Let φ be as above. Then the lemma shows that φ is a homomorphism, by our previous calculations. It is easy to see that φ is onto, since $G = HK$. In particular, if $g \in G$, then $g = hk$ for some $h \in H$, $k \in K$, and

$$\varphi(h, k) = hk = g.$$

To see that φ is one-to-one, observe that

$$\ker \varphi = \{(h, k) \in H \times K : hk = e\}.$$

Thus if $(h, k) \in \ker \varphi$, then $h = k^{-1}$. But then $h, k \in H \cap K = \{e\}$, so $h = k = e$. Therefore, $\ker \varphi = \{e\}$, so φ is one- to-one. It follows that φ is an isomorphism, and $G \cong H \times K$. □

The proof of Theorem 2.11.8 not only tells us that G is isomorphic to $H \times K$, but it shows us that the isomorphism only relies upon multiplication inside G. Therefore, we call G the **internal direct product** of H and K in this case.

Let's apply Theorem 2.11.8 to an example, which will foreshadow the upcoming results on finite abelian groups.

Example 2.11.10. A long time ago, we defined the Klein 4-group V_4. It was the group consisting of 4 elements e, a, b, and c, with the condition that

$$a^2 = b^2 = c^2 = e,$$

and

$$ab = c, ac = b, bc = a.$$

Since V_4 is abelian, any subgroup is normal. Take $H = \{e, a\}$ and $K = \{e, b\}$. Then $HK = V_4$ (since $ab = c$) and $H \cap K = \{e\}$, so

$$V_4 = H \times K$$

by the theorem. But $|H| = |K| = 2$, so $H, K \cong \mathbb{Z}_2$. Thus

$$V_4 \cong \mathbb{Z}_2 \times \mathbb{Z}_2.$$

It's not hard to see this explicitly either, via the correspondence

$$\begin{aligned} e &\longleftrightarrow (0, 0) \\ a &\longleftrightarrow (1, 0) \\ b &\longleftrightarrow (0, 1) \\ c &\longleftrightarrow (1, 1). \end{aligned}$$

We should mention finally that the theorem we have proven also holds if we have any finite number of normal subgroups. The proof is essentially the same, but slightly more involved, and we won't reproduce it here.

Theorem 2.11.11. *Let G be a group, and let $N_1, N_2, \ldots, N_m$ be normal subgroups of G. Suppose that*

1. *$N_1 N_2 \cdots N_m = G$, and*
2. *$N_1 \cap N_2 \cap \cdots \cap N_m = \{e\}$.*

Then $G \cong N_1 \times N_2 \times \cdots \times N_m$.

2.12 The Classification of Finite Abelian Groups

As we discussed at the end of the last section, we would like to use direct products to obtain classification results for finite groups. Shorly, we will obtain a fairly satisfying answer to some of the questions we've been asking, at least in the case of abelian groups. Without further ado, we'll state the theorem and discuss it.

Theorem 2.12.1 (Fundamental Theorem of Finite Abelian Groups)**.** *Let G be a finite abelian group. Then*

$$G \cong \mathbb{Z}_{p_1^{r_1}} \times \mathbb{Z}_{p_2^{r_2}} \times \cdots \times \mathbb{Z}_{p_k^{r_k}},$$

where $p_1, p_2, \ldots, p_k$ are (not necessarily distinct) primes. That is, G is isomorphic to a direct product of cyclic groups, where the order of each factor is a power of a prime.

This theorem looks sort of daunting, and it's still not really clear how to use it. It's even less clear how to prove it. We won't go through the proof, since it's extremely long and technical. However, we will talk about the theorem and how to apply it to specific examples. First of all, where do the primes in the theorem come from? We begin by factoring $|G|$ into primes:

$$|G| = p_1^{m_1} p_2^{m_2} \cdots p_j^{m_j}.$$

Then G breaks down as a direct product of groups of order $p_i^{m_i}$, which are not necessarily cyclic. But each one decomposes as a direct product of cyclic groups. (This is really the content of the proof.) In short, we have the following procedure for determining all the abelian groups of a given order.

1. Factor $n = |G|$ into primes: $n = p_1^{m_1} p_2^{m_2} \cdots p_k^{m_k}$.
2. Determine all possible groups of the form

$$\mathbb{Z}_{p_i^{r_1}} \times \mathbb{Z}_{p_i^{r_2}} \times \cdots \times \mathbb{Z}_{p_i^{r_j}}, \quad r_1 + r_2 + \cdots + r_j = m_i$$

for $1 \leq i \leq k$. (This amounts to finding all *partitions* of m_i.)

3. Determine all the possible ways of combining these groups together (via direct products) to get a group of order n.

Example 2.12.2. Determine all abelian groups of order 360, up to isomorphism.

Solution. First factor 360 into primes: $360 = 2^3 \cdot 3^2 \cdot 5$. Then determine all possible groups for each prime power:

- **Order 2^3:** $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$, $\mathbb{Z}_2 \times \mathbb{Z}_4$, $\mathbb{Z}_8$.
- **Order 3^2:** $\mathbb{Z}_3 \times \mathbb{Z}_3$, $\mathbb{Z}_9$.
- **Order 5:** $\mathbb{Z}_5$.

Therefore, the possible groups are:

$$\begin{aligned}
&\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_3 \times \mathbb{Z}_3 \times \mathbb{Z}_5 \\
&\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_9 \times \mathbb{Z}_5 \\
&\mathbb{Z}_2 \times \mathbb{Z}_4 \times \mathbb{Z}_3 \times \mathbb{Z}_3 \times \mathbb{Z}_5 \\
&\mathbb{Z}_2 \times \mathbb{Z}_4 \times \mathbb{Z}_9 \times \mathbb{Z}_5 \\
&\mathbb{Z}_8 \times \mathbb{Z}_3 \times \mathbb{Z}_3 \times \mathbb{Z}_5 \\
&\mathbb{Z}_8 \times \mathbb{Z}_9 \times \mathbb{Z}_5
\end{aligned}$$

□

Some readers might notice something odd here—there's one obvious abelian group of order 360 that doesn't appear to be on this list, namely $\mathbb{Z}_{360}$. It is here, but it's in disguise. You'll prove the following result in the exercises:

Proposition 2.12.3. *If* $\gcd(m, n) = 1$, *then* $\mathbb{Z}_m \times \mathbb{Z}_n \cong \mathbb{Z}_{mn}$.

Since 5, 8, and 9 are pairwise relatively prime, the last group on our list is really $\mathbb{Z}_{360}$ in disguise. We could even apply this result to the other groups to write them in a more compact way, but it isn't necessary to do so.

Example 2.12.4. Suppose we want to determine all abelian groups of order 63. We first write $63 = 3^2 \cdot 7$. We see that there are only two abelian groups of this order:

$$\mathbb{Z}_3 \times \mathbb{Z}_3 \times \mathbb{Z}_7 \cong \mathbb{Z}_3 \times \mathbb{Z}_{21}$$

and

$$\mathbb{Z}_9 \times \mathbb{Z}_7 \cong \mathbb{Z}_{63}.$$

Nonabelian Groups

Things are quite a bit harder when one tries to classify nonabelian groups. We'll stick to finite groups here, since there's really not much hope in general for infinite nonabelian groups.

In the nonabelian case, it is still possible sometimes to write a group G as a direct product of two of its subgroups, say

$$G \cong H \times K.$$

However, recall that we need H and K to both be normal. In the abelian case, there is no problem since every subgroup is normal. However, it is entirely possible that a nonabelian group G might not possess enough normal subgroups to be able to carry out this construction. The next best thing would be to find two subgroups H and K, where one of them is normal, such that $G = HK$ and $H \cap K = \{e\}$. Then we can express G as the **semidirect product** of H and K, written

$$G \cong H \rtimes K.$$

We won't go into how the semidirect product is constructed, though the idea is that it is like the direct product, but with a "twisted" binary operation. Thus it is still possible to break down a nonabelian group into a combination of two smaller groups, provided that we can find subgroups H and K as described above. Sometimes, however, even this isn't enough. There are cases where we can't even write a group as a semidirect product—the best we can do is find a normal subgroup N of G such that N and G/N are easy to understand. Then we say that G is an **extension** of N by G/N. Thus the classification of finite nonabelian groups depends on the understanding of all possible extensions of finite groups.

Once we have ways of breaking finite groups down into smaller pieces, it is feasible that we could reach a point where there are no smaller pieces. That is, we come to a stage where there are no longer any normal subgroups to consider; in other words, we arrive at a **simple group**. In short, we are saying that every finite group is built out of finite simple groups via direct products, semidirect products, and group extensions. Therefore, if we can understand all possible finite simple groups, then we will understand all finite groups. This is an ambitious goal for us, but not for mathematicians as a whole. In the 1970s, group theorists collectively managed to classify *all* finite simple groups. It took them about 10,000 pages of incredibly hard mathematics, but they did indeed accomplish it. They found that there are three families of finite simple groups:

- Finite cyclic groups of prime order, i.e., $\mathbb{Z}_p$.
- Alternating groups A_n for $n \geq 5$.
- Simple groups of Lie type. (Whatever those are...)

There are also 26 groups which do not fit into these categories, called the **sporadic groups**. The largest of the sporadic groups is called the **Monster group**, which is downright huge, with

$$808,017,424,794,512,875,886,459,904,961,710,757,005,754,368,000,000,000,$$

or approximately 8×10^{53} elements. Its discoverers, Fischer and Griess, wanted to call it the "friendly giant," which would conveniently have the abbreviation FG.

Additional Exercises for Chapter 2

2.8. In each case, determine whether $*$ defines a binary operation on the given set. If not, give reason(s) why $*$ fails to be a binary operation.

(a) $*$ defined on $\mathbb{Z}^+$ by $a * b = a - b$.

(b) $*$ defined on $\mathbb{Z}^+$ by $a * b = a^b$.

(c) $*$ defined on $\mathbb{Z}$ by $a * b = a/b$.

(d) $*$ defined on $\mathbb{R}$ by $a * b = c$, where c is at least 5 more than $a + b$.

2.9. Determine whether the binary operation $*$ is associative, and state whether it is commutative or not.

(a) $*$ defined on $\mathbb{Z}$ by $a * b = a - b$.

(b) $*$ defined on $\mathbb{Q}$ by $a * b = ab + 1$.

(c) $*$ defined on $\mathbb{Z}^+$ by $a * b = a^b$.

2.10. Let $S = \{a, b, c, d\}$ and define a binary operation $*$ on S via the following Cayley table:

$*$	a	b	c	d
a	a	c	b	d
b	c	a	d	b
c	b	d	a	c
d	d	b	c	a

Is this binary operation commutative? Is it associative?

2.11. Compute the Cayley table for $\langle \mathbb{Z}_6, +_6 \rangle$.

2.12. Suppose that $*$ is an associative and commutative binary operation on a set S. Show that the subset

$$H = \{a \in S : a * a = a\}$$

of S is closed under $*$. (The elements of H are called **idempotents** for $*$.)

2.13. Which of the following are groups? Why? (That is, either verify that the axioms hold, or explain why one of them fails.)

(a) $\mathbb{R}^+$ under addition. (Here $\mathbb{R}^+$ denotes the set of all *positive* real numbers.)

(b) The set $3\mathbb{Z}$ of integers that are multiples of 3, under addition.

(c) $\mathbb{R}-\{1\}$ under the operation $a*b = a+b-ab$.

(d) $\mathbb{Z}$ under the operation $a * b = a + b - 1$.

(e) The set of all rational numbers with denominator divisible by 5 (when written in lowest terms) under the operation $a * b = a + b$.

(f) Any set containing more than one element under the operation $a * b = a$.

2.14. The following table defines a binary operation on the set $S = \{a, b, c\}$.

$*$	a	b	c
a	a	b	c
b	b	b	c
c	c	c	c

Is $\langle S, * \rangle$ a group?

2.15. Let G be the set of all real-valued functions f on the real line which have the property that $f(x) \neq 0$ for all $x \in \mathbb{R}$. In other words,

$$G = \{f : \mathbb{R} \to \mathbb{R} : f(x) \neq 0 \text{ for all } x \in \mathbb{R}\}.$$

Define the product $f \times g$ of two functions $f, g \in G$ by

$$(f \times g)(x) = f(x)g(x) \text{ for all } x \in \mathbb{R}.$$

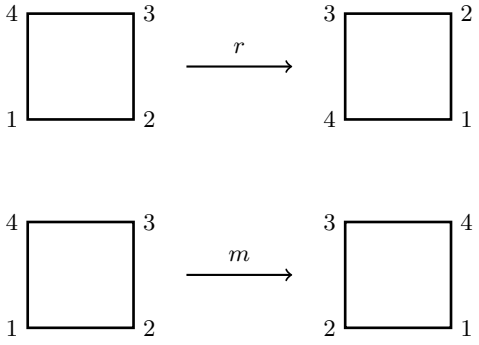
With this operation, does G form a group? Prove or disprove.

2.16. Let G be a group in which $a * a = e$ for all $a \in G$, show that G is abelian.

2.17. We saw earlier that any group of order 1, 2, or 3 is abelian. Show that any group of order 4 must be abelian. (**Hint:** Try to write down all the possible group tables in this case. Up to reordering the elements of the group, there will only be two possibilities.)

2.18.* Try to extend the previous problem by showing that any group of order 5 must also be abelian.

2.19. Let D_4 be the 4th dihedral group, which consists of symmetries of the square. Let $r \in D_4$ denote counterclockwise rotation by 90°, and let m denote reflection across the vertical axis.



Check that

$$rm = mr^{-1},$$

and conclude that D_4 is a nonabelian group of order 8.

2.20. We mentioned earlier that elements of D_n can be thought of as permutations of the vertices of the regular n-gon. For example, the rotation r of the square mentioned in the last problem can be identified with the permutation

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

Write the reflection m as a permutation $\mu \in S_4$, and compute the product $\rho\mu$ in S_4. Then compute $rm \in D_4$, and write it as a permutation σ. Check that $\sigma = \rho\mu$. (In other words, this identification of symmetries of the square with permutations respects the group operations.)

2.21. Recall that if $*$ is a binary operation on a set S, an element x of S is an **idempotent** if $x * x = x$. Prove that a group has exactly one idempotent element.

2.22. Consider the group $\langle \mathbb{Z}_{30}, +_{30} \rangle$ under addition.

(a) Find the orders of the elements 3, 4, 6, 7, and 18 in $\mathbb{Z}_{30}$.

(b) Find all the generators of $\langle \mathbb{Z}_{30}, +_{30} \rangle$.

2.23. Determine whether each of the following subsets is a subgroup of the given group. If not, state which of the subgroup axioms fails.

(a) The set of real numbers $\mathbb{R}$, viewed as a subset of the complex numbers $\mathbb{C}$ (under addition).

(b) The set $\pi\mathbb{Q}$ of rational multiples of π, as a subset of $\mathbb{R}$.

(c) The set of $n \times n$ matrices with determinant 2, as a subset of $\mathrm{GL_n}(\mathbb{R})$.

(d) The set $\{i, m_1, m_2, m_3\} \subseteq D_3$ of reflections of the equilateral triangle, along with the identity transformation.

2.24. Show that if G is a finite group and $|G|$ is even, then there is an element $a \in G$ such that $a \neq e$ and $a^2 = e$.

2.25. Let a and b be elements of a group G. Show that if ab has finite order n, then ba also has order n.

2.26. Let G be a group and let $a \in G$. An element $b \in G$ is called a *conjugate* of a if there exists an element $x \in G$ such that $b = xax^{-1}$. Show that any conjugate of a has the same order as a.

2.27. Let G be a group. If H and K are subgroups of G, show that $H \cap K$ is also a subgroup of G.

2.28. Let r and s be positive integers, and define

$$H = \{nr + ms : n, m \in \mathbb{Z}\}.$$

(a) Show that H is a subgroup of $\mathbb{Z}$.

(b) We saw earlier that every subgroup of $\mathbb{Z}$ is cyclic. Therefore, $H = \langle d \rangle$ for some $d \in \mathbb{Z}$.

What is this integer d? Prove that the d you've found is in fact a generator for H.

2.29. We proved earlier that every subgroup of a cyclic group is cyclic. The following statement is almost the converse of this:

"Let G be a group. If every *proper* subgroup of G is cyclic, then G is cyclic."

Find a counterexample to the above statement.

2.30. Prove that any subgroup of an abelian group is abelian.

2.31. Let X be a set, and recall that S_X is the group consisting of the bijections from S to itself, with the binary operation given by composition of functions. (If X is finite, then S_X is just the symmetric group on n letters, where X has n elements.) Given $x_1 \in X$, define

$$H = \{f \in S_X : f(x_1) = x_1\}.$$

Show that $H \leq S_X$.

2.32. Let G be a group. Define

$$Z(G) = \{a \in G : ax = xa \text{ for all } x \in G\}.$$

In other words, the elements of $Z(G)$ are exactly those that commute with *every* element of G. Prove that $Z(G)$ is a subgroup of G, called the **center** of G.

2.33. Show that if H and K are subgroups of an *abelian* group G, then

$$\{hk : h \in H \text{ and } k \in K\}$$

is a subgroup of G.

2.34. We know that if p is a prime number, then the cyclic group $\mathbb{Z}_p$ has no proper subgroups as a consequence of Lagrange's theorem. This problem will have you investigate a "converse" to this result.

(a) If G is a finite group which has no proper subgroups (other than $\{e\}$), prove that G must be cyclic.

(b) Extend the result of (a) by showing that if G has no proper subgroups, then G is not only cyclic, but

$$|G| = p$$

for some prime number p.

2.35.* Let G be a group, and let H be a subgroup of G.

(a) Let a be some fixed element of G, and define

$$aHa^{-1} = \{aha^{-1} : h \in H\}.$$

This set is called the **conjugate** of H by a. Prove that aHa^{-1} is a subgroup of G.

(b) Define the **normalizer** of H in G to be

$$N(H) = \{a \in G : aHa^{-1} = H\}.$$

Prove that $N(H)$ is a subgroup of G.

2.36. Determine if each mapping is a homomorphism. State why or why not. If it is a homomorphism, find its kernel, and determine whether it is one-to-one and/or onto.

(a) Define $\varphi : \mathbb{Z} \to \mathbb{R}$ by $\varphi(n) = n$. (Both are groups under addition here.)

(b) Let G be a group, and define $\varphi : G \to G$ by $\varphi(a) = a^{-1}$ for all $a \in G$.

(c) Let G be an *abelian* group, and define $\varphi : G \to G$ by $\varphi(a) = a^{-1}$ for all $a \in G$.

(d) Let G be a group, and define $\varphi : G \to G$ by $\varphi(a) = a^2$ for all $a \in G$.

2.37. Consider the subgroup $H = \{i, m_1\}$ of the dihedral group D_3. Find all the left cosets of H, and then find all of the right cosets of H. Observe that the left and right cosets do not coincide.

2.38. Find the cycle decomposition and order of each of the following permutations.

(a) $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 3 & 1 & 4 & 2 & 7 & 6 & 9 & 8 & 5 \end{pmatrix}$

(b) $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 7 & 6 & 5 & 4 & 3 & 2 & 1 \end{pmatrix}$

2.39. Determine whether each permutation is even or odd.

(a) $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 2 & 4 & 5 & 1 & 3 & 7 & 8 & 9 & 6 \end{pmatrix}$

(b) $(1\ 2\ 3\ 4\ 5\ 6)(7\ 8\ 9)$

(c) $(1\ 2\ 3\ 4\ 5\ 6)(1\ 2\ 3\ 4\ 5\ 7)$

(d) $(1\ 2)(1\ 2\ 3)(4\ 5)(5\ 6\ 8)(1\ 7\ 9)$

2.40. Let G and G' be groups, and suppose that $|G| = p$ for some prime number p. Show that any group homomorphism $\varphi : G \to G'$ must either be the trivial homomorphism or a one-to-one homomorphism.

2.41. Let $\varphi : G \to G'$ be a group homomorphism. If G is abelian and φ is onto, prove that G' is abelian.

2.42. Let G be an abelian group, n a positive integer, and define $\varphi : G \to G$ by $\varphi(x) = x^n$.

(a) Show that φ is a homomorphism.

(b) Suppose that G is a finite group and that n is relatively prime to $|G|$. Show that φ is an automorphism of G.

2.43. Let $V_4 = \{e, a, b, c\}$ denote the Klein 4-group. Since $|V_4| = 4$, Cayley's theorem tells us that V_4 is isomorphic to a subgroup of S_4. In this problem you will apply techniques from the proof of the theorem to this specific example in order to determine which subgroup of S_4 is matched up to V_4.

Suppose we label the elements of the Klein 4-group using the numbers 1 through 4, in the following manner:

e	a	b	c
1	2	3	4

Now multiply every element by a in order, i.e.,

e	a	b	c	$\longrightarrow$	a	e	c	b
1	2	3	4		2	1	4	3

Then multiplication by a determines a permutation of V_4 (by the proof of Cayley's theorem). This corresponds to an element of S_4 via the labels that we have given the elements of V_4. Do this for every element x of V_4. That is, write down the permutation in S_4 (in cycle notation) that is obtained by multiplying every element of V_4 by x.

2.44.* Let G be a group, and let $\mathrm{Aut}(G)$ denote the set of all automorphisms of G. We can define a binary operation on $\mathrm{Aut}(G)$ by:

$$\theta\psi = \theta \circ \psi$$

for $\theta, \psi \in \mathrm{Aut}(G)$.

(a) Prove that if $\theta, \psi \in \mathrm{Aut}(G)$, then $\theta\psi \in \mathrm{Aut}(G)$. (That is, show that we have indeed defined a binary operation by checking that $\mathrm{Aut}(G)$ is closed.

(b) If $\theta \in \mathrm{Aut}(G)$, then θ is in particular a bijection, so it has an inverse θ^{-1}. Prove that θ^{-1} is a homomorphism, so that $\theta^{-1} \in \mathrm{Aut}(G)$ for all $\theta \in \mathrm{Aut}(G)$.

(c) Use parts (a) and (b) to show that $\mathrm{Aut}(G)$ is itself a group under composition.

2.45.* Let G be a group with identity element e, and let X be a set. A **(left) action of G on X** is a function $G \times X \to X$, usually denoted by

$$(g, x) \mapsto g \cdot x$$

for $g \in G$ and $x \in X$, satisfying:

1. $g_1 \cdot (g_2 \cdot x) = (g_1 g_2) \cdot x$ for all $g_1, g_2 \in G$ and all $x \in X$.
2. $e \cdot x = x$ for all $x \in X$.

Intuitively, a group action assigns a permutation of X to each group element. (You will explore this idea in part (d) below.)

Finally, there are two important objects that are affiliated to any group action. For any $x \in X$, the **orbit of x under G** is the subset

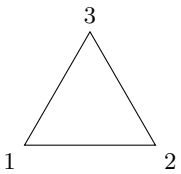
$$\mathrm{orb}(x) = \{g \cdot x : g \in G\}$$

of X, and the **stabilizer of x** is the subset

$$G_x = \{g \in G : g \cdot x = x\}$$

of G.

(a) (Warm up.) We have already seen that it is possible to view the elements of the dihedral group D_3 as permutations of the vertices of a triangle, labeled as below:



Thus D_3 acts on the set $X = \{1, 2, 3\}$ of vertices by permuting them. Determine the orbit and stabilizer of each vertex under this action.

(b) (Another example.) Let G be a group, let $X = G$, and define a map $G \times X \to G$ by

$$(g, x) \mapsto g \cdot x = gx$$

for all $g \in G$ and $x \in X$, i.e., the product of g and x as elements of G. Verify that this defines a group action of G on itself. (This action is called **left translation**.) Given $x \in X = G$, what are $\text{orb}(x)$ and G_x?

(c) Given an action of a group G on a set X, define a relation $\sim$ on X by $x \sim y$ if and only if there exists $g \in G$ such that $y = g \cdot x$. Prove that $\sim$ is an equivalence relation on X. What are the equivalence classes?

(d) Prove that for every $x \in X$, the stabilizer G_x is a subgroup of G.

(e) Given a fixed $g \in G$, define a function $\sigma_g : X \to X$ by

$$\sigma_g(x) = g \cdot x.$$

Show that σ_g is bijective, so σ_g defines a permutation of X. (Compare this to the proof of Cayley's theorem.)

(f) Recall that S_X denotes the group of permutations of X under composition. Define a function $\varphi : G \to S_X$ by

$$\varphi(g) = \sigma_g$$

for all $g \in G$. Prove that φ is a homomorphism. (**Note:** The proof of Cayley's theorem is a special case of this phenomenon, with G acting on itself by left translation.)

Parts (d) and (e) above show that a group action gives an alternative way of viewing a group as a collection of symmetries (or permutations) of some object. Cayley's theorem provides a specific example, where a group is viewed as a collection of permutations of itself. Group actions provide one of the most interesting ways in which groups are used in practice.

2.46. We proved that the kernel of any homomorphism is a normal subgroup. Conversely, you will show that any normal subgroup is the kernel of some homomorphism. That is, let G be a group with N a normal subgroup of G, and define a function $\pi : G \to G/N$ by

$$\pi(g) = Ng$$

for all $g \in G$. Prove that π is a homomorphism, and that $\ker \pi = N$.

2.47. Recall that $\mathbb{R}^\times$ denotes the group of nonzero real numbers (under multiplication), and let $N = \{-1, 1\}$. Show that N is a normal subgroup of $\mathbb{R}^\times$, and that $\mathbb{R}^\times/N$ is isomorphic to the group of positive real numbers under multiplication. (**Hint:** Use the First Isomorphism Theorem.)

2.48. Let $\varphi : \mathbb{Z}_8 \to \mathbb{Z}_4$ be given by

$$\varphi(x) = [x]_4,$$

i.e., the remainder of x mod 4. Find $\ker \varphi$. To which familiar group is $\mathbb{Z}_8/\ker \varphi$ isomorphic?

2.49. If G is a group and $M \trianglelefteq G$, $N \trianglelefteq G$, prove that $M \cap N \trianglelefteq G$. (You proved in an earlier exercise that $M \cap N$ is a subgroup of G, so you only need to prove that it is normal.)

2.50. Let G be a group. Recall from an earlier exercise that the **center** of G is the set $Z(G)$ defined by

$$Z(G) = \{x \in G : xa = ax \text{ for all } a \in G\}.$$

You proved that $Z(G)$ is a subgroup of G.

(a) Prove that $Z(G) \trianglelefteq G$.

(b) If $G/Z(G)$ is cyclic, prove that G is abelian.

2.51. Let G be a group and $H \leq G$. If $[G : H] = 2$, prove that H is normal in G.

2.52.* Let G be a group, and recall that $\text{Aut}(G)$ is the set of all automorphisms of G. You proved on the last homework that $\text{Aut}(G)$ forms a group under composition.

(a) Given $g \in G$, define a function $\theta_g : G \to G$ by

$$\theta_g(a) = gag^{-1}$$

for all $a \in G$. Show that $\theta_g \in \text{Aut}(G)$. (Such an automorphism is called an **inner automorphism**.)

(b) Let $\mathrm{Inn}(G)$ denote the set of all inner automorphisms of G. Then $\mathrm{Inn}(G) \subset \mathrm{Aut}(G)$ by part (a). Show that $\mathrm{Inn}(G)$ is actually a subgroup of $\mathrm{Aut}(G)$.

(c) Prove that $\mathrm{Inn}(G)$ is a normal subgroup.

2.53. Classify all abelian groups of order 600 up to isomorphism.

2.54. Let G be a group, and let $H \leq G$.

(a) If G is abelian, prove that G/H is abelian. (**Hint:** You may want to use a result from an earlier exercise.)

(b) Prove that if G is cyclic, then G/H is also cyclic.

2.55. Prove that if G_1 and G_2 are abelian groups, then $G_1 \times G_2$ is abelian.

2.56. If G_1 and G_2 are groups, prove that $G_1 \times G_2 \cong G_2 \times G_1$.

2.57. Let $G = \mathbb{Z}_4 \times \mathbb{Z}_6$, and let $H = \langle (1,0) \rangle$. Find all the right cosets of H in G, and compute the quotient group G/H. (That is, identify it with a more familiar group.)

2.58.* Let D_n denote the dihedral group of order $2n$, let $r \in D_n$ denote the counterclockwise rotation by $2\pi/n$ radians, and let m denote any reflection of the regular n-gon. Recall that the rotation subgroup

$$H = \{e, r, r^2, \ldots, r^{n-1}\}$$

is a normal subgroup of D_n. Let $K = \{e, m\}$; we saw in class that K is a subgroup, but it is not normal. You will need two facts regarding D_n (which you do not need to prove):

1. Every element of D_n can be written as $r^i m^j$, with $0 \leq i \leq n-1$ and $j = 0$ or 1. Thus $D_n = HK$.
2. You proved earlier (in a special case) that $mr = r^{-1}m$.

Define a group G as follows: the elements of G are pairs (r^i, m^j) (so that $G = H \times K$ *as sets*), but the binary operation on G is "twisted" in some sense. More specifically, we define

$$\begin{aligned}(r^i, e)(r^j, e) &= (r^{i+j}, e)\\ (r^i, m)(r^j, e) &= (r^{i-j}, m)\\ (r^i, e)(r^j, m) &= (r^{i+j}, m)\\ (r^i, m)(r^j, m) &= (r^{i-j}, e)\end{aligned}$$

for all i, j between 1 and $n-1$. Now define $\varphi : G \to D_n$ by

$$\varphi(r^i, m^j) = r^i m^j.$$

Prove that φ is an isomorphism of G onto D_n. (This shows that D_n is not quite a direct product of the subgroups H and K, since K isn't normal. However, things can be made to work if we modify the multiplication on $H \times K$ slightly.)

Chapter 3

Ring Theory

We're now done with our study of groups (at least for the sake of groups). The plan now is to move on to more complicated algebraic structures, namely objects called **rings**.

3.1 Rings

You'll probably agree that groups were kind of strange at first. We had a set with a single binary operation that satisfied certain nice properties, and we were able to come up with many examples of such things. Some of these examples were ones we already knew, such as $\mathbb{Z}$, $\mathbb{R}$, and $\mathbb{Q}$ under addition. However, these examples all really come equipped with *two* binary operations, namely *addition* and *multiplication*.

Example 3.1.1. We already know that $\mathbb{Z}$ forms an abelian group under addition. What nice properties are satisfied by multiplication on $\mathbb{Z}$?

- associativity
- commutativity
- identity
- distributivity (if $a, b \in \mathbb{Z}$ and $n \in \mathbb{Z}$, then $n(a+b) = na + nb$)

These four properties really tell us that $\mathbb{Z}$ is a model example of a ring.

Definition 3.1.2. A **ring** is a set R equipped with two binary operations, denoted by $+$ and $\cdot$, satisfying the following conditions.

1. $\langle R, + \rangle$ is an abelian group, meaning that
   (a) $+$ is associative and commutative;

(b) there is an additive identity $0 \in R$ such that $0 + a = a$ for all $a \in R$; and

(c) for each $a \in R$, there is an additive inverse $-a \in R$ so that $a + (-a) = 0$.

2. The multiplication operation $\cdot$ is associative.

3. (Distributive law) For all $a, b, r \in R$, we have

$$r \cdot (a + b) = r \cdot a + r \cdot b$$

and

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

You'll probably notice that we left two things off this list. We did not require that multiplication be commutative, or that there is even a multiplicative identity. Rings that have these properties are special, and thus have special names.

Definition 3.1.3. A **commutative ring** is a ring R for which

$$a \cdot b = b \cdot a$$

for all $a, b \in R$.

Definition 3.1.4. A **ring with identity**[1] (also called a **ring with unity** or a **unital ring**) is a ring R which contains an element $1 \in R$ (with $1 \neq 0$) satisfying

$$1 \cdot a = a \cdot 1 = a$$

for all $a \in R$.

Before we proceed, let's look at some familiar (and less familiar) examples of rings.

Example 3.1.5.

1. We already saw that $\mathbb{Z}$ is a ring, and actually a commutative ring with identity.

2. Similarly, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are all commutative rings with identity with respect to their usual operations.

3. For any n, $\mathbb{Z}_n$ is a commutative ring with identity with respect to modular addition and multiplication.

4. Let's try a noncommutative example. Recall that $M_n(\mathbb{R})$ is the set of all $n \times n$ matrices with real coefficients. Then $M_n(\mathbb{R})$ is a ring with respect to matrix addition and multiplication. It is noncommutative, but it has an identity, namely the identity matrix I.

5. Let $2\mathbb{Z}$ be the set of all even integers. Then $2\mathbb{Z}$ is a commutative ring with respect to the usual arithmetic. It does not have an identity, however, since $1 \notin 2\mathbb{Z}$.

6. Let $C([0,1],\mathbb{R}) = \{f : [0,1] \to \mathbb{R} : f \text{ is continuous}\}$. Then $C([0,1],\mathbb{R})$ is a ring with respect to pointwise addition and multiplication:

$$f + g(x) = f(x) + g(x)$$

and

$$fg(x) = f(x)g(x).$$

It is also commutative, and its multiplicative identity is the function that is identically 1.

7. This will be one of the main examples that we'll study. Let $\mathbb{Q}[x]$ be the set of all polynomials with rational coefficients. Then $\mathbb{Q}[x]$ forms a ring: we add two polynomials by adding their coefficients, say

$$\left[2x^2 + 3x + \frac{1}{2}\right] + \left[5x^3 + \frac{7}{2}x + 2\right] = 5x^3 + 2x^2 + \frac{13}{2}x + \frac{3}{2}.$$

We multiply them by "foiling":

$$(x^3 + 1)(3x^2 + 4x + 2) = 3x^5 + 4x^4 + 2x^3 + 3x^2 + 4x + 2.$$

This turns $\mathbb{Q}[x]$ into a commutative ring with identity. (We could also do this with any commutative ring in place of $\mathbb{Q}$.)

3.2 Basic Facts and Properties of Rings

We'll soon start investigating particular types of rings. Before we can do this, we should prove some relatively simple facts about rings which will be needed in computations.

Proposition 3.2.1. *Let R be a ring. Then for all $a, b \in R$, we have:*

(a) $0 \cdot a = a \cdot 0 = 0$;

(b) $(-a)b = a(-b) = -(ab)$;

(c) $(-a)(-b) = ab$; *and*

(d) *if* $1 \in R$, *then* $(-1)a = a(-1) = -a$.

Proof. (a) If $a \in R$, then we have

$$0 \cdot a = (0 + 0) \cdot a = 0 \cdot a + 0 \cdot a$$

by the right distributive law. But this means that $0 \cdot a$ is an additive idempotent in the abelian group $\langle R, + \rangle$. There is only one such element, so $0 \cdot a = 0$. A similar argument works to show that $a \cdot 0 = 0$.

(b) Let $a, b \in R$. Then by distributivity,

$$ab + (-a)b = (a + (-a))b = 0 \cdot b = 0,$$

so $(-a)b = -(ab)$. The same sort of argument works to show that $a(-b) = -(ab)$.

(c) If we apply part (b), we have

$$(-a)(-b) = -((-a)b) = -(-(ab)) = ab.$$

(d) Again by part (b), $(-1)a = -(1 \cdot a) = -a$, and $a(-1) = -(a \cdot 1) = -a$. □

As opposed to groups, there are many different types of rings that one can consider depending on how the multiplication behaves. In particular, some strange things can happen regarding the multiplication in a ring.

Example 3.2.2. Let $R = M_2(\mathbb{R})$, and let $A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ and $B = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$. Then

$$AB = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

but neither A nor B is the zero matrix. In other words, it is possible to multiply two nonzero elements of a ring and obtain 0. Such anomalies have a special name in ring theory.

Definition 3.2.3. An element $a \neq 0$ of a (commutative) ring R is called a **zero divisor** if there exists $b \in R$ such that $ab = 0$.

Many of the results that we will obtain will involve commutative rings which are free of zero divisors.

Definition 3.2.4. A commutative ring R is called an **integral domain** if R contains no zero divisors. Equivalently, if $a, b \in R$ with $ab = 0$, then either $a = 0$ or $b = 0$.

Example 3.2.5. 1. The rings $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are all integral domains.

2. When n is composite, $\mathbb{Z}_n$ is not an integral domain. For example, in $\mathbb{Z}_6$,
$$2 \cdot 3 = 0.$$

3. The function ring $C([0,1], \mathbb{R})$ is not an integral domain. Let
$$f(x) = \begin{cases} 0 & x < 1/2 \\ x - 1/2 & x \geq 1/2 \end{cases}$$
and
$$g(x) = \begin{cases} x - 1/2 & x < 1/2 \\ 0 & x \geq 1/2. \end{cases}$$
Then $f(x)g(x) = 0$ for all $x \in \mathbb{R}$, but neither function is identically zero.

Another thing you may have noticed is that elements of rings (with identity) do not necessarily possess multiplicative inverses. The ones that do have a special name.

Definition 3.2.6. Let R be a ring with identity. An element $a \in R$ is called a **unit** if there is a $b \in R$ such that
$$ab = ba = 1.$$
(We usually write $b = a^{-1}$.) The set of all units is denoted by $R^\times$, called the **unit group**[2] of R.

Example 3.2.7. Let's compute the unit groups in some rings.

1. What are the units in $\mathbb{Z}$? The only integers that have multiplicative inverses are 1 and -1, so
$$\mathbb{Z}^\times = \{1, -1\}.$$

2. What is the unit group in $\mathbb{Q}$? Every nonzero rational number has a multiplicative inverse, so
$$\mathbb{Q}^\times = \mathbb{Q} - \{0\}.$$
Conveniently enough, this is the notation that we have already used for the group of nonzero rational numbers under multiplication. Similarly, we have $\mathbb{R}^\times = \mathbb{R} - \{0\}$ and $\mathbb{C}^\times = \mathbb{C} - \{0\}$.

3. Let's consider the ring $\mathbb{Z}_n$ for $n \in \mathbb{Z}$. We proved a long time ago that the elements of $\mathbb{Z}_n$ that have multiplicative inverses are precisely those which are relatively prime to n. Therefore,
$$\mathbb{Z}_n^\times = \{a \in \mathbb{Z}_n : \gcd(a, n) = 1\}.$$
Note that if p is prime, then $\mathbb{Z}_p^\times = \{1, 2, \ldots, p-1\} = \mathbb{Z}_p - \{0\}$.

Of course 0 never has a multiplicative inverse, but we've seen that it's possible that all other elements do. Such rings are special.

Definition 3.2.8. A ring R with identity is called a **division ring** if every nonzero $a \in R$ is a unit. Equivalently, $R^\times = R - \{0\}$.

Definition 3.2.9. A commutative division ring is called a **field**.

Example 3.2.10.
1. The integral domains $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are all fields.
2. Since only 1 and -1 are units, $\mathbb{Z}$ is not a field.
3. If p is prime, $\mathbb{Z}_p$ is a field. It is an example of a **finite field**.

Exercise 3.1. Let R be a commutative ring with identity. Show that if $u \in R$ is a unit, then u cannot be a zero divisor. As a consequence, any field is an integral domain.

3.2.1 The Quaternions

You may have noticed that we produced several examples of fields (and hence of division rings), but no division rings that weren't fields. There is an example of such a thing, and it's an interesting one to both mathematicians and physicists.

Example 3.2.11. The quaternions were invented by the Irish mathematician and physicist William Rowan Hamilton in the middle of the 19th century. Motivated by the complex numbers and their geometric interpretations, he had tried to define multiplication on triples of real numbers. (That is, he tried to turn $\mathbb{R}^3$ into a ring.) This didn't quite work (and, in fact, is impossible), but he had a breakthrough one day while walking around campus. It was so exciting to him that he actually carved his initial identity into a stone bridge.

Hamilton's idea was to consider three elements i, j, and k, each of which behave like the imaginary unit in $\mathbb{C}$:

$$i^2 = j^2 = k^2 = ijk = -1.$$

From this single identity, it is possible to deduce that

$$ij = k, jk = i, ki = j,$$

and that these elements *anticommute*:

$$ij = -ji, kj = -kj, ik = -ki.$$

The **quaternions** are then defined to be the set

$$\mathbb{H} = \{a_1 + a_2 i + a_3 j + a_4 k : a_1, a_2, a_3, a_4 \in \mathbb{R}\}.$$

In other words, one can think of $\mathbb{H}$ as a four-dimensional real vector space with basis $\{1, i, j, k\}$. The addition operation works just like addition of vectors in $\mathbb{R}^4$: for example,

$$(1 + 4i + 7j + 2k) + (-2 + 6i - 9j - 4k) = -1 + 10i - 2j - 2k.$$

In general, we would have

$$(a_1 + a_2 i + a_3 j + a_4 k) + (b_1 + b_2 i + b_3 j + b_4 k) = (a_1 + b_1) + (a_2 + b_2)i + (a_3 + b_3)j + (a_4 + b_4)k.$$

Multiplication works like multiplication of complex numbers, but more complicated. We simply multiply everything out, and then use the quaternion identities listed above to simplify the result: for example,

$$\begin{aligned}(3 - 5k)(1 + 4i - 5j) &= 3 + 12i - 15j - 5k - 20ki + 25kj \\ &= 3 + 12i - 15j - 5k + 20j - 25i \\ &= 3 - 13i + 5j - 5k\end{aligned}$$

There is a general formula for the coefficients of the product of two quaternions, but it's usually easier to simply multiply on a case-by-case basis:

$$(a_1 + a_2 i + a_3 j + a_4 k)(b_1 + b_2 i + b_3 j + b_4 k) = c_1 + c_2 i + c_3 j + c_4 k,$$

where

$$\begin{aligned}c_1 &= a_1 b_1 - a_2 b_2 - a_3 b_3 - a_4 b_4 \\ c_2 &= a_1 b_2 + a_2 b_1 + a_3 b_4 - a_4 b_3 \\ c_3 &= a_1 b_3 - a_2 b_4 + a_3 b_1 + a_4 b_2 \\ c_4 &= a_1 b_4 + a_2 b_3 - a_3 b_2 + a_4 b_1.\end{aligned}$$

It's possible to check that multiplication is associative and distributive, though we won't do here. Also, note that for example we have

$$\begin{aligned}(1 + 3i - 4j - 10k)(1 - 3i + 4j + 10k) &= 1 - 3i + 4j + 10k + 3i - 9i^2 + 12ij + 30ik \\ &\quad - 4j + 12ji - 16j^2 - 40jk - 10k + 30ki - 40kj - 100k^2 \\ &= 1 + 9 + 12k - 30j - 12k + 16 - 40i + 30j + 40i + 100 \\ &= 1 + 9 + 16 + 100 \\ &= 126.\end{aligned}$$

In general, one can verify that

$$(a_1 + a_2i + a_3j + a_4k)(a_1 - a_2i - a_3j - a_4k) = a_1^2 + a_2^2 + a_3^2 + a_4^2.$$

This fact allows us to define the inverse of any nonzero quaternion: if $a_1 + a_2i + a_3j + a_4k \neq 0$, then the quantity

$$\alpha = a_1^2 + a_2^2 + a_3^2 + a_4^2$$

is nonzero, and it can be checked that

$$(a_1 + a_2i + a_3j + a_4k)\left(\frac{a_1}{\alpha} - \frac{a_2}{\alpha}i - \frac{a_3}{\alpha}j - \frac{a_4}{\alpha}k\right) = 1.$$

Thus $\mathbb{H}^\times = \mathbb{H} - \{0\}$, and $\mathbb{H}$ is a division ring. It is not a field, however, since we have $ij \neq ji$, for example.

Remark 3.2.12. Hamilton's quaternions may seem somewhat contrived, but they are more than just a method for multiplying vectors in $\mathbb{R}^4$ together. They are actually important in physics, since they can be used to model rotations in $\mathbb{R}^3$.

3.3 Ring Homomorphisms and Ideals

Since rings can be thought of as abelian groups with an additional binary operation, it might be expected that there are analogues to various facts and results from group theory. In particular, one might guess that the ideas of subgroup, homomorphism, normal subgroup, and quotient group can be ported over to rings. Of course they can, and many results will come simply because we have already done much of the work in our study of abelian groups.

The first object to mention is the analogue of a subgroup, which is quite naturally called a **subring**.

Definition 3.3.1. Let R be a ring. A **subring** of R is a subset $S \subseteq R$ that is a ring with respect to the operations inherited from R.

Subrings will not be tremendously important, and we'll be more interested in their "normal" brethren. However, we'll still give a couple of examples of importance.

Example 3.3.2.

1. We saw earlier that $\mathbb{Z}$ and $2\mathbb{Z}$ are both rings, and since $2\mathbb{Z} \subseteq \mathbb{Z}$, $2\mathbb{Z}$ is a subring of $\mathbb{Z}$.
2. Similarly, $\mathbb{Z}$ is a subring of $\mathbb{Q}$.
3. Any ring R automatically contains two subrings, namely $\{0\}$ and R itself.

The next idea that we want to generalize is that of a homomorphism. We know that a group homomorphism is a function between groups that preserves the binary operation. What should we then expect a ring homomorphism to satisfy?

Definition 3.3.3. Let R and R' be rings. A **homomorphism** is a function $\varphi : R \to R'$ with the property that

$$\varphi(a+b) = \varphi(a) + \varphi(b)$$

and

$$\varphi(ab) = \varphi(a)\varphi(b)$$

for all $a, b \in R$.

Example 3.3.4. 1. Let $\varphi : \mathbb{Z} \to \mathbb{Z}_n$ be given by reduction mod n:

$$\varphi(a) = [a]_n = a \bmod .$$

We have already seen that this is a homomorphism of abelian groups, but it is also not hard to see that it preserves multiplication, and is thus a ring homomorphism.

2. Define $\mathrm{ev}_0 : \mathbb{Q}[x] \to \mathbb{Q}$ by $\mathrm{ev}_0(f) = f(0)$. This map evaluates a polynomial at 0, and is thus called an **evaluation homomorphism**. It is a ring homomorphism, since

$$\mathrm{ev}_0(f+g) = (f+g)(0) = f(0) + g(0) = \mathrm{ev}_0(f) + \mathrm{ev}_0(g)$$

and

$$\mathrm{ev}_0(fg) = (fg)(0) = f(0)g(0) = \mathrm{ev}_0(f)\mathrm{ev}_0(g).$$

Note that a ring homomorphism is also a group homomorphism with respect to the underlying abelian groups $\langle R, + \rangle$ and $\langle R', + \rangle$. Then we should of course be able to say a few things about ring homomorphisms using our knowledge of groups. First, recall that the image of any group homomorphism is a subgroup of the codomain. What more can we say if we are dealing with rings?

Proposition 3.3.5. *If $\varphi : R \to R'$ is a ring homomorphism, then $\varphi(R)$ is a subring of R'.*

Also, we can still talk about the kernel of a homomorphism in the ring case, and things work much the same way as they do with groups.

Definition 3.3.6. Let $\varphi : R \to R'$ be a ring homomorphism. The **kernel** of φ is the set

$$\ker \varphi = \{x \in R : \varphi(x) = 0\}.$$

Here we have defined the kernel in terms of the additive identity because we want to extend the idea of the kernel from group theory. That is, we want to consider φ first as a homomorphism of abelian groups, and then show that we can say more about the kernel by viewing φ as a ring homomorphism. With this in mind, we know that R forms an abelian group under addition, so of course the kernel is an additive subgroup of R. Can we say more? If $a, b \in \ker \varphi$, then

$$\varphi(ab) = \varphi(a)\varphi(b) = 0 \cdot 0 = 0,$$

so $\ker \varphi$ is a subring of R. In fact, it's even stronger than this: if $a \in \ker \varphi$ and $r \in R$, then

$$\varphi(ra) = \varphi(r)\varphi(a) = \varphi(r) \cdot 0 = 0$$

and

$$\varphi(ar) = \varphi(a)\varphi(r) = 0 \cdot \varphi(r) = 0,$$

so $ra, ar \in \ker \varphi$. This says that $\ker \varphi$ "absorbs" elements of R under multiplication. Therefore, it is actually much more than a subring.

Definition 3.3.7. Let R be a ring. An **ideal** of R is a subring $I \subseteq R$ with the property that $ra, ar \in I$ for all $a \in I$ and $r \in R$. We write $I \trianglelefteq R$ to denote that I is an ideal of R.

Of course we have already verified the following fact in the discussion above.

Theorem 3.3.8. *Let $\varphi : R \to R'$ be a ring homomorphism. Then $\ker \varphi$ is an ideal of R.*

Below are some specific examples of ideals.

Example 3.3.9.

1. Every ring R always contains two trivial ideals: $\{0\}$ and R itself.
2. For any $n \in \mathbb{Z}$, $n\mathbb{Z}$ is an ideal of $\mathbb{Z}$. We already know that $n\mathbb{Z}$ is a subgroup of the additive group $\mathbb{Z}$, and if $r \in \mathbb{Z}$ and $na \in n\mathbb{Z}$, then

   $$r(na) = n(ra) \in n\mathbb{Z},$$

   so $n\mathbb{Z}$ is an ideal. We should also note that $n\mathbb{Z}$ is the kernel of the homomorphism given by reduction mod n.

3. We saw that $\mathbb{Z}$ is a subring of $\mathbb{Q}$, but it is not an ideal.

4. We saw that $\text{ev}_0 : \mathbb{Q}[x] \to \mathbb{Q}$ is a homomorphism. What is its kernel? In order to have
$$\text{ev}_0(f) = f(0) = 0,$$
we must have that f has no constant term, i.e., that $f(x) = x \cdot g(x)$ for some other polynomial $g \in \mathbb{Q}[x]$. Therefore,
$$\ker \text{ev}_0 = \{x \cdot g : g \in \mathbb{Q}[x]\}.$$

5. Let $C([0,1], \mathbb{R})$ be the ring of real-valued continuous functions on $[0,1]$, and define
$$I = \{f \in C([0,1], \mathbb{R}) : f(0) = 0\}.$$
Then $I \trianglelefteq C([0,1], \mathbb{R})$: it is clearly closed under addition, and if $f \in I$ and $g \in C(\mathbb{R})$, then
$$(gf)(0) = g(0)f(0) = g(0) \cdot 0 = 0,$$
so $gf = I$. Again, we could have also appealed to the fact that $I = \ker \text{ev}_0$, the homomorphism given by evaluation at 0.

6. If F is a field, we claim that F has no nontrivial ideals. Suppose that $I \trianglelefteq F$ and $I \neq \{0\}$. Then there is a nonzero element $a \in I$, and a must be a unit. Since I is an ideal, $1 = a^{-1}a \in I$. But if $1 \in I$, then $r = r \cdot 1 \in I$ for any $r \in R$. Therefore, $I = F$.

7. Just as any element of a group generates a cyclic subgroup, any element of a ring generates an ideal. That is, given a ring R and $a \in R$, we define
$$(a) = \{r_1 a s_1 + r_2 a s_2 + \cdots + r_n a s_n : r_i, s_i \in R\}.$$
Then $(a) \trianglelefteq R$. If we assume that R is commutative (which is the case we'll usually be looking at), then this ideal is pretty easy to describe:
$$(a) = \{ra : r \in R\}.$$
These ideals are usually given a special name.

Definition 3.3.10. An ideal of the form (a) for $a \in R$ is called the **principal ideal** generated by a.

3.4 Quotient Rings

So far we've talked about some constructs for rings that mimic those of groups. We have one left—we need to generalize the idea of a quotient group to rings. Just as we needed to consider normal subgroups when we built quotient groups, we need to consider ideals here. We will see why we need to do this along the way.

Let R be a ring and $I \trianglelefteq R$. Then I is an additive subgroup of the *abelian* group R (hence normal), and we can form the quotient group R/I. Recall that the elements of R/I are the cosets of I in R, which all have the form

$$I + a = \{x + a : x \in I\}.$$

The group operation on $I + a$ is the addition that is inherited naturally from R:

$$(I + a) + (I + b) = I + (a + b).$$

The obvious question to ask in this context is: can we turn R/I into a ring? The natural thing to do would be to define the product of two cosets to be

$$(I + a)(I + b) = I + ab.$$

However, is this operation well-defined? We saw that things were not all well with groups, and we needed to assume that the subgroup was normal to make the group operation well-defined. In this context, it will be the fact that I is an ideal that saves us. Suppose that $I + a' = I + a$ and $I + b' = I + b$. Then we need $I + a'b' = I + ab$, or $a'b' - ab \in I$. Well, $a' - a \in I$ and $b' - b \in I$, so $a' = x + a$ and $b' = y + b$ for some $x, y \in I$. Then

$$\begin{aligned}(I + a')(I + b') &= (I + (a + x))(I + (b + y)) \\ &= I + (a + x)(b + y) \\ &= I + (ab + ay + xb + xy).\end{aligned}$$

Since I is an ideal, $ay, xb, xy \in I$, so this coset is simply $I + ab$. Therefore, the desired definition of multiplication on R/I is well-defined, so we can make R/I into a ring. Of course one would need to check that all of the necessary properties carry over from R, but this is fairly routine.

Definition 3.4.1. The ring R/I is called the **quotient ring** of R by I.

There are two facts regarding quotient groups that should also hold for rings. The first is that every ideal is the kernel of some homomorphism.

Theorem 3.4.2. *Let R be a ring and I an ideal of R. There is a homomorphism $\psi : R \to R/I$ given by*

$$\psi(a) = a + I$$

such that $\ker \psi = I$.

The second is of course the ubiquitous *First Isomorphism Theorem.*

Theorem 3.4.3 (First Isomorphism Theorem)**.** *Let $\varphi : R \to R'$ be a surjective ring homomorphism. Then there is an isomorphism $\tilde{\varphi} : R/\ker\varphi \to R'$ given by*

$$\tilde{\varphi}(a + \ker\varphi) = \varphi(a).$$

Most of the work for both of these proofs has already been done in the context of groups. In particular, we have already verified that these two maps are well-defined homomorphisms of abelian groups, so we simply need to see that they respect multiplication. This is not hard to do; for example, to prove the First Isomorphism Theorem, we just need to see that

$$\begin{aligned}\tilde{\varphi}((a + \ker\varphi)(b + \ker\varphi)) &= \tilde{\varphi}(ab + \ker\varphi) \\ &= \varphi(ab) \\ &= \varphi(a)\varphi(b) \\ &= \tilde{\varphi}(a + \ker\varphi)\tilde{\varphi}(b + \ker\varphi).\end{aligned}$$

Example 3.4.4. Let $\mathrm{ev}_0 : \mathbb{Q}[x] \to \mathbb{Q}$ be given by evaluation at 0. This is a surjective homomorphism, since if $a \in \mathbb{Q}$, the constant polynomial $f \equiv a$ evaluates to a at 0. We have already seen that the kernel is given by $I = \{x \cdot g : g \in \mathbb{Q}[x]\}$. Therefore,

$$\mathbb{Q}[x]/\ker \mathrm{ev}_0 \cong \mathbb{Q},$$

the field of rational numbers.

There is one more fact regarding quotient rings which carries over from group theory. We saw that there was a *correspondence theorem*, which gave us a bijection between subgroups of a quotient group G/N and subgroups of G containing N. There is a similar result for quotient rings, albeit with ideals playing the role of normal subgroups.

Theorem 3.4.5 (Correspondence Theorem)**.** *Let R be a ring and I an ideal in R. There is a one-to-one correspondence between the ideals of R/I and the ideals of R that contain I.*

We're going to use this theorem in a few moments to prove some interesting results about quotient rings. To do this, we should really parse what the theorem says. If J is an ideal of R/I, then there is a unique ideal J' of R with $I \subseteq J' \subseteq R$. In fact, this ideal is precisely

$$\psi^{-1}(J) = \{r \in R : r + I \in J\}.$$

Likewise, if K is an ideal of R that contains I, so $I \subset K \subset R$, then there is a unique corresponding ideal K' of R/I. This ideal is actually given by

$$\psi(K) = \{a + I : a \in K\} = K/I.$$

3.4.1 Maximal Ideals

To complete our discussion of ideals, we will now introduce a special kind of ideal known as a *maximal ideal*. Our motivation is the following question: given a ring R and an ideal I, when is the quotient ring R/I a field? This question might seem strange at this point, but it will become important in our upcoming study of polynomial rings.

We saw previously that if F is a field, then the only ideals of F are $\{0\}$ and F itself. We'll start by proving the converse of this statement.

Theorem 3.4.6. *Let R be a commutative ring with identity, and suppose that the only ideals of R are $\{0\}$ and R. Then R is a field.*

Proof. We need to show that if $a \in R$ and $a \neq 0$, then a is a unit. Let $I = (a)$. Since $a \neq 0$, $I \neq \{0\}$, so we must have $I = R$. In particular, $1 \in I$. But then $1 = ra$ for some $r \in R$, so a is a unit with $r = a^{-1}$. Therefore, R is a field. □

We'll now use this fact to define maximal ideals. Our goal here is two-fold: we want to gain some tools that will be useful in our study of polynomials, but we also want to capture some of the nice properties of the integers in more general rings.

We'll start with the first goal. In studying polynomials, we'll sometimes want to form a quotient ring in order to obtain a field. That is, given a commutative ring R with identity, we want to find an ideal I of R such that R/I is a field. (We'll see why we want this sort of thing soon enough.) What would we need to know about I in order for this to work? Well, in light of the Correspondence Theorem and Theorem 3.4.6, it makes sense that I must satisfy the following definition.

Definition 3.4.7. Let R be a commutative ring with identity. An an ideal M of R is said to be **maximal** if the only ideals of R containing M are R and M.

With this definition in hand, we can strengthen Theorem 3.4.6 and completely characterize when R/I is a field.

Theorem 3.4.8. *Let R be a commutative ring with identity, and let M be an ideal of R. Then R/M is a field if and only if M is maximal.*

Proof. Suppose first that R/M is a field, but let's assume that there is an ideal I of R that contains M. By the Correspondence Theorem, the set

$$J = \{a + M \in R/M : a \in I\}$$

is an ideal in R/M. But R/M is a field, so J must be either $\{0+M\}$ or the whole ring R/M. In the first case, we must have $I = M$, and in the second $I = R$. Therefore, M is maximal.

On the other hand, suppose that M is maximal. Then the only ideals of R which contain M are M and R. As in the last case, I claim that the only possible ideals of R/M are of the form J as above, so the only ideals of R/M are the trivial ones. Thus R/M has no nontrivial ideals, so it is a field by Theorem 3.4.6. □

Example 3.4.9.

1. What are the maximal ideals in $\mathbb{Z}$? The ideals of $\mathbb{Z}$ are all of the form $n\mathbb{Z}$ for $n \in \mathbb{Z}$, and what would it mean for $n\mathbb{Z}$ to be maximal? If $n\mathbb{Z} \subseteq m\mathbb{Z}$, this means that $n \in m\mathbb{Z}$, so $m \mid n$. For $n\mathbb{Z}$ to be maximal, n cannot have any proper divisors, so it must be prime. Thus the maximal ideals of $\mathbb{Z}$ are exactly those of the form $p\mathbb{Z}$, where p is prime. Of course we have already seen that $\mathbb{Z}_p = \mathbb{Z}/p\mathbb{Z}$ is a finite field.

2. Recall that $I = \ker \mathrm{ev}_0 = (x)$ is an ideal in $\mathbb{Q}[x]$, and that $\mathbb{Q}[x]/I \cong \mathbb{Q}$ is a field. Therefore, $I = (x)$ is a maximal ideal in $\mathbb{Q}[x]$.

3.5 Polynomial Rings

Now that we have laid the foundations of basic ring theory, we will turn our attention to a specific class of rings—those whose elements are *polynomials*. In the process, we will finally be able to tie everything that we've learned together with the discussion that we had at the beginning of this class. We discussed the history of abstract algebra, and we mentioned that many algebraic techniques and constructions originally grew out of questions about polynomials and their roots. These algebraic ideas have taken on a life of their own since, but this is where their origins lie.

Classically, people studied polynomials with rational coefficients, and we will eventually focus on this case. However, we will begin by exploring polynomials whose coefficients come from an arbitrary commutative ring. Let's start by talking about what a polynomial is in a formal sense. Let R be a commutative ring with identity,

and let x be an **indeterminate**. That is, x is a sort of placeholder (or variable) which can be formally multiplied by itself and by ring elements. A **polynomial** over R is a formal expression of the type

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0,$$

where $a_i \in R$ for all i. The a_i are called the **coefficients** of the polynomial. We denote the set of all polynomials with coefficients in R by $R[x]$. One can add and multiply elements of $R[x]$ in the obvious ways, and it is straightforward to check the following:

Theorem 3.5.1. *$R[x]$ is a commutative ring with identity.*

It is also fairly easy to see that R can be viewed as a subset of $R[x]$, simply by identifying an element $a \in R$ with the constant polynomial $p(x) = a$. Thus R is a subring of $R[x]$, and the identity of $R[x]$ is the identity $1 \in R$.

Polynomial rings of course make sense for commutative rings with identity (and even for noncommutative rings), but most of the time we will be concerned with polynomials over a field. Therefore, we will fix a field F and consider the polynomial ring $F[x]$. This ring has particularly nice properties, and we will spend most of our time trying to gather all of them.

Polynomials vary wildly, and it can be hard to analyze them in general. However, there is one thing that we can always talk about, and it will give us a lot of control over polynomials.

Definition 3.5.2. Let $p(x) = a_n x^n + \cdots + a_1 x + a_0$ be a polynomial in $F[x]$ with $a_n \neq 0$. The integer n is called the **degree** of p, denoted by $\deg(p(x))$. We define the degree of the zero polynomial to be $\deg(0) = -\infty$.

An easy but useful fact is that the degree of a product of two polynomials is directly related to the degrees of the factors.

Proposition 3.5.3. *If $p(x), q(x) \in F[x]$, then*

$$\deg(p(x)q(x)) = \deg(p(x)) + \deg(q(x)).$$

Proof. Let $n = \deg(p(x))$ and $m = \deg(q(x))$, so the leading terms of p and q are $a_n x^n$ and $b_m x^m$, respectively. Then the leading term of pq is

$$(a_n x^n)(b_m x^m) = a_n b_m x^{n+m},$$

so $\deg pq = n + m = \deg(p) + \deg(q)$. □

This gives us another important fact, which is that $F[x]$ is always an integral domain. (More generally, $R[x]$ is an integral domain whenever R is.)

Theorem 3.5.4. *If F is a field, then $F[x]$ is an integral domain.*

Proof. Let $p, q \in F[x]$ with $p \neq 0$ and $q \neq 0$. Then $\deg(p), \deg(q) \geq 0$, so by the previous proposition, $\deg(pq) \geq 0$ as well. But this means that $pq \neq 0$, so $F[x]$ has no zero divisors. □

It is certainly nice that $F[x]$ is an integral domain, though it is not altogether unexpected. In reality, $F[x]$ is much more than that, and its properties will tell us much about polynomials in general. These properties will largely stem from the fact that $F[x]$ has its own version of the division algorithm—one can always divide one polynomial into another to obtain a quotient and remainder. This is done via the familiar method of polynomial long division.

Example 3.5.5. Consider the polynomials $p(x) = 2x^2 + 1, q(x) = x^4 - 7x + 1 \in \mathbb{Q}[x]$. We can divide p into q, and write

$$x^4 - 7x + 1 = (2x^2 + 1)\left(\frac{1}{2}x^2 - \frac{1}{4}\right) + \left(-7x + \frac{5}{4}\right).$$

Note that the degree of the remainder, $-7x + 5/4$, is less than that of p. This is a model for the division algorithm, and we appeal to examples like this in lieu of a formal proof to convince ourselves that it works.

Theorem 3.5.6 (Division Algorithm)**.** *Given two polynomials $f(x), g(x) \in F[x]$, there exist $q, r \in F[x]$ such that*

$$f(x) = q(x)g(x) + r(x),$$

with either $\deg(r(x)) < \deg(g(x))$, or $r(x) = 0$.

Using the Division Algorithm, we can prove that $F[x]$ is a particular type of integral domain, called a *principal ideal domain*.

Definition 3.5.7. An integral domain R is called a **principal ideal domain** (abbreviated **PID**) if every ideal in R is principal.

You have already seen one example of a PID, which is the prototype. Recall that every ideal in $\mathbb{Z}$ has the form $n\mathbb{Z}$ for some $n \in \mathbb{Z}$, hence it is principal. More generally, polynomial rings over fields are always PIDs.

Theorem 3.5.8. *If F is a field, then $F[x]$ is a PID. That is, if I is an ideal in $F[x]$, then there exists $h(x) \in F[x]$ such that*

$$I = (h(x)) = \{f(x)h(x) : f(x) \in F[x]\}.$$

Proof. Choose a polynomial $h(x) \in I$ with minimal degree; that is, $\deg(h(x)) \leq \deg(t(x))$ for all $t \in I$. Given $t \in I$, the division algorithm implies that there exist $q, r \in F[x]$ such that

$$t(x) = q(x)h(x) + r(x),$$

with either $r(x) = 0$ or $\deg(r(x)) < \deg(h(x))$. If we rewrite this, we have

$$r(x) = t(x) - q(x)h(x),$$

and since $h \in I$ and I is an ideal, $q(x)h(x) \in I$. But $t \in I$ as well, so $r \in I$. Since h has minimal degree in I, we must have $r = 0$. Therefore, $t(x) = q(x)h(x)$, and $t \in (h(x))$. Clearly $(h(x)) \subseteq I$, so it follows that $I = (h(x))$ and $F[x]$ is a PID. □

The fact that $F[x]$ is a PID is nice, but it's hard to see how it immediately relates to properties of polynomials. Though we won't prove it, this gives us information regarding the factorization of polynomials. Specifically, it tells us that any polynomial with coefficients in a field factors uniquely into a product of irreducible polynomials.

Definition 3.5.9. A polynomial $f \in F[x]$ is said to be **irreducible** if whenever

$$f(x) = g(x)h(x)$$

for $g, h \in F[x]$, then either $\deg(g) = 0$ or $\deg(h) = 0$.

An alternative way of saying that a polynomial is irreducible is that it cannot be written as a product of two polynomials of smaller degree.

Example 3.5.10. The polynomial $x^2 + 1$ is irreducible over $\mathbb{Q}$, since it cannot be factored into linear polynomials. It is reducible over $\mathbb{C}$, however, since $x^2 + 1 = (x+i)(x-i)$. On the other hand, x^2-1 is reducible over $\mathbb{Q}$, since $x^2-1 = (x+1)(x-1)$.

Theorem 3.5.11. *Let F be a field. Every polynomial $f \in F[x]$ can be written as a product of irreducible polynomials. That is, there are irreducible polynomials $p_1, \ldots, p_n \in F[x]$ and positive integers $m_1, \ldots, m_n$ such that*

$$f(x) = p_1(x)^{m_1} p_2(x)^{m_2} \cdots p_n(x)^{m_n}.$$

Furthermore, the factorization is unique: if $f(x) = q_1(x)^{k_1}q_2(x)^{k_2}\cdots q_n(x)^{k_j}$ *is another factorization into irreducibles, then* $n = j$ *and there is a reordering of the factors such that* $p_i(x) = a_i q_i(x)$ *for some* $a_i \in F$*, and* $m_i = k_i$ *for all* i*.*

This theorem can be proven directly, but it is quite tedious. It can also be proven by observing that this theorem is simply a statement of the fact that $F[x]$ is a special kind of ring known as a **unique factorization domain**, or **UFD**. It can then be proven that any PID is also a UFD. This proof requires a bit of extra technology, and we will not go into it here. We will simply accept the fact that polynomials factor uniquely into irreducibles.

Since irreducibility is obviously an important concept, we should develop some tools to help recognize it. We'll specialize to the case of polynomials with rational coefficients, and produce two results for checking whether a polynomial in $\mathbb{Q}[x]$ irreducible. The first provides an answer to a question that seems fairly innocuous, but has a surprisingly difficult resolution. Given a polynomial $f \in \mathbb{Z}[x]$ that is irreducible over $\mathbb{Z}$, does it remain irreducible over $\mathbb{Q}$? In other words, is it possible for a polynomial with integer coefficients to factor in $\mathbb{Q}[x]$, but not in $\mathbb{Z}[x]$? The answer is of course no, but the proof requires work. Because of this, we won't write it down.

Theorem 3.5.12 (Gauss's lemma)**.** *Let* $f(x) \in \mathbb{Z}[x]$*, and suppose that* $f(x) = g(x)h(x)$ *for some* $g, h \in \mathbb{Q}[x]$*. Then there exists* $u \in \mathbb{Q}$ *such that* $f(x) = g_1(x)h_1(x)$*, with* $g_1(x) = ug(x)$ *and* $h_1(x) = u^{-1}h(x)$*.*

A consequence of Gauss's lemma is a tool called Eisenstein's criterion. It allows us to check whether a polynomial with integer coefficients is irreducible over $\mathbb{Q}$. It looks like a strange criterion, but the proof should give some indication as to why it works.

Theorem 3.5.13 (Eisenstein's Criterion)**.** *Let* $f(x) = a_n x^n + \cdots + a_1 x + a_0$ *be a polynomial in* $\mathbb{Z}[x]$*, and suppose that* $p \in \mathbb{Z}$ *is a prime number such that:*

1. $p \mid a_i$ *for* $0 \le i \le n-1$*;*
2. $p \nmid a_n$*; and*
3. $p^2 \nmid a_0$*.*

Then f *is irreducible over* $\mathbb{Q}$*.*

Proof. Suppose to the contrary that f is not irreducible over $\mathbb{Q}$; that is, there are polynomials $g, h \in \mathbb{Q}[x]$ with $\deg(g), \deg(h) \ge 1$ and $f(x) = g(x)h(x)$. Then by

Gauss's lemma there are polynomials $g_1, h_1 \in \mathbb{Z}[x]$ such that $f(x) = g_1(x)h_1(x)$ and $\deg(g_1) = \deg(g)$, $\deg(h_1) = \deg(h)$. Write

$$g_1(x) = b_r x^r + \cdots + b_1 x + b_0$$

and

$$h_1(x) = c_s x^s + \cdots + c_1 x + c_0$$

with $1 \leq r, s < n$ and $b_r, c_s \neq 0$. Note that $a_0 = b_0 c_0$, and since $p \mid a_0$ but $p^2 \nmid a_0$, p must divide either b_0 or c_0 (but not both). Assume, without loss of generality, that $p \mid c_0$. Now observe that $a_n = b_r c_s$, and since $p \nmid a_n$, p cannot divide c_s. Therefore, there is some smallest index l, with $0 < l \leq s$, for which $p \nmid c_l$. Consider the coefficient a_l of f:

$$a_l = b_0 c_l + b_1 c_{l-1} + \cdots + b_{l-1} c_1 + b_l c_0.$$

We know that p divides a_l (since $l < n$) and that p divides the coefficients $c_0, c_1, \ldots, c_{l-1}$ of h_1. Therefore, we must have $p \mid b_0 c_l$. But we know that p divides neither b_0 nor c_l, so we have a contradiction. It follows that f must be irreducible over $\mathbb{Q}$. □

3.6 Roots of Polynomials

Most of our work regarding polynomials has been done with a specific goal in mind. We'd like to study roots of polynomials, especially polynomials over $\mathbb{Q}$. First, given a polynomial $f \in \mathbb{Q}[x]$, what does it mean for $\alpha \in \mathbb{Q}$ to be a **root** of f? It means that if we *evaluate* f at α, we get 0:

$$f(\alpha) = a_n \alpha^n + \cdots + a_1 \alpha + a_0 = 0.$$

Here's the big question: given a polynomial $f \in \mathbb{Q}[x]$, does f necessarily have a root in $\mathbb{Q}$? The answer is certainly no—for example, the quadratic polynomial $x^2 - 2$ has no roots in $\mathbb{Q}$.

Our primary goal moving forward will therefore be the following. Given a polynomial $f \in \mathbb{Q}[x]$ that does not have a root in $\mathbb{Q}$, can we find a field K containing $\mathbb{Q}$ such that f has a root in K? Also, can we find the smallest such field? To go one step further, can we find the smallest field containing $\mathbb{Q}$ which also contains *all* the roots of f? These are the main questions that we intend to address in the next couple of sections, and then we'll give a brief overview of the branch of algebra that comes from these ideas, called *Galois theory*.

Let's try to push our previous example a bit further in order to motivate the ensuing discussion.

Example 3.6.1. Consider the polynomial $f(x) = x^2 - 2$ in $\mathbb{Q}[x]$. We already mentioned that f does not have a root in $\mathbb{Q}$. What is the smallest field that contains

a root of f? Certainly $\mathbb{R}$ contains a root, but this is a massive field compared to $\mathbb{Q}$. Perhaps there is a smaller field which also contains a root. We could look at, say,

$$\mathbb{Q}(\sqrt{2}) = \left\{a + b\sqrt{2} : a, b \in \mathbb{Q}\right\}.$$

It's not hard to check that this set forms a field, and that $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt{2})$. Furthermore, it contains a root of f, namely $\sqrt{2}$.

Based on this example, the recipe is easy enough: we just take a root of the polynomial and "adjoin" it to $\mathbb{Q}$. It sounds easy enough, right? Not so much. There are some bigger philosophical questions lurking in the shadows here. First, what is $\sqrt{2}$, really? It is a real number with the property that it squares to 2. That is, it is *defined* to be a root of $x^2 - 2$. Before we can go tacking it on to $\mathbb{Q}$, we need to know that it even exists. That is, we need to prove that if $f \in \mathbb{Q}[x]$, there exists some field K with $\mathbb{Q} \subseteq K$ containing a root of f. If you know some real analysis, then you know that such a real number exists. However, this is not a course on analysis, and we will try to construct $\sqrt{2}$ *algebraically*.

Let's try to be more formal now. We'll introduce some terminology that will phrase our questions in a more mathematically precise manner.

Definition 3.6.2. Let K and F be fields with F a subfield of K. Then we say that K is a **field extension** of F.

Then the two basic goals that we have in mind can be rephrased in the following way.

1. Given a polynomial $f \in \mathbb{Q}[x]$, can we find the smallest extension K of $\mathbb{Q}$ that contains a root of f?

2. Can we find the smallest extension L of $\mathbb{Q}$ in which f **splits**? (That is, all the roots of f are contained in L.)

We'll start by addressing the first question. The first step is to obtain a field that contains a root of f, and then to determine a way of measuring how "small" it is relative to $\mathbb{Q}$.

Let F be a field and $f(x) \in F[x]$. Then we seek a field extension K of F such that f has a root in K. That is, since $F \subseteq K$, we can view f as a polynomial in $K[x]$, and we want to find an element $\alpha \in K$ such that $f(\alpha) = 0$. Let's try to simplify things first. Recall that since F is a field, every polynomial in $F[x]$ factors uniquely into a product of irreducible polynomials. In particular,

$$f(x) = p_1(x)^{m_1} p_2(x)^{m_2} \cdots p_n(x)^{m_n},$$

where the p_i are irreducible in $F[x]$. If we view $f \in K[x]$ and $\alpha \in K$ is a root of f, then we have

$$f(\alpha) = p_1(\alpha)^{m_1} \cdots p_n(\alpha)^{m_n} = 0$$

in K. Since K is an integral domain, we must have $p_i(\alpha) = 0$ for some i. That is, to produce a root of f, we simply need to construct a root of one of its irreducible factors.

The punchline of the above discussion is that we can restrict ourselves to irreducible polynomials over F. That is, we simply need to show that if $p \in F[x]$ is irreducible, then there is an extension K of F in which p has a root. How should we produce this extension? We want to somehow obtain a field from $F[x]$. The way to do this is to form a quotient by a maximal ideal. That is, if we can find a maximal ideal M that is somehow related to $p(x)$, then the quotient $F[x]/M$ would be a field that might be a candidate for K. It turns out that the appropriate choice for M is actually the principal ideal generated by $p(x)$.

Theorem 3.6.3. *Let F be a field. If $p(x)$ is an irreducible polynomial in $F[x]$, then the ideal $(p(x))$ is maximal in $F[x]$.*

Proof. Suppose that I is an ideal of $F[x]$ with $(p(x)) \subseteq I$. Since $F[x]$ is a PID, I is principal, so $I = (g(x))$ for some $g \in F[x]$. Since $p(x) \in (g(x))$, we have

$$p(x) = t(x)g(x)$$

for some $t \in F[x]$. But p is irreducible, so either $\deg(t(x)) = 0$ or $\deg(g(x)) = 0$. If $\deg(g(x)) = 0$, then g is constant, i.e. a unit in $F[x]$, and $(g(x)) = F[x]$. On the other hand, if $\deg(t(x)) = 0$, then t is constant, and g and p generate the same ideal. That is, $(g(x)) = (p(x))$. Therefore, either $I = (p(x))$ or $I = F[x]$, so $(p(x))$ is maximal. □

We are now ready to obtain an answer to our first question. If $p \in F[x]$ is irreducible, then there is an extension field of F, namely $F[x]/(p(x))$, in which p has a root.

Theorem 3.6.4. *If $p(x) \in F[x]$ is irreducible, then there is an extension K of F and an $\alpha \in K$ such that $p(\alpha) = 0$.*

Proof. We have already said that our candidate for K is $F[x]/(p(x))$, and we know that this is a field since $(p(x))$ is maximal. We just need to check that it contains an isomorphic copy of F and a root of p.

Define $\varphi : F \to K$ by $\varphi(a) = a + (p(x))$. To make things easier, we'll write $\overline{a} = a + (p(x))$ to denote the coset containing a in the quotient ring $F[x]/(p(x))$. It is easy to see that $\ker \varphi = \{0\}$, since

$$\varphi(a) = \overline{a} = 0$$

implies that $p(x)$ divides a, which is impossible. Thus φ is one-to-one, and F is isomorphic to a subfield of K.

Now we need to produce a root of p, and this is where things get a little weird. We first need to view p as a polynomial with coefficients in $F[x]/(p(x))$, or it wouldn't even make sense for p to have a root in K. If $p(x) = a_n x^n + \cdots + a_1 x + a_0$, then we can use the injection of F into K to view p as an element of $K[y]$: we simply apply φ to each coefficient, and we obtain a polynomial

$$\overline{p}(y) = \overline{a_n} y^n + \overline{a_{n-1}} y^{n-1} + \cdots + \overline{a_1} y + \overline{a_0} \in K[y].$$

Now that we've done some bookkeeping, we can get on with constructing the root. Strangely enough, the root is actually going to be $\alpha = x + (p(x)) = \overline{x}$. Let's try to see why. Viewing $p \in K[y]$, we have

$$\begin{aligned}
\overline{p}(\alpha) &= \overline{p}(\overline{x}) \\
&= \overline{a_n}(\overline{x})^n + \cdots + \overline{a_1}(\overline{x}) + \overline{a_0} \\
&= \overline{p}(\overline{x}) \\
&= p(x) + (p(x)) \\
&= 0.
\end{aligned}$$

Therefore, α is indeed a root of p. □

This proof is a little strange. It at least allows us to show that there is a field containing a root of p, but it's a very odd field to work with. Let's try to rephrase things in a slightly better way. Suppose that E is an extension of F containing a root α of p—we know now that at least one of these exists. Let $F(\alpha)$ be the subfield of E generated by α (i.e., the smallest field containing α). Then we have a map $\psi : F[x] \to F(\alpha)$ given by

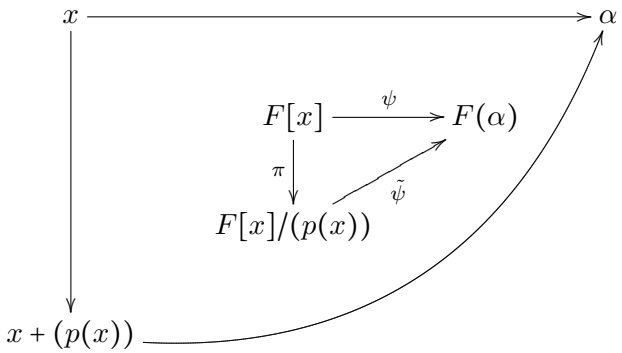
$$\psi(f(x)) = f(\alpha),$$

i.e.,

$$\psi(b_m x^m + \cdots + b_1 x + b_0) = b_m \alpha^m + \cdots + b_1 \alpha + b_0.$$

This is a ring homomorphism from F onto $F(\alpha)$. What is its kernel? It's actually equal to $(p(x))$, so the First Isomorphism Theorem gives an induced isomorphism $\tilde{\psi} : F[x]/(p(x)) \to F(\alpha)$. Thus the field we constructed in the theorem is really

nothing more than just F with α "adjoined" to it. We can even draw a diagram that illustrates this relationship.



In the argument above, we asserted that the kernel of "evaluation at α" was the ideal $(p(x))$. In order to actually know this, we need a couple of facts. First, we'll introduce some terminology.

Definition 3.6.5. Let K be an extension of F. An element $\alpha \in K$ is said to be **algebraic over** F if α is a root of some polynomial $f \in F[x]$.

In particular, the α that we constructed above is algebraic over F. The next fact tells us that any algebraic element always has a particular irreducible polynomial for which it is a root.

Theorem 3.6.6. *Suppose that α is algebraic over F. Then there is a unique monic irreducible polynomial $m_{\alpha,F} \in F[x]$ which has α as a root, called the* ***minimal polynomial*** *of α over F. If a polynomial $f \in F[x]$ has α as a root, then $m_{\alpha,F}$ divides f.*

In particular, if α is a root of an irreducible polynomial $p \in F[x]$ with leading coefficient a_n, then $m_{\alpha,F} = \frac{1}{a_n}p$. Thus $m_{\alpha,F}$ and p generate the same ideal in $F[x]$, and

$$F(\alpha) \cong F[x]/(m_{\alpha,F}(x)).$$

Note also that the minimal polynomial depends on the field that one is working over.

Example 3.6.7. Let's consider the polynomial $p(x) = x^2 - 2 \in \mathbb{Q}[x]$ again. This polynomial is irreducible as a result of Eisenstein's criterion, and it is monic, so it is the minimal polynomial of $\sqrt{2}$ over $\mathbb{Q}$. That is, $m_{\sqrt{2},\mathbb{Q}}(x) = x^2 - 2$. Since $\mathbb{Q}(\sqrt{2})$ is the smallest field containing $\sqrt{2}$, we have

$$\mathbb{Q}[x]/(x^2 - 2) \cong \mathbb{Q}(\sqrt{2}).$$

However, what is the minimal polynomial of $\sqrt{2}$ over $\mathbb{Q}(\sqrt{2})$? Since $\sqrt{2} \in \mathbb{Q}(\sqrt{2})$, we actually have

$$m_{\sqrt{2},\mathbb{Q}(\sqrt{2})}(x) = x - \sqrt{2}.$$

3.7 Field Extensions

Before we go further with our study of polynomials, we need to introduce some basic properties of field extensions. The main concept is that of the *degree* of a field extension, which will give us a great deal of information in computations. To motivate the idea, let's think about field extensions in general. Suppose that K is an extension of a field F. Then K is a field in particular, but we could strip it down and just observe that there are two things we can do in K. We can add elements of K: if $a, b \in K$, then

$$a + b \in K.$$

We can also multiply by elements of F: if $x \in F$ and $a \in K$, then

$$xa \in K.$$

Furthermore, this scalar multiplication by elements of F distributes over addition. What are we really saying that K is in this case? These two operations mean that we can view K as a *vector space* over F. Of course a vector space over a field comes with a number associated to it, namely the *dimension* of the vector space.

Definition 3.7.1. Let K be an extension of a field F. The **degree** of K over F is defined to be the dimension of K as a vector space over F:

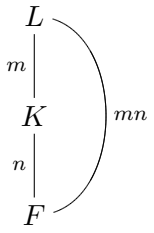
$$[K : F] = \dim_F(K).$$

This notation should look familiar, since we used the same sort of symbols to denote the index of a subgroup. We'll see later that this comparison is indeed warranted.

Definition 3.7.2. An extension K of F is said to be **finite** if $[K : F]$ is finite.

We'll mostly be interested in finite extensions, since we'll be able to associate an actual number to the extension to gauge how "large" it is in relation to $\mathbb{Q}$. Before we do some examples, let's state one useful fact about degrees of extensions. Though we won't prove it, it is reasonably intuitive. If we have an "extension of an extension," i.e. fields $F \subseteq K \subseteq L$, how should the degree of L over F relate to the degrees of L

over K and K over F? More specifically, if we have a tower of fields



then we would expect the degrees to multiply.

Theorem 3.7.3. *Let K be a finite extension of F, and let L be a finite extension of K (so $F \subseteq K \subseteq L$). Then*

$$[L : F] = [L : K][K : F].$$

In particular, the degree of the subfield K over F divides the degree of L.

This should again evoke a result from group theory—it looks similar to Lagrange's theorem.

Example 3.7.4. Let's compute the degrees of some field extensions.

1. We'll start with a sort of easy one. What is $[\mathbb{C} : \mathbb{R}]$? Since $\mathbb{C}$ can be viewed as a two-dimensional vector space over $\mathbb{R}$, the degree is 2.

2. What is $[\mathbb{R} : \mathbb{Q}]$? Trick question—this is not a finite extension. As a rational vector space, $\mathbb{R}$ is infinite-dimensional, so $[\mathbb{R} : \mathbb{Q}] = \infty$.

3. Let $K = \mathbb{Q}(\sqrt{2})$. We defined

$$\mathbb{Q}(\sqrt{2}) = \left\{a + b\sqrt{2} : a, b \in \mathbb{Q}\right\},$$

   so a basis for K over $\mathbb{Q}$ is $\{1, \sqrt{2}\}$. Therefore,

$$[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 2.$$

4. Here's another example with polynomials. Consider the polynomial $f(x) = x^3 - 2$ in $\mathbb{Q}[x]$. A root of f is $\alpha = \sqrt[3]{2}$, so the smallest field containing a root of f is $\mathbb{Q}(\sqrt[3]{2})$. Let's try to find a basis for this field over $\mathbb{Q}$. Certainly $\mathbb{Q}(\sqrt[3]{2})$ will need to contain the elements

$$1, \sqrt[3]{2}, (\sqrt[3]{2})^2, (\sqrt[3]{2})^3, \cdots$$

and any rational linear combinations of these. But note that $(\sqrt[3]{2})^3 = 2 \in F$, so we gain nothing new by including this element. That is, the elements

$$1, \sqrt[3]{2}, (\sqrt[3]{2})^2, (\sqrt[3]{2})^3$$

are linearly dependent over $\mathbb{Q}$. However, we claim that $\left\{1, \sqrt[3]{2}, (\sqrt[3]{2})^2\right\}$ is linearly independent, and thus forms a basis for $\mathbb{Q}(\sqrt[3]{2})$. To see this, suppose that these elements are linearly dependent, so there are $a, b, c \in \mathbb{Q}$ such that

$$a + b\sqrt[3]{2} + c(\sqrt[3]{2})^2 = 0.$$

Then $\sqrt[3]{2}$ is a root of the polynomial

$$g(x) = cx^2 + bx + c.$$

Then the minimal polynomial of $\sqrt[3]{2}$ must divide $g(x)$. But the minimal polynomial is $x^3 - 2$, which has degree 3, so the only way it can divide g is if $g = 0$. That is, $a = b = c = 0$, and the given elements are linearly independent. Thus we have shown that

$$[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3.$$

What do you notice about the degree of the field extension in each of the last two examples? The degree equals the degree of the minimal polynomial of the given algebraic element. The argument that we gave in the last example can be generalized to obtain a proof of the following theorem.

Theorem 3.7.5. *Let F be a field, and suppose that α is algebraic over F. Then $F(\alpha)$ is a finite extension of F and*

$$[F(\alpha) : F] = \deg(m_{\alpha,F}),$$

the degree of the minimal polynomial of α over F.

As we mentioned, the last example above gives us the idea of the proof, so we won't write down the proof here. However, we should observe that it also tells us how to find a nice basis for $F(\alpha)$ over F: if $n = \deg(m_{\alpha,F})$, then the set

$$\left\{1, \alpha, \alpha^2, \cdots, \alpha^{n-1}\right\}$$

is linearly independent and spans $F(\alpha)$. Therefore, we have the following corollary.

Corollary 3.7.6. *Let F be a field with α algebraic over F. Then*

$$F(\alpha) = \left\{a_0 + a_1\alpha + a_2\alpha^2 + \cdots + a_{n-1}\alpha^{n-1} : a_i \in F\right\}.$$

3.8 The Splitting Field of a Polynomial

In the previous sections we introduced a few problems related to polynomials with coefficients in a field F. The first was that if $p \in F[x]$ is an irreducible polynomial, is there a field extension K of F which contains a root of p? We answered this with a resounding yes, and in fact the smallest field is $F(\alpha)$, the field generated by F and α. Moreover, we actually quantified the "size" of $F(\alpha)$ compared to F: we saw last time that

$$[F(\alpha) : F] = \deg m_{\alpha,F},$$

the degree of the minimal polynomial of α over F.

The second question that we posed involved finding *all* the roots of a polynomial $f \in F[x]$. That is, can we find the smallest field extension of F containing all the roots of f? Equivalently, we're looking for the smallest field over which f **splits**, i.e., f factors completely into linear factors:

$$f(x) = (x - \alpha_1)^{m_1}(x - \alpha_2)^{m_2} \cdots (x - \alpha_k)^{m_k}.$$

This field is, conveniently enough, called the **splitting field** of f. What would we need to do to show that it exists? Well, we know that there is an extension $F(\alpha)$ containing a root α of f, and we can write

$$f(x) = (x - \alpha)^m g(x)$$

for some $g \in F(\alpha)[x]$. If g doesn't have any roots in $F(\alpha)$, we can extend further to obtain a root β of g, so

$$f(x) = (x - \alpha)^m (x - \beta)^k h(x)$$

for some $h \in F(\alpha, \beta)[x]$. Thus we obtain a tower of fields

$$\begin{array}{c} F(\alpha, \beta) \\ | \\ F(\alpha) \\ | \\ F \end{array}$$

If h doesn't have a root in $F(\alpha, \beta)$, we could extend further to obtain a root of h. This process needs to terminate eventually, since we would inevitably run up against the degree of f. Therefore, we have an idea as to why the splitting field exists, but we need to prove it formally. The proper way to do it is with a proof by induction.

Theorem 3.8.1. *Let F be a field and let $f \in F[x]$. Then there is an extension L of F over which f splits. Equivalently, L contains all the roots of f.*

Proof. We proceed by induction on the degree of f. If $\deg(f) = 1$, then f has only one root, which lies in F, so f splits over the field F. Suppose then that any polynomial of degree less than n has a splitting field, and suppose that $\deg(f) = n$. Then there is an extension K of F which contains a root α of f. That is,

$$f(x) = (x - \alpha)^m g(x)$$

for some $g \in K[x]$. If g also splits over K, then we are done. If not, then since $\deg(g) = n - 1$, there is an extension L of K over which g splits. But then $\alpha \in L$, so f splits over L as well. □

Once we know that there is a field over which f splits, we can talk about the *smallest* such field.

Definition 3.8.2. The smallest field containing all the roots of f is called the **splitting field** of f.

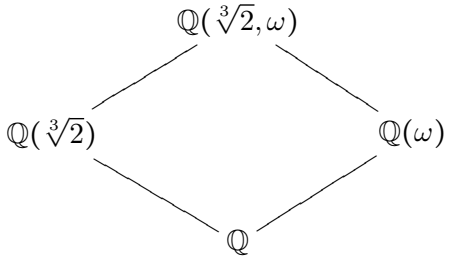
Example 3.8.3. Let's try to compute some examples of splitting fields.

1. Last time we constructed the field $\mathbb{Q}(\sqrt{2})$ to obtain a root of $p(x) = x^2 - 2$. It turns out that this is actually the splitting field of p, since

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

   over $\mathbb{Q}(\sqrt{2})$.

2. The other example that we had was $\mathbb{Q}(\sqrt[3]{2})$, which gave a root of $x^3 - 2$. This is *not* a splitting field, since the other roots of $x^3 - 2$ (which are actually $\sqrt[3]{2}\left(-\frac{1}{2} + i\frac{\sqrt{3}}{2}\right)$ and $\sqrt[3]{2}\left(-\frac{1}{2} - i\frac{\sqrt{3}}{2}\right)$) do not lie in $\mathbb{Q}(\sqrt[3]{2})$. If we let $\omega = \sqrt[3]{2}\left(-\frac{1}{2} + i\frac{\sqrt{3}}{2}\right)$, then the splitting field of $x^3 - 2$ is $\mathbb{Q}(\sqrt[3]{2}, \omega)$. What is its degree? Let's draw a picture first:



The degree $[\mathbb{Q}(\sqrt[3]{2} : \mathbb{Q}]$ is 3, while $[\mathbb{Q}(\sqrt[3]{2}, \omega) : \mathbb{Q}(\sqrt[3]{2})] = 2$, so the degree of the splitting field over $\mathbb{Q}$ is 6.

In light of these examples, what is the worst possible case for the degree of the splitting field of a polynomial over F? That is, if $f \in F[x]$ and L is its splitting field, what is the largest possible degree of L over F? The worst thing that could happen is that when we attach one root, we are left with

$$f(x) = (x - \alpha)g(x)$$

with $\deg(g) = n-1$. Then when we attach a root of g, we are left with an irreducible polynomial of degree $n-2$. Continuing in this fashion, we see that in the worst case,

$$[L : F] = n(n-1)(n-2)\cdots 2 \cdot 1 = n!.$$

In general, we must have $[L : F] \leq n!$, but the "generic" polynomial will have a splitting field of degree $n!$. We'll see in the next section (in a very brief sense) that this has major consequences for the solvability of the quintic equation in terms of radicals.

3.9 A Preview of Galois Theory

We began this course with a short lecture on the history behind abstract algebra. Most of that discussion went into the study of polynomials, and how renaissance mathematicians had tried to find formulas for the roots of polynomials. They succeeded in finding solutions for quadratic, cubic, and quartic polynomials, but the case of quintic (and higher degree) polynomials went unsolved for a couple of centuries.

As we mentioned in that initial lecture, it was shown in the 1820s that no formula exists for the roots of a general quintic polynomial in terms of the usual arithmetic operations and radicals. This was proven independently by Évariste Galois and Niels Henrik Abel. Both men developed what is now known as a **group** in order to analyze polynomials and gain information about their roots. Their work has developed over the last two centuries into a well-studied branch of algebra which bears Galois' name: it is known as **Galois theory**.

Now we're going to talk about the ideas that Galois used and take a small glimpse into Galois theory. In the process we'll see that algebra is really much more than just a collection of facts about groups, rings, and fields—things really fit together quite nicely, and algebra is actually a very beautiful subject. We won't use the same language and notation that Galois used; algebra has come a long way since his days, and we'll phrase things in more modern terms. We also won't be very rigorous—Galois theory is generally a whole course to itself, so we won't prove anything. The interested reader can just sit back and enjoy the ideas without being saddled with the details.

3.9.1 The Idea of Galois

In order to discuss Galois' ideas, we'll start by looking at an example involving a particular polynomial over the rationals.

Example 3.9.1. Let $f(x) = x^4 - 5x^2 + 6 \in \mathbb{Q}[x]$. What are the roots of f? Well, this polynomial factors as $f(x) = (x^2-2)(x^2-3)$, so the roots are $\pm\sqrt{2}$ and $\pm\sqrt{3}$. Now we'll introduce Galois' key idea: he decided to permute the roots of a polynomial and to see what happened. This might seem like a strange thing to do, and it is a testament to Galois' genius that he actually thought of this. He didn't allow just any old permutation of the roots, however. There was one rule: a root of f could only be sent to another root of the same irreducible factor of f. This requirement basically ensures that the permutations respect certain symmetries of the polynomial in question. Therefore, what are the valid permutations of the roots of $(x^2-2)(x^2-3)$? There are four in all:

$$\iota = \begin{cases} \sqrt{2} \mapsto \sqrt{2} \\ -\sqrt{2} \mapsto -\sqrt{2} \\ \sqrt{3} \mapsto \sqrt{3} \\ -\sqrt{3} \mapsto -\sqrt{3} \end{cases} \qquad \sigma_1 = \begin{cases} \sqrt{2} \mapsto -\sqrt{2} \\ -\sqrt{2} \mapsto \sqrt{2} \\ \sqrt{3} \mapsto \sqrt{3} \\ -\sqrt{3} \mapsto -\sqrt{3} \end{cases}$$

$$\sigma_2 = \begin{cases} \sqrt{2} \mapsto \sqrt{2} \\ -\sqrt{2} \mapsto -\sqrt{2} \\ \sqrt{3} \mapsto -\sqrt{3} \\ -\sqrt{3} \mapsto \sqrt{3} \end{cases} \qquad \sigma_3 = \begin{cases} \sqrt{2} \mapsto -\sqrt{2} \\ -\sqrt{2} \mapsto \sqrt{2} \\ \sqrt{3} \mapsto -\sqrt{3} \\ -\sqrt{3} \mapsto \sqrt{3} \end{cases}$$

Suppose we labeled the roots as $r_1 = \sqrt{2}$, $r_2 = -\sqrt{2}$, $r_3 = \sqrt{3}$, and $r_4 = -\sqrt{3}$. Then we can express these permutations in cycle notation as

$$\begin{aligned} \iota &= (1)(2)(3)(4) \\ \sigma_1 &= (1\ 2)(3)(4) = (1\ 2) \\ \sigma_2 &= (1)(2)(3\ 4) = (3\ 4) \\ \sigma_3 &= (1\ 2)(3\ 4). \end{aligned}$$

What do you notice about the set $\{\iota, \sigma_1, \sigma_2, \sigma_3\}$? It forms a subgroup of S_4 (which is actually isomorphic to the Klein 4-group $\mathbb{Z}_2 \times \mathbb{Z}_2$). Thus we've taken a polynomial f and associated a group to it, which we'll call the **Galois group** of f, denoted by $\mathrm{Gal}(f)$.

So we have a group associated to a polynomial—what's the big deal? How does the group tell us anything about the polynomial? To answer this, we need to translate everything into the language of field extensions, which is the modern approach to Galois theory.

3.9.2 Modern Galois Theory

Before we can talk about fields, we need an actual field to work with. The natural option is the splitting field of the polynomial.

Example 3.9.2. What is the splitting field of $f(x) = (x^2 - 2)(x^2 - 3)$? It's the smallest extension of $\mathbb{Q}$ containing $\sqrt{2}$ and $\sqrt{3}$, which is $K = \mathbb{Q}(\sqrt{2}, \sqrt{3})$. The "standard basis" for K as a vector space over $\mathbb{Q}$ is

$$\left\{1, \sqrt{2}, \sqrt{3}, \sqrt{2}\sqrt{3} = \sqrt{6}\right\}.$$

The permutations of the roots that we worked out earlier can be thought of as a "change of basis" for K. For example, the permutation σ_3 takes this basis to

$$\left\{1, -\sqrt{2}, -\sqrt{3}, \sqrt{6}\right\}.$$

In other words, each element of $\mathrm{Gal}(f)$ gives us a map from K back to itself which is actually a ring homomorphism. In fact, it is even more than that—it is an automorphism of K which fixes the subfield $\mathbb{Q}$. The set of all such automorphisms is denoted by

$$\mathrm{Gal}(K/\mathbb{Q}) = \{\sigma \in \mathrm{Aut}(K) : \sigma(a) = a \text{ for all } a \in \mathbb{Q}\},$$

and called the **Galois group** of K over $\mathbb{Q}$.[3] It turns out that these automorphisms correspond exactly to Galois' permutations of the roots of f. That is,

$$\mathrm{Gal}(K/\mathbb{Q}) = \mathrm{Gal}(f).$$

Here's a natural question to consider: what is the degree of K over $\mathbb{Q}$? We saw already that a basis for K over $\mathbb{Q}$ is $\{1, \sqrt{2}, \sqrt{3}, \sqrt{6}\}$, so $[K : \mathbb{Q}] = 4$. Note that $|\mathrm{Gal}(K/\mathbb{Q})| = 4$, so we have

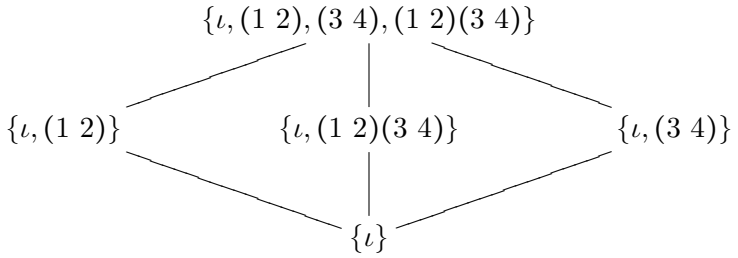
$$|\mathrm{Gal}(K/\mathbb{Q})| = [K : \mathbb{Q}].$$

It turns out that this is always the case when K is the splitting field of a polynomial. Now we're getting somewhere—we have a tight relationship between the splitting field of f and the Galois group. We can actually push this relationship even further, and see that it is even more intimate than it appears at first glance.

The point of assigning a group to a polynomial or a field is that the *structure* of the group tells us about the structure of the polynomial or field. We spent quite a bit of time studying the structure of groups, and analyzing their subgroups in

[3] Actually, any automorphism of K must fix $\mathbb{Q}$, so the Galois group is really the full automorphism group of K.

particular. In our example, what are the subgroups of $\mathrm{Gal}(K/\mathbb{Q})$? The subgroup lattice looks like:



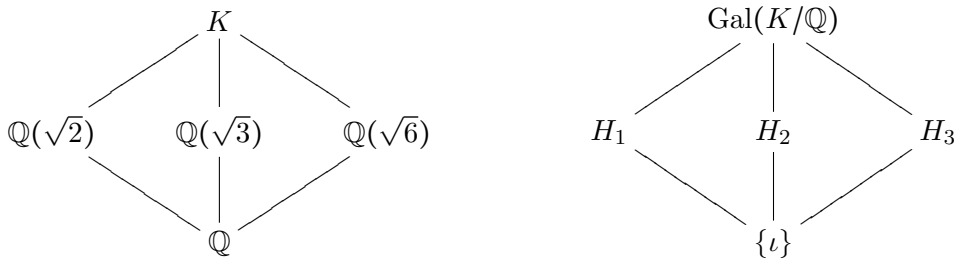
Let's look at the subgroup $H_1 = \{\iota, (3\ 4)\}$. What happens if we apply either of these permutations to an element of K? In either case, $\sqrt{2}$ and $-\sqrt{2}$ are fixed, so for any element of the form $a + b\sqrt{2}$, we have

$$\sigma(a + b\sqrt{2}) = a + b\sqrt{2}$$

for all $\sigma \in H_1$. That is, H_1 **fixes** all elements of the form $a + b\sqrt{2}$. Equivalently, H_1 fixes the subfield $\mathbb{Q}(\sqrt{2})$ of K. If we define

$$\mathrm{Gal}(K/\mathbb{Q}(\sqrt{2})) = \left\{\sigma \in \mathrm{Gal}(K/\mathbb{Q}) : \sigma(\alpha) = \alpha \text{ for all } \alpha \in \mathbb{Q}(\sqrt{2})\right\}$$

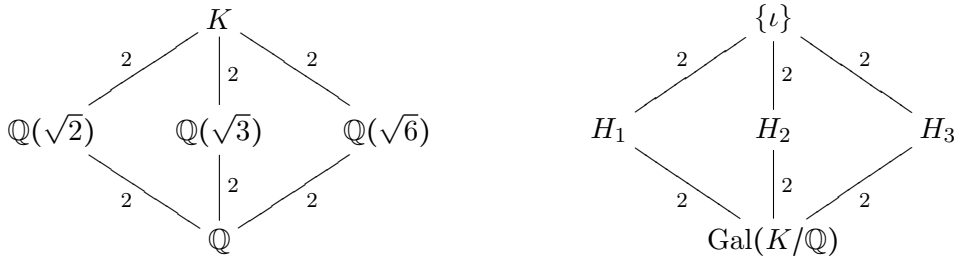
then $H_1 = \mathrm{Gal}(K/\mathbb{Q}(\sqrt{2}))$. Similarly, $H_2 = \{\iota, (1\ 2)\}$ fixes the subfield $\mathbb{Q}(\sqrt{3})$ and $H_3 = \{\iota, (1\ 2)(3\ 4)\}$ fixes $\mathbb{Q}(\sqrt{6})$. These are actually all the subfields of K, so we have a one-to-one correspondence between subfields of K and subgroups of the Galois group $\mathrm{Gal}(K/\mathbb{Q})$:



Now we're beginning to see how the structure of a Galois group might shed some light on a polynomial and its splitting field.

We can actually do even better than we have done so far. Let's redraw the above diagram, but with the subgroup lattice inverted, and write in the degrees of all the

extensions and the indices of all the subgroups:



Do you notice anything special? It's not so enlightening here, since all the degrees and indices are just 2. However, in general we have that if F is a subfield of K, then

$$[F : \mathbb{Q}] = [\mathrm{Gal}(K/\mathbb{Q}) : \mathrm{Gal}(K/F)].$$

The main facts to take away from this discussion are the following: if K is the splitting field of a separable polynomial over $\mathbb{Q}$ (also called a **Galois extension**), there is a group $G = \mathrm{Gal}(K/\mathbb{Q})$ such that:

1. $|\mathrm{Gal}(K/\mathbb{Q})| = [K : \mathbb{Q}]$
2. There is a one-to-one correspondence between subgroups of $\mathrm{Gal}(K/\mathbb{Q})$ and subfields of K.
3. If F is a subfield of K, then
$$[F : \mathbb{Q}] = [\mathrm{Gal}(K/\mathbb{Q}) : \mathrm{Gal}(K/F)].$$

These three statements actually form a major chunk of an important theorem, called the Fundamental Theorem of Galois Theory. The theorem actually says even more, but one needs to be looking at a nonabelian Galois group before such things can appear. The fact to which we are alluding is the following: if K is a Galois extension of $\mathbb{Q}$ and F is a subfield of K, then F is itself a Galois extension if and only if the subgroup $\mathrm{Gal}(K/F)$ is normal in $\mathrm{Gal}(K/\mathbb{Q})$.

Hopefully this discussion has shown you that there is a very strong relationship between groups and fields. The structure of the splitting field of a polynomial is perfectly reflected in the structure of the associated Galois group, and one can therefore cull information about the polynomial from the group. This gives one indication as to why it is important to understand the structure of groups well. As a specific example, the proof of the insolvability of the quintic equation comes from studying the structure of the Galois group. It turns out that a polynomial is solvable in terms of radicals if and only if its Galois group is what is called a **solvable group**. One can then easily come up with fifth degree polynomials which have S_5 as their Galois groups, and S_5 is easily proven to be a nonsolvable group.

We don't have the tools to write down a full proof here, but this gives an indication of the true power of Galois theory.

Additional Exercises for Chapter 2

3.2. Let R be an integral domain. If $a, b, c \in R$ with $a \neq 0$ and $ab = ac$, show that $b = c$.

3.3. Find the following products of quaternions.

(a) $(i + j)(i - j)$.

(b) $(1 - i + 2j - 2k)(1 + 2i - 4j + 6k)$.

(c) $(2i - 3j + 4k)^2$.

(d) $i(\alpha_0 + \alpha_1 i + \alpha_2 j + \alpha_3 k) - (\alpha_0 + \alpha_1 i + \alpha_2 j + \alpha_3 k)i$.

3.4. Recall that if $a \in \mathbb{Q}$, we can always write a in **lowest terms** (or **reduced form**) as $a = m/n$, where $\gcd(m, n) = 1$. Define

$$R = \left\{ \frac{m}{n} \in \mathbb{Q} : \gcd(m, n) = 1 \text{ and } n \text{ is odd} \right\}.$$

That is, R is the set of all rationals which, when written in lowest terms, have an odd denominator. Show that R is a ring under the usual addition and multiplication of rational numbers. Determine which elements of R are units. (This ring R is actually quite important in higher algebra. It is usually denoted by $\mathbb{Z}_{(2)}$, and called the *localization* of $\mathbb{Z}$ at 2.)

3.5. Show that any field is an integral domain.

3.6. Let R be a finite integral domain with identity $1 \in R$. Show that R is actually a field.

3.7. Let R be a ring, and suppose that I and J are ideals in R. Prove that $I \cap J$ is an ideal in R.

3.8. Let R be a commutative ring, and fix $a \in R$. Define the **annihilator** of a to be the set

$$\mathrm{Ann}(a) = \{x \in R : xa = 0\}.$$

Prove that $\mathrm{Ann}(a)$ is an ideal of R.

3.9. Let R be a commutative ring. An element $a \in R$ is said to be **nilpotent** if there is a positive integer n such that $a^n = 0$. The set

$$\mathrm{Nil}(R) = \{a \in R : a \text{ is nilpotent}\}$$

is called the **nilradical** of R. Prove that the nilradical is an ideal of R. [**Hint:** You may need to use the fact that the usual binomial theorem holds in a commutative ring. That is, if $a, b \in R$ and $n \in \mathbb{Z}^+$, then

$$(a + b)^n = \sum_{k=0}^{n} a^{n-k} b^k.$$

This should help with checking that $\mathrm{Nil}(R)$ is closed under addition.]

3.10. Let R and S be two rings with identity, and let 1_R and 1_S denote the multiplicative identities of R and S, respectively. Let $\varphi : R \to S$ be a nonzero ring homomorphism. (That is, φ does not map every element of R to 0.)

(a) Show that if $\varphi(1_R) \neq 1_S$, then $\varphi(1_R)$ must be a zero divisor in S. Conclude that if S is an integral domain, then $\varphi(1_R) = 1_S$.

(b) Prove that if $\varphi(1_R) = 1_S$ and $u \in R$ is a unit, then $\varphi(u)$ is a unit in S and

$$\varphi(u^{-1}) = \varphi(u)^{-1}.$$

3.11. Let R be a commutative ring with identity.

(a) Fix $\alpha \in R$. Define the **evaluation homomorphism at** α to be the map $\mathrm{ev}_\alpha : R[x] \to R$ given by: if $p(x) = a_n x^n + \cdots + a_1 x + a_0$ is in $R[x]$, then

$$\mathrm{ev}_\alpha(p) = a_n \alpha^n + \cdots + a_1 \alpha + a_0.$$

Show that ev_α is indeed a ring homomorphism.

(b) Determine the kernel of ev_α.

(c) Suppose now that $R[x]$ is a PID. Show that the kernel of ev_α is a maximal ideal, and conclude that R must be a field in this case.

3.12. Determine whether each of the following polynomials is irreducible over the given field.

(a) $3x^4 + 5x^3 + 50x + 15$ over $\mathbb{Q}$.

(b) $x^2 + 7$ over $\mathbb{Q}$.

(c) $x^2 + 7$ over $\mathbb{C}$.

Appendix A

Set Theory

This appendix is intended to expose you to some of the terminology and symbols that we will use throughout this course. Much of this information corresponds to the contents of Section 0 in Saracino's *Abstract Algebra: A First Course*.

A.1 Sets

Informally, a **set** is just a collection of objects. These could be numbers, people, physical objects, or pretty much anything else. For example, we could talk about

"the set of all Presidents of the United States"

or

"the set of all even integers."

There is actually a precise definition of a set, but we're not going to worry about it. (In general, problems can arise if one is not careful. If you're curious about this, look up *Russell's Paradox*.)

The objects that make up a set are called its **elements**. If we have a set S and a is an element of S, we write

$$\boxed{a \in S}.$$

If a is not an element of S, we instead write

$$\boxed{a \notin S}.$$

We can often define a set by simply listing all of its elements. When doing this, we usually enclose the list in braces: $\{\cdot\}$.

Example A.1.1. Define

$$S = \{1, 3, 5, 7, 18\}.$$

Then S is a set, and $3 \in S$. The number $15 \notin S$, since it is not on the list that we used to define S.

Example A.1.2. There is a special set $\varnothing$, called the **empty set**, which is defined to be the set which contains no elements.

Quite often it is infeasible to simply list all the elements of a set. For example, it would be impossible to list all of the even integers. In situations like this, we can describe a set by specifying that its elements should satisfy some defining property. We write

$$S = \{n : n \text{ is an even integer}\},$$

read "the set of all n such that n is an even integer." This sort of notation is called **set-builder notation**. (**Note:** Some authors use a vertical line | in place of the colon, but the meaning is the same.)

Example A.1.3. In linear algebra you learned about the span of a set of vectors v_1, v_2, v_3 in $\mathbb{R}^n$. In set-builder notation, one can write

$$\text{span}(v_1, v_2, v_3) = \{a_1v_1 + a_2v_2 + a_3v_3 : a_1, a_2, a_3 \text{ are real numbers }\}.$$

A.2 Constructions on Sets

There are a few operations on sets which will be important to know. The first will be the notion of a subset, and the others will allow us to build new sets out of old ones.

Definition A.2.1. Let A and B be sets. We say that B is a **subset** of A if every element of B is also an element of A. We write

$$\boxed{B \subseteq A}.$$

It is entirely possible a subset B of A could actually be all of A, i.e. $B = A$. If we want to emphasize that this is not the case, we will write

$$\boxed{B \subsetneq A}$$

and say that B is a **proper** subset of A.

Example A.2.2. Let

$$A = \{\text{Presidents of the United States}\},$$

and let

$$B = \{\text{George Washington, Franklin Pierce, Barack Obama}\}.$$

Then $B \subseteq A$, and in fact B is a proper subset of A.

Example A.2.3. Let

$$S = \{1, 3, 5, 7, 18\}.$$

Then the set $T = \{1, 7\}$ is a subset of S.

Example A.2.4. Let A be any set. Since the empty set $\varnothing$ contains no elements, all of its elements are contained in A. (Some would say that a statement like this is *vacuously true*.) Therefore, $\varnothing \subseteq A$. In other words, the empty set is a subset of *any* other set.

The next few constructions will allow us to construct new sets by combining two other sets in some way. They are called the union, intersection, difference, and Cartesian product.

Definition A.2.5. Let A and B be sets. The **union** of A and B, written

$$\boxed{A \cup B},$$

is the set whose elements consist of all the elements of A and all the elements of B. That is,

$$\boxed{A \cup B = \{x : x \in A \text{ or } x \in B\}}.$$

Example A.2.6. Let $A = \{\text{red}, \text{orange}\}$ and $B = \{\text{tiger, lion, bear}\}$. Then

$$A \cup B = \{\text{red, orange, tiger, lion, bear}\}.$$

Example A.2.7. Let $A = \{1, 2, 3\}$ and $B = \{2, 4, 6\}$. Then

$$A \cup B = \{1, 2, 3, 4, 6\}.$$

Remark A.2.8. Note that A and B are always subsets of $A \cup B$. Also, if A is any set, $A \cup \varnothing = A$.

Definition A.2.9. The **intersection** of two sets A and B, written

$$\boxed{A \cap B},$$

is the set whose elements are all the elements which lie in both A and B. That is,

$$\boxed{A \cap B = \{x : x \in A \text{ and } x \in B\}}.$$

Example A.2.10. If $A = \{1, 2, 3\}$ and $B = \{2, 4, 6\}$, then

$$A \cap B = \{2\}.$$

Example A.2.11. Let $A = \{\text{even integers}\}$ and $B = \{\text{odd integers}\}$. Then

$$A \cap B = \{x : x \text{ is both an even and an odd integer}\} = \varnothing,$$

since no integer is both even and odd.

Remark A.2.12. We always have $A \cap B \subset A$ and $A \cap B \subset B$. Also, if A is any set, $A \cap \varnothing = \varnothing$.

Definition A.2.13. The **difference** of A and B, written

$$\boxed{A - B},$$

consists of all of the elements of A which do not lie in B. That is,

$$\boxed{A - B = \{a \in A : a \notin B\}}.$$

Example A.2.14. Let $A = \{1, 2, 3, 4\}$ and $B = \{2\}$. Then

$$A - B = \{1, 3, 4\}.$$

Example A.2.15. Let $A = \{\text{integers}\}$ and $B = \{\text{even integers}\}$. Then

$$A - B = \{\text{odd integers}\}.$$

Definition A.2.16. The **Cartesian product** of A and B is

$$\boxed{A \times B = \{(a, b) : a \in A \text{ and } b \in B\}}.$$

That is, $A \times B$ is the set of all ordered pairs where the first coordinate is an element of A and the second is an element of B.

Example A.2.17. Let $A = \{1, 2, 3\}$ and $B = \{3, 4\}$. Then

$$A \times B = \{(1, 3), (1, 4), (2, 3), (2, 4), (3, 3), (3, 4)\}$$

A.3 Set Functions

We'll now give the definition of a function between sets. There are specific types of functions that we will eventually need to understand, but we will talk about them when the time comes.

Definition A.3.1. Let A and B be sets. A **function** (or **mapping**) f from A to B, written

$$\boxed{f : A \to B},$$

assigns to each element $a \in A$ exactly one element $f(a) \in B$.

Example A.3.2. Let $A = \{\text{real numbers}\}$ and $B = \{\text{nonnegative real numbers}\}$, and define $f : A \to B$ by

$$f(x) = |x|.$$

Then f is a function from A to B.

A.4 Notation

In addition to the notation that we have already highlighted here, there are some commonly used sets which have fairly standard symbols. We'll list them here for reference.

- $\mathbb{N}$: the set of all **natural numbers**, i.e. the nonnegative integers.
- $\mathbb{Z}$: the set of all integers.
- $\mathbb{Q}$: the set of all **rational numbers**, i.e., numbers that can be expressed as a ratio m/n, where m and n are integers.
- $\mathbb{R}$: the set of all real numbers.
- $\mathbb{C}$: the set of all complex numbers.
- $\mathbb{Z}^+$, $\mathbb{Q}^+$, $\mathbb{R}^+$: the set of all positive elements of $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$, respectively.

Appendix B

Techniques for Proof Writing

This guide includes some things that I like to keep in mind when I am writing proofs. They will hopefully become second-nature after a while, but it helps to actively think about them when one is first learning to write proofs.

B.1 Basic Proof Writing

There are some books out there that are designed to help you learn to read and write proofs. (A commonly–used one is *How to Read and Do Proofs* by Daniel Solow. It should be on reserve at Baker-Berry soon, and the library has two additional older copies available.) However, they can be somewhat dry and slow in getting off the ground. For me, the best way to learn to write proofs is to dive in and try to write up proofs of some reasonably simple statements. With this in mind, there are two fundamental aspects of proof writing that need to be mastered.

Logic: The first step to writing a proof, and probably the biggest hurdle for most people, is determining the logical steps needed to verify the given statement. In other words, this involves laying down an outline of the argument that you intend to make. There are some things to keep in mind when trying to do this.

- **What am I being asked to prove?** Often this will require you to unravel a definition or two to figure out what you're *really* trying to prove.
- **What are the hypotheses?** You're usually given some assumptions, and then you are asked to deduce something from them. This would usually be written in the form

  "If ..., then ..."

  Again, look back at the definitions and decide what the hypotheses are really saying.

- **What theorems might help?** Try to think of definitions and theorems that are related to the given statement. Determine which ones might help you get from the hypotheses to the desired result.
- **Put it all together.** Try to piece together the theorems that you've found in a logical way to deduce the result.

Example B.1.1. Let's try to put these ideas into action.

> **Prove:** If a and b are relatively prime and $a \mid bc$, then $a \mid c$.

Let's think about what we need to do here.

- What are we being asked to prove? We want to show that $a \mid c$, which means that we need to show that there is an integer n such that $c = na$.
- What are the hypotheses? There are two: we are told that a and b are relatively prime **and** that $a \mid bc$. The first means that

  $$\gcd(a, b) = 1,$$

  and the second tells us that there exists $m \in \mathbb{Z}$ such that

  $$bc = ma.$$

- What theorems do we have at our disposal? One of the theorems we proved regarding gcds was Bézout's lemma (or the extended Euclidean algorithm), which said that

  $$\gcd(a, b) = ax + by$$

  for some $x, y \in \mathbb{Z}$.
- Let's put it together:

  $$\gcd(a, b) = 1 \underset{\text{Bézout}}{\implies} 1 = ax + by \underset{\text{multiply by } c}{\implies} acx + bcy = c$$

  Now use the other hypothesis:

  $$\begin{aligned} a \mid bc &\implies bc = ma \\ &\implies bcy = may \\ &\implies acx + bcy = acx + may \\ &\implies c = a(cx + my) \\ &\implies a \mid c \end{aligned}$$

In this example we didn't actually write a proof. We simply outlined the argument, which is the backbone of the proof. Now we need to turn it into something readable. This brings us to the second major aspect of proof writing.

Style: Once you have your argument laid out, the next thing you need to do is to write it up in a nice way. Here are some tips for doing this.

- **Write in proper English.** Use complete sentences, with proper grammar and punctuation. The goal is to make it easy for the reader to understand. If you are unsure of how a particular sentence looks, read it back to yourself and think about how it would sound to the reader.
- **Be clear and precise.** Try to say what you mean in as simply as possible, while still using proper mathematical language. Be careful how you say things, and explain yourself at each step. If there is a step that you have to think about, or that you think may give the reader pause, explain it.
- **Don't say too much (or too little).** Again, explain yourself thoroughly, but don't overdo it. Get to the point, and avoid using overly ornate or mellifluous language. At this stage in the game, it's okay to err on the side of writing too much, but try to not overdo it.

These are all things that will become much easier with practice. Also, reading proofs in the book (or seeing them in class) will give you a better idea of how people tend to talk when they are writing proofs.

Example B.1.2. Let's write up a proper proof of the example.

Prove: If a and b are relatively prime and $a \mid bc$, then $a \mid c$.

Proof. Since a and b are relatively prime, $\gcd(a, b) = 1$, and Bézout's lemma lets us write

$$ax + by = 1$$

for some $x, y \in \mathbb{Z}$. If we multiply both sides by c, we get

$$acx + bcy = c.$$

We are assuming that $a \mid bc$, so there is an $m \in \mathbb{Z}$ such that $bc = ma$. Then

$$bcy = may,$$

so

$$c = acx + bcy = acx + may.$$

Factoring out a, we get

$$c = a(cx + my).$$

But $cx + my \in \mathbb{Z}$, and this is precisely what it means for a to divide c. □

Here are two more examples of simple proof–writing exercises. We will approach them in the same manner as the previous example.

Example B.1.3. Prove: Let $a, b \in \mathbb{Z}$, and let $m > 0$ be an integer. Then

$$\gcd(ma, mb) = m \cdot \gcd(a, b).$$

Let's outline our plan of attack.

- What are we trying to prove? We need to show that $\gcd(ma, mb) = m \cdot \gcd(a, b)$. Our plan will be to show that $\gcd(ma, mb) \leq m \cdot \gcd(a, b)$, and that $m \cdot \gcd(ma, mb) \leq \gcd(ma, mb)$.
- What are the hypotheses? We are simply given that $a, b, m \in \mathbb{Z}$, and that $m > 0$.
- What theorems or definitions might be useful? We know that $d = \gcd(a, b)$ divides a and b, so md divides both ma and mb. Also, Bézout's lemma says that there are integers x and y satisfying

  $$ax + by = d,$$

  so

  $$max + mby = md.$$

- Put it all together: $md \mid ma$ and $md \mid mb \implies md$ is a (positive) common divisor of $ma, mb \implies md \leq \gcd(ma, mb)$. Also,

  $$max + mby = md \implies \gcd(ma, mb) \mid md,$$

  so $\gcd(ma, mb) \leq md$. Thus $\gcd(ma, mb) = m \cdot \gcd(a, b)$.

Now we'll write it up.

Proof. Let $d = \gcd(a, b)$. Since d divides both a and b, md divides both ma and mb. Since $m > 0$, md is a positive common divisor of ma and mb, so it must be smaller than the greatest common divisor. That is, $md \leq \gcd(ma, mb)$. Also, Bézout's lemma implies that there are integers x and y satisfying

$$ax + by = d.$$

Multiplying both sides by m, we get

$$max + mby = md.$$

Since $\gcd(ma, mb)$ divides both ma and mb, it divides the left side of this equation. Thus $\gcd(ma, mb)$ divides md, so we must have

$$\gcd(ma, mb) \leq md.$$

Therefore, $md = \gcd(ma, mb)$, or $m \cdot \gcd(a, b) = \gcd(ma, mb)$. □

Example B.1.4. Prove: The equation

$$ax + by = c$$

has integer solutions x and y if and only if $\gcd(a, b)$ divides c.
There are two directions here, so we need to handle them one at a time.

- For the first direction, what are we being asked to prove? We need to show that $\gcd(a, b)$ divides c.
- What are the hypotheses? We are given that there are integers x and y such that $ax + by = c$.
- What theorems or definitions might be useful? We'll use the definition of the greatest common divisor, namely that it dives a and b. If we let $d = \gcd(a, b)$, we can write

  $$a = ed \quad \text{and} \quad b = fd$$

  for some integers e and f.
- Now let's put it together.

  $$\begin{aligned} a = ed \quad \text{and} \quad b = fd &\implies c = ax + by = edx + fdy \\ &\implies c = d(ex + fy) \\ &\implies d \text{ divides } c \end{aligned}$$

- What do we need to do for the other direction? We assume that $\gcd(a, b)$ divides c, and we show that $ax + by = c$ has integer solutions.
- What can we use? First, if $d = \gcd(a, b)$ divides c, we can write $c = kd$ for some $k \in \mathbb{Z}$. Second, we have Bézout's lemma: there exist $x_0, y_0 \in \mathbb{Z}$ such that

  $$ax_0 + by_0 = d.$$

- Now put it together:

  $$\begin{aligned} ax_0 + by_0 = d &\implies kax_0 + kby_0 = kd = c \\ &\implies a(kx_0) + b(ky_0) = c \end{aligned}$$

  so we can take $x = kx_0$ and $y = ky_0$.

Now we'll try to write it up nicely.

Proof. Suppose first that there are integers $x, y \in \mathbb{Z}$ such that $ax + by = c$. Let $d = \gcd(a, b)$. Since d divides both a and b, there are integers $e, f \in \mathbb{Z}$ such that $a = ed$ and $b = fd$. Then

$$ax + by = edx + fdy = d(ex + fy).$$

But $ax + by = c$, so

$$c = d(ex + fy),$$

and d divides c.

Conversely, suppose that d divides c. Then there is an integer k satisfying $c = kd$. By Bézout's lemma, there exist $x_0, y_0 \in \mathbb{Z}$ such that

$$ax_0 + by_0 = d.$$

Thus

$$k(ax_0 + by_0) = kd,$$

or

$$a(kx_0) + b(ky_0) = c.$$

If we set $x = kx_0$ and $y = ky_0$, then $ax + by = c$, so we are done. □

As a final note on style, people usually use a symbol to indicate the end of a proof. The most common is a simple square, which can be open or filled: □ or ■. (The default in LaTeX is an open square.) Some people will also use an open or closed diamond, or double or triple hatch marks (// or ///). Older proofs sometimes end with Q.E.D, which is an abbreviation of the Latin "quod erat demonstrandum," or "which was to be demonstrated."

B.2 Proof by Contradiction

The proofs we've written so far are *direct proofs*: we started with the hypotheses and we made a chain of logical deductions to eventually prove the given statement. This is generally the most desirable way to prove something, but it may not always work. Even if it does work, it may not be the best way. To this end, there is another proof technique called **proof by contradiction**.

Proof by contradiction, or in Latin, *reduction ad absurdum*, is an alternative to writing a direct proof. Instead of assuming the hypotheses and directly proving the result, you assume the hypotheses and you *assume that the result is not true.* You then try to make logical deductions until you arrive at a *contradiction*, which is a sort of logical conundrum. This contradiction should then lead you to conclude that there is a faulty assumption somewhere, and the only possibility is the assumption that the result is false. In summary:

Proof by Contradiction: Suppose that you are asked to prove a statement of the form

"If A, then B."

To prove this by contradiction:

1. Assume (A) and (not B).
2. Investigate the logical implications of these assumptions. (Use any theorems, definitions, etc. that you know.)
3. Arrive at a contradiction.
4. Conclude that B must be true after all.

There is one thing that should be noted before we continue. A proof by contradiction is not generally considered to be an aesthetically pleasing proof, and the technique should always be used as a last resort. That is, you should always try first to prove something directly, and then attempt a contradiction proof if a direct proof is too difficult. However, contradiction is sometimes the only way, and sometimes it may even give a nicer proof than those that can be obtained directly.

The following example is the oldest known proof by contradiction. There are actually many other known proofs of this statement, but the contradiction proof is still well–known due to its simplicity.

Example B.2.1. Prove that there are infinitely many prime numbers.

Proof. Suppose not—that is, let's assume that there are only finitely many prime numbers, say

$$p_1, p_2, \ldots, p_n.$$

Consider the integer

$$N = p_1 p_2 \ldots p_n + 1.$$

Observe that none of the primes $p_1, \ldots, p_n$ divide N. Since we are assuming that these are all the prime numbers, N has no prime divisors. But every integer can be written as a product of primes, so we have arrived at a contradiction. Therefore, our assumption that there are only finitely many primes must be faulty. We can thus conclude that there must be infinitely many primes. □

Here is another example. The details are a little more straightforward, and you simply need to "follow your nose" after making the necessary assumptions.

Example B.2.2. Suppose that $a \in \mathbb{Z}$. Prove that if a^2 is even, then a is also even.

Proof. Assume that a^2 is even, but that a is odd. Then

$$a = 2n + 1$$

for some $n \in \mathbb{Z}$. If we compute a^2, we get

$$a^2 = (2n+1)^2 = 4n^2 + 4n + 1 = 2(2n^2 + 2n) + 1,$$

which is odd. But we are assuming that a^2 is even, so we have arrived at a contradiction. Therefore, our assumption that a is odd must be invalid, and we can conclude that a must be even. □

Here's one for you to try on your own if you want more practice. It is a commonly-studied proof by contradiction, so you can probably find the proof written down in any number of places.

Exercise B.1. Prove that $\sqrt{2}$ is an irrational number.

B.3 Mathematical Induction

There is another proof technique, called the *Principle of Mathematical Induction*, which is used in special situations. One generally employs it to prove a statement that depends on an integer n. For example, you might be asked to prove that some formula, written in terms of n, holds for all $n \in \mathbb{Z}$. If you are lucky, you might be able to prove it directly. However, it's possible to envision, at least in principle, a systematic way of writing down such a proof. You could prove the result for $n = 1$, and then use this fact to *induce* the result for $n = 2$. You could then prove it for $n = 3$, then $n = 4$, and so on, proving each case by using the previous one. Obviously we can't actually write a proof this way—it would require a lot of work, and we'd have to prove an infinite number of cases. Fortunately, mathematical induction gives us the ability to do all of this in one fell swoop.

Principle of Mathematical Induction: Suppose that you are asked to prove that a statement $P(n)$, depending on $n \in \mathbb{Z}^+$, is true for all n. To prove this via induction, there are two steps:

Base case: Prove that $P(1)$ is true.

Inductive step: Assume that $P(n-1)$ is true, and use this to prove that $P(n)$ is true.

Example B.3.1. Prove that for all $n \in \mathbb{Z}^+$,

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

Proof. We need to check first that the formula holds for $n = 1$. The left side is simply 1, and the right side is

$$\frac{1}{2} \cdot 1 \cdot (1+1) = \frac{1}{2} \cdot 1 \cdot 2 = 1,$$

so the formula holds for $n = 1$.

Now we need to handle the inductive step. Assume that the formula holds for $n-1$, i.e., that

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

Then

$$\begin{aligned} 1 + 2 + \cdots + n &= 1 + 2 + \cdots + (n-1) + n \\ &= \frac{1}{2}(n-1)n + n \\ &= \left(\frac{1}{2}(n-1) + 1\right)n \\ &= \left(\frac{1}{2}(n-1+2)\right)n \\ &= \frac{1}{2}n(n+1). \end{aligned}$$

□

Now try the following example, which actually relates to abstract algebra.

Example B.3.2. If G is an abelian group and $a, b \in G$, prove that

$$(ab)^n = a^n b^n$$

for all $n \in \mathbb{Z}^+$.

Proof. For $n = 1$, we simply have

$$(ab)^1 = ab = a^1 b^1,$$

so the base case holds. Now assume that $(ab)^{n-1} = a^{n-1}b^{n-1}$. Then

$$(ab)^n = (ab)^{n-1}(ab) = a^{n-1}b^{n-1}ab$$

by assumption. Since G is abelian,

$$a^{n-1}b^{n-1}ab = a^{n-1}ab^{n-1}b = a^nb^n.$$

Therefore, the result holds by induction. □

If you want to read more about mathematical induction, or if you want to try other problems, the latter half of Section 0 in Saracino is devoted to induction. There are more examples, and there are several exercises that would allow you to practice proofs by induction.

B.4 Proof by Contrapositive

So far we've practiced some different techniques for writing proofs. We started with direct proofs, and then we moved on to proofs by contradiction and mathematical induction. The method of contradiction is an example of an **indirect proof**: one tries to skirt around the problem and find a clever argument that produces a logical contradiction. This is not the only way to perform an indirect proof—there is another technique called **proof by contrapositive**.

Suppose that we are asked to prove a **conditional statement**, or a statement of the form

"If A, then B."

We know that we can try to prove it directly, which is always the more enlightening and preferred method. If a direct proof fails (or is too hard), we can try a contradiction proof, where we assume $\neg B$ and A, and we arrive at some sort of fallacy. It's also possible to try a proof by contrapositive, which rests on the fact that a statement of the form

"If A, then B." $(A \implies B)$

is logically equivalent to

"If $\neg B$, then $\neg A$." $(\neg B \implies \neg A)$

The second statement is called the **contrapositive** of the first. Instead of proving that A implies B, you prove directly that $\neg B$ implies $\neg A$.

> **Proof by contrapositive:** To prove a statement of the form "If A, then B," do the following:
>
> 1. Form the contrapositive. In particular, negate A and B.
> 2. Prove directly that $\neg B$ implies $\neg A$.

There is one small caveat here. Since proof by contrapositive involves negating certain logical statements, one has to be careful. If the statements are at all complicated, negation can be quite delicate. However, sometimes the given proposition already contains certain negative statements, and contrapositive is the natural choice.

Example B.4.1. Prove by contrapositive: Let $a, b, n \in \mathbb{Z}$.If $n \nmid ab$, then $n \nmid a$ and $n \nmid b$.

Proof. We need to find the contrapositive of the given statement. First we need to negate "$n \nmid a$ and $n \nmid b$." This is an example of a case where one has to be careful, the negation is

$$\text{“}n \mid a \textbf{ or } n \mid b.\text{”}$$

The "and" becomes an "or" because of DeMorgan's law. The initial hypothesis is easy to negate: $n \mid ab$. Therefore, we are trying to prove

$$\text{“If } n \mid a \text{ or } n \mid b, \text{ then } n \mid ab.\text{”}$$

Suppose that n divides a. Then $a = nc$ for some $c \in \mathbb{Z}$, and

$$ab = ncb = n(cb),$$

so $n \mid ab$. Similarly, if $n \mid b$, then $b = nd$ for some $d \in \mathbb{Z}$, and

$$ab = and = n(ad),$$

so $n \mid ab$. Therefore, we have proven the result by contraposition. □

Here's another example. In this one, a direct proof would be awkward (and quite difficult), so contrapositive is the way to go.

Example B.4.2. Prove by contrapositive: Let $x \in \mathbb{Z}$. If $x^2 - 6x + 5$ is even, then x is odd.

Proof. Suppose that x is even. Then we want to show that $x^2 - 6x + 5$ is odd. Write $x = 2a$ for some $a \in \mathbb{Z}$, and plug in:

$$\begin{aligned} x^2 - 6x + 5 &= (2a)^2 - 6(2a) + 5 \\ &= 4a^2 - 12a + 5 \\ &= 2(2a^2 - 6a + 2) + 1. \end{aligned}$$

Thus $x^2 - 6x + 5$ is odd. □

B.5 Tips and Tricks for Proofs

The four proof techniques that we've talked about are really the only ones that people ever use. However, there are some general tips regarding the types of statements that you may be asked to prove. You've probably seen many of these via example in these notes, but we'll list them here anyway.

If and only if: Sometimes you are asked to prove something of the form "A if and only if B" or "A is equivalent to B." The usual way to do this is to prove two things: first, prove that "A implies B," and then prove that "B implies A." Use any of the possible techniques to prove these two implications.

Uniqueness: You are often asked to prove that some object satisfying a given property is unique. We've seen before that the standard trick is to assume that there is another object satisfying the property, and then show that it actually equals the original one.

Existence: This sort of proof often goes hand in hand with uniqueness. You are given some specified property, and then asked to show that an object exists which has that property. There are often two ways to do this. One can offer up a **constructive** proof, in which the object is explicitly constructed. A nonconstructive proof is the exact opposite—it shows that the object exists, but it gives no indication as to what the object looks like.

Existence and Uniqueness: As stated before, existence and uniqueness go hand in hand. Sometimes you will be asked to prove that an object satisfying some property exists and is unique. This can be done in either order. Sometimes it is easy to prove that the object exists, and then to show that it is unique. However, the existence proof may seem daunting, and it is often helpful to prove uniqueness first. The uniqueness proof may give some hints as to what the object must look like.