Introduction to Abstract Algebra
(Draft)

Christopher J Leininger

March 23, 2017

Contents

Note to students.

Abstract algebra is a part of *pure mathematics*, which involves building an understanding "from the ground up". We make some basic assumptions, which we call *axioms*, and then **prove** consequences of these assumptions, which take the form of *theorems* (stand-alone, important consequences), *propositions* (useful facts), *lemmas* (facts which serve as "stepping stones" to propositions or theorems), and *corollaries* (facts which are "easily" deduced from theorems, propositions, or lemmas).

It is natural to ask: *Where do the axioms in pure mathematics come from?* The short (and mostly uninformative) answer is that they can come from anywhere. When trying to study a physical, biological, chemical, or mathematical system, there are so many features that we can get lost in the sea of information. It is often helpful to filter out some of the "noise" by looking for the key features of the system, and then seeing what consequences we can draw. Deciding what the key features are depends on what aspects of the system one is interested in. Even with this in mind, it can take a very long time to figure out what's important and what's noise. Quite often, people recognize similarities in different systems and it is precisely these similarities which turn out to be relevant features. In this case, it is useful to remove the noise entirely by working with in an *abstract* setting in which we do not assume to be studying any particular system, but instead we simultaneously study **ALL** systems that share the common features. These common features become the axioms, and so we are reduced to studying the consequences of the axioms.

As with most pure mathematics, abstract algebra requires a familiarity with the basic foundations of mathematics. **Everything** will be defined in terms of sets and functions. Indeed, the central abstract algebraic objects we will study—*groups*, *rings*, *fields*, and *vector spaces*—are comprised of sets together with operations (defined by functions) satisfying certain axioms. Consequently, most of the first chapter of this book is dedicated to a review of sets, functions, operations, and relations. Without being fluent in the language of sets and functions, the student is doomed to fail. The first chapter will also fix notation and terminology to be used throughout to ensure that we are "speaking the same dialect of this language".

The first chapter also includes a mostly self-contained discussion of the basic properties of arithmetic in the integers, including addition, multiplication, divisibility, primes, etc. and modular arithmetic. This is in the first chapter for a couple reasons. First of all, the integers (together with its arithmetic) provides a concrete examples in abstract algebra, as well as "building blocks" for other, more complicated examples. Second, when a group, ring, etc. is finite, the size is an integer, and its arithmetic properties—especially the prime factorization—can have a profound effect on its structure. Third, the proofs we supply will serve to illustrate what a proof should look like, as well as giving examples of proof techniques like *proof by contradiction* and *proof by induction*.

Daggers Throughout these notes, you will see that certain theorems, propositions, lemmas, and corollaries are marked with a dagger †. You should know how these proofs go as you will be asked to provide a proof for at least one of them on each of the exams. These proofs are short, but they are plentiful. There are two things to point out here: (1) you should be able to write proofs, and having some specific examples to work with is useful and (2) you should not memorize the proofs verbatim, but instead you should understand the key ideas, and be able to turn those ideas into a proof. At the end of every section, there are exercises. You should do all the exercises. There is also a brief list of things you should know and be able to do. These will serve as the "review sheets" for the exams.

These notes are not a comprehensive introduction to Abstract Algebra. Instead, they are a single source for the material covered in the Math 417 course at the U of I, organized for the Spring '16 section M13.

Chapter 1

Preliminaries

1.1 Basics: Sets, functions, and operations.

1.1.1 Sets

We will not attempt a formal definition of what a set is, but will rather take the naive approach and simply view a **set** as any collection of **objects**. The objects are also called the **elements** of the set, and sometimes even the **points** of the set. Although not a hard-and-fast rule, sets will *usually* be denoted with capital Roman letters A, B, C, D, X, Y, Z, etc, and the elements of a set will *usually* be denoted with lower case Roman letters a, b, c, x, y, z, etc. We write $a \in A$ to mean "a is an element of A", and $a \notin A$ to mean "a is not an element of A". If A is a set, and $a, b \in A$, then we write $a = b$ to mean that a and b are the same element in A.

The **natural numbers** will be denoted by

$$\mathbb{N} = \{0, 1, 2, 3, \ldots\},$$

and the **integers** will be denoted

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

The **positive integers** will be denoted $\mathbb{Z}_+$ and is the same thing as $\mathbb{N}$, but without 0.

The integers are formed from the natural numbers by adding the negatives of all the positive natural numbers. Other sets of interest are the **rational numbers** $\mathbb{Q}$ and the **real numbers** $\mathbb{R}$, which can also be defined from $\mathbb{N}$ (though this is a bit more complicated). We will assume the readers familiarity with all of these number systems, $\mathbb{N}, \mathbb{Z}, \mathbb{Q}$, and $\mathbb{R}$, as well as with their **arithmetic**, meaning addition, multiplication, subtraction, and division, where these are defined. We will discuss these examples and arithmetic in more detail later.

With these examples, we have illustrated one method for specifying a set. Namely, we simply listed the elements inside the *braces* { } (these are also sometimes called *curly brackets*). Actually, we didn't list the elements, but instead we listed enough of the elements to deduce the pattern, then wrote "..." meaning that the pattern should be continued indefinitely. For the integers, the notation indicates that we are to continue the pattern indefinitely both forward and backward. Similarly, we might also denote the set of integers between 1 and 100 as

$$\{1, 2, 3, \ldots, 99, 100\}.$$

Using "..." is typically an imprecise way to describe a set because it assumes the reader will deduce the pattern the author has in mind. In the examples just given, this is not an issue as the author provided sufficient descriptive information—"the natural numbers", "the integers", and "the integers between 1 and 100"—for the reader to know what the sets in question are. Indeed, explicitly listing the elements (using "...") is only there to help remind the reader what these sets are.

If A and B are sets, and all the elements of A are also elements of B, then we say that A **is a subset of** B, or that A **is contained in** B, and we write $A \subset B$ or $B \supset A$. We will use these two notations interchangeably. We note that $A \subset B$ also allows the *possibility* that $A = B$, meaning that these two sets are the same (they consist of the exact same collection of elements). Indeed, a useful way to prove that two sets A and B are equal is to prove both $A \subset B$ and $B \subset A$.

If $A \subset B$ and $A \neq B$, then we say that A **is a proper subset of** B, and we will sometimes express this by writing $A \subsetneq B$. For example, $\mathbb{Z}_+ \subsetneq \mathbb{N}$ and $\mathbb{N} \subsetneq \mathbb{Z}$. The **empty set**, denoted $\emptyset$, is the set with no elements. We declare that the empty set is a subset of every set: $\emptyset \subset A$ for every set A.

We will frequently describe a set in terms of properties that its elements satisfy. This is most often done with a subset of some larger set. More precisely, if S is a set and P is some property that an element of S may or may not satisfy, then the subset $A \subset S$ of **elements in** S **satisfying the property** P is denoted

$$A = \{a \in S \mid a \text{ satisfies property } P\}.$$

This should be read as "the set of elements a in S such that a satisfies property P". For example, the set E of even integers can be expressed as

$$E = \{a \in \mathbb{Z} \mid a \text{ is even }\}.$$

Given two sets A and B, we can form the **union** of A and B

$$A \cup B = \{x \mid x \in A \text{ or } x \in B\}.$$

We should clarify that "$x \in A$ or $x \in B$" allows the possibility that x is in *both* A and B (i.e. it is required to be in at least one of the sets, but could be in both). The **intersection** of A and B is defined by

$$A \cap B = \{x \mid x \in A \text{ and } x \in B\}.$$

If $A \cap B = \emptyset$, then we say that A and B are **disjoint**.

The **difference** of A and B is defined to be

$$A - B = \{x \mid x \in A \text{ and } x \notin B\}.$$

In words, this is the set of all elements which are in A but not in B. For example, $\mathbb{Z} - \mathbb{N}$ is the set of *negative integers*. We note that we are not assuming $B \subset A$ to define $A - B$. For example, if E is the set of even integers, then $E - \mathbb{N}$ is the set of negative even integers.

We often want to work with more than two sets at once, taking their union or intersection. For three sets, we can just name them A, B, C, for example, and similarly for four sets, or five, or six... At some point we run out of letters, or it becomes to unwieldy, or impossible (if there are infinitely many sets). In this case, we will often describe a collection of sets $\mathcal{F} = \{A_j\}_{j \in J}$ in terms of an **index set** J. The index set J is an auxiliary set we use for the purpose of keeping track of the sets A_j in our collection: for every $j \in J$, there is exactly one set A_j in our collection $\mathcal{F}$. When the index set is the natural numbers, then this takes the slightly more familiar form of a **sequence** of sets $\{A_j\}_{j \in \mathbb{N}}$, sometimes written $\{A_j\}_{j=0}^{\infty}$. For example, for every $j \in \mathbb{N}$, we can consider the set $X_j \subset \mathbb{N}$ consisting of all natural numbers from 0 to j. Then $\{X_j\}_{j=0}^{\infty}$ is the collection of all sets X_j, for all $j \in \mathbb{N}$.

Now, given a collection of sets $\{A_j\}_{j \in J}$ indexed by a set J, we define the **union** and **intersection** of the collection of sets, respectively, just as in the case of two sets:

$$\bigcup_{\alpha \in J} A_\alpha = \{x \mid x \in A_\alpha \text{ for some } \alpha \in J\} \quad \text{and} \quad \bigcap_{\alpha \in J} A_\alpha = \{x \mid x \in A_\alpha \text{ for all } \alpha \in J\}.$$

It sometimes happens that our collection of sets $\mathcal{F}$ does not have a specified index set. In this case, we can still define the union and intersection of the sets in $\mathcal{F}$, and we write this as

$$\bigcup_{A \in \mathcal{F}} A = \{x \mid x \in A \text{ for some } A \in \mathcal{F}\} \quad \text{and} \quad \bigcap_{A \in \mathcal{F}} A = \{x \mid x \in A \text{ for all } A \in \mathcal{F}\}.$$

There is really no difference with the previous example, as one could view the set of sets $\mathcal{F}$ as the index set.

Another way of forming new sets from old is the **Cartesian product**. For two sets A, B, the **Cartesian product** (or **product**) of A and B is defined as the set of ordered pairs of elements in A and B:

$$A \times B = \{(x, y) \mid x \in A \text{ and } y \in B\}.$$

When $A = B$, $A \times A$ is also written as A^2. In the case that $A = \mathbb{R}$, then $A^2 = \mathbb{R}^2$ is the familiar Cartesian coordinate plane consisting of ordered pairs of real numbers.

We may also construct the Cartesian product of three or more sets as well. If $\{A_j\}_{j=1}^n$ is a collection of sets indexed by the set $\{1, \ldots, n\}$ for $n \in \mathbb{Z}_+$, then the product of these sets is the set of n–tuples of elements

$$A_1 \times A_2 \times \ldots \times A_n = \{(a_1, a_2, \ldots, a_n) \mid a_j \in A_j \text{ for } j \in \{1, \ldots, n\}\}.$$

Again, if $A_j = A$ for all j, then $A_1 \times \ldots \times A_n = A^n$. When $A = \mathbb{R}$, then $A^n = \mathbb{R}^n$, the usual n–dimensional space. We can similarly define a product of a collection of sets indexed by any set J, but will not bother with this here.

1.1.2 Logic

Given statements P and Q, a sentence of the form "If P, then Q" will have its usual (logical) mathematical meaning, which is: "If P is true, then Q is true, and if P is false, then Q could be either true or false (i.e. we no nothing about the validity of Q)." The statement P is called the **hypothesis** and the statement Q is called the **conclusion**. Shorthand for "If P, then Q" is $P \Rightarrow Q$, which we also read as "P implies Q".

Example 1.1.1. The following sentence about integers x is logically true:

$$\textit{If } x \neq 0, \textit{ then } x^2 > 0.$$

This is true because whenever an integer x is nonzero, its square is necessarily positive. Another statement about integers which is logically true is the following:

$$\textit{If } x^2 < 0, \textit{ then } x = 2.$$

This is true because the hypothesis, $x^2 < 0$ never happens for an integer x. In this case, we sometimes say that the sentence is **vacuously true**. This provides us with a logical justification for the claim $\emptyset \subset A$ for any set A. Indeed, $\emptyset \subset A$ means "If $x \in \emptyset$, then $x \in A$", which is vacuously true, since $x \in \emptyset$ never happens.

If a statement $P \Rightarrow Q$ is true, then the **contrapositive**, $(\textit{not } Q) \Rightarrow (\textit{not } P)$, is also true. Here, whereas $P \Rightarrow Q$ means "if P is true, then Q is true", the statement $(\textit{not } Q) \Rightarrow (\textit{not } P)$ means "if Q is false, then P is false". We can likewise form the **converse** of $P \Rightarrow Q$, which is $Q \Rightarrow P$. Unlike the contrapositive, the validity of the converse is independent of that of the original statement. That is, it could be that $P \Rightarrow Q$ is true, but $Q \Rightarrow P$ is false. When both statements $P \Rightarrow Q$ and $Q \Rightarrow P$ are true, then we say that $P \Leftrightarrow Q$ is true. In this case, we also say that P and Q are **equivalent** statements.

Quite often, instead of saying "$P \Rightarrow Q$ is true", we will simply say "$P \Rightarrow Q$". Similarly, instead of "$P \Leftrightarrow Q$ is true", we say "$P \Leftrightarrow Q$" or "P iff Q" which is shorthand for "P if and only if Q". We also use the logical quantifier "for all", written $\forall$, and "there exists", written $\exists$.

The kinds of statements we will typically prove have the form $P \Rightarrow Q$ and $P \Leftrightarrow Q$. To prove the latter, it suffices to prove $P \Rightarrow Q$ and $Q \Rightarrow P$. Proving $P \Rightarrow Q$, often involves directly deducing that the validity of statement P implies the validity of the statement Q.

Two other important types of proofs that you should be able to understand and construct are **proof by contradiction** and **proof by mathematical induction**. We will see examples throughout this chapter.

1.1.3 Functions

Given two sets A and B, a **function** (also called a **map** or **mapping**) with **domain** A and **range** B is a rule σ associating to every element $a \in A$, an element $\sigma(a) \in B$. We typically write $\sigma\colon A \to B$ or $A \xrightarrow{\sigma} B$ as shorthand. We call $\sigma(a)$ the **image** of a, and for any subset $C \subset A$, the **image** of C is the subset of B defined by

$$\sigma(C) = \{\sigma(a) \mid a \in C\} \subset B.$$

If $D \subset B$, then the **preimage** of D, denoted $\sigma^{-1}(D)$, is the subset

$$\sigma^{-1}(D) = \{a \in A \mid \sigma(a) \in D\}.$$

If $D = \{b\} \subset B$, we also write $f^{-1}(b) = f^{-1}(\{b\})$. In this case, preimage $f^{-1}(b)$ is also called the **fiber** over b. Two functions $\sigma, \tau\colon A \to B$ are equal if $\sigma(a) = \tau(a)$ for all $a \in A$.

If you are wondering what a "rule" is, then here is the more precise definition. A function $\sigma\colon A \to B$ is determined by a subset $\mathcal{G}_\sigma \subset A \times B$ with the property that for every $a \in A$, there exists a unique element $\sigma(b) \in B$ so that $(a, \sigma(b)) \in \mathcal{G}_\sigma$. Of course $\mathcal{G}_\sigma$ is precisely what we call the **graph** of σ. It is a convenient tool for making precise the notion of a function in terms of sets and logic.

Given three sets A, B, C and functions $\sigma\colon A \to B$ and $\tau\colon B \to C$, we can form the **composition**

$$\tau \circ \sigma\colon A \to C$$

by the equation $\tau \circ \sigma(a) = \tau(\sigma(a))$ for all $a \in A$.

Remark 1.1.2. We are following the convention you likely learned in calculus in which we compose functions *from right to left*. In other books on abstract algebra, functions are sometimes composed *from left to right*. Both choices have advantages and disadvantages. We have chosen the convention that deviates least from what you are likely familiar with.

Proposition 1.1.3. † *For functions $\sigma\colon A \to B$, $\tau\colon B \to C$, and $\rho\colon C \to D$, we have*

$$\rho \circ (\tau \circ \sigma) = (\rho \circ \tau) \circ \sigma.$$

Proof. Let $a \in A$ be any element. We need only show that the two functions on either side of the equality in the proposition have the same value on a. We compute from the definition:

$$\rho \circ (\tau \circ \sigma)(a) = \rho((\tau \circ \sigma)(a)) = \rho(\tau(\sigma(a))).$$

On the other hand

$$(\rho \circ \tau) \circ \sigma(a) = (\rho \circ \tau)(\sigma(a)) = \rho(\tau(\sigma(a))).$$

Since these are equal, the proposition is proved. □

Now suppose $\sigma\colon A \to B$ is a function. We say that σ is **injective** (or **one-to-one**) if for all $a, b \in A$ with $\sigma(a) = \sigma(b)$, then $a = b$. The function σ is **surjective** (or **onto**) if for all $b \in B$, there exists $a \in A$ so that $\sigma(a) = b$. That is, the image of the domain is the entire range: $\sigma(A) = B$. Finally, σ is **bijective** (or we say σ is a **bijection**) if it is both injective and surjective.

Example 1.1.4. For any set A, the identity function $\mathrm{id}_A\colon A \to A$ is defined by $\mathrm{id}_A(a) = a$ for all $a \in A$. This is a bijection. Furthermore, for any functions $\sigma\colon A \to B$ and $\tau\colon B \to A$ we have $\sigma \circ \mathrm{id}_A = \sigma$ and $\mathrm{id}_A \circ \tau = \tau$. When the set A is understood from the context, we also write $\mathrm{id} = \mathrm{id}_A$.

Example 1.1.5. Consider the squaring function $\sigma\colon \mathbb{Z}_+ \to \mathbb{Z}_+$; that is $\sigma(x) = x^2$ for every positive integer $x \in \mathbb{Z}_+$. If $\sigma(x) = \sigma(y)$, then since x and y are both positive, we know that $x = y$. Thus, σ is injective. On the other hand, $2 \neq \sigma(x)$ for any $x \in \mathbb{Z}_+$, so σ is not surjective.

Example 1.1.6. The function $\tau\colon \mathbb{Z} \to \mathbb{Z}$ given by $\tau(x) = x + 1$, for every integer $x \in \mathbb{Z}$, is a bijection. To prove this, we first suppose $\tau(x) = \tau(y)$. Then $x + 1 = \tau(x) = \tau(y) = y + 1$, and hence $x = y$, so τ is injective. For any $y \in \mathbb{Z}$, $y - 1$ is also an integer, and furthermore, $\tau(y - 1) = (y - 1) + 1 = y$, so $y \in \tau(\mathbb{Z})$. Since y was an arbitrary integer, it follows that $\tau(\mathbb{Z}) = \mathbb{Z}$, and so τ is surjective, and hence bijective.

Lemma 1.1.7.† *Suppose $\sigma\colon A \to B$ and $\tau\colon B \to C$ are functions. If σ and τ are injective, so is $\tau \circ \sigma$. If σ and τ are surjective, so is $\tau \circ \sigma$. Consequently, the composition of bijections is a bijection.*

Proof. Suppose first that σ and τ are injective, and let $a, b \in A$ be such that $\tau \circ \sigma(a) = \tau \circ \sigma(b)$. Then $\tau(\sigma(a)) = \tau(\sigma(b))$, and by injectivity of τ, we see that $\sigma(a) = \sigma(b)$. But then by injectivity of σ, it follows that $a = b$, and hence $\tau \circ \sigma$ is injective. Next, suppose σ and τ are surjective and let $c \in C$ be any element. Surjectivity of τ implies that there exists $b \in B$ so that $\tau(b) = c$, while surjectivity of σ provides an element $a \in A$ with $\sigma(a) = b$. Then $\tau \circ \sigma(a) = \tau(\sigma(a)) = \tau(b) = c$. Therefore, $\tau \circ \sigma$ is surjective. □

Proposition 1.1.8. *A function $\sigma\colon A \to B$ is a bijection if and only if there exists a function $\tau\colon B \to A$ so that $\tau \circ \sigma = \mathrm{id}_A$ and $\sigma \circ \tau = \mathrm{id}_B$. In this case, the function τ, called the* **inverse** *of σ, is unique and is denoted $\tau = \sigma^{-1}$.*

Given that a function is a bijection if and only if it admits an inverse, bijections are also sometimes called **invertible** functions. Also note that from the proposition, if σ is a bijection, so is σ^{-1}.

Proof. First, suppose that σ is a bijection. Then for any $b \in B$, surjectivity implies that there exists $a \in A$ so that $\sigma(a) = b$. Moreover, because σ is injective, there is only one such a, and we define $\tau(b) = a$.

Now for any $a \in A$, let $b = \sigma(a)$. By definition, $\tau(b) = a$ and thus

$$\tau \circ \sigma(a) = \tau(\sigma(a)) = \tau(b) = a = \mathrm{id}_A(a).$$

On the other hand, for any $b \in B$, if we let $a = \tau(b)$ so again by definition of τ, $\sigma(a) = b$, then

$$\sigma \circ \tau(b) = \sigma(\tau(b)) = \sigma(a) = b = \mathrm{id}_B(b).$$

Therefore, $\tau \circ \sigma = \mathrm{id}_A$ and $\sigma \circ \tau = \mathrm{id}_B$, as required.

Next suppose that there exists $\tau\colon B \to A$ so that $\tau \circ \sigma = \mathrm{id}_A$ and $\sigma \circ \tau = \mathrm{id}_B$. For all $b \in B$, observe that $\tau(b) \in A$ and $\sigma(\tau(b)) = \mathrm{id}_B(b) = b$, and hence σ is surjective. Furthermore, if $a_1, a_2 \in A$ and $\sigma(a_1) = \sigma(a_2)$, then

$$a_1 = \mathrm{id}_A(a_1) = \tau(\sigma(a_1)) = \tau(\sigma(a_2)) = \mathrm{id}_A(a_2) = a_2.$$

So, σ is also injective, hence a bijection.

Finally, supposing there are two functions $\tau_1, \tau_2\colon B \to A$ so that $\tau_1 \circ \sigma = \mathrm{id}_A = \tau_2 \circ \sigma$ and $\sigma \circ \tau_1 = \mathrm{id}_B = \sigma \circ \tau_2$, then by Proposition 1.1.3, we have

$$\tau_1 = \tau_1 \circ \mathrm{id}_B = \tau_1 \circ (\sigma \circ \tau_2) = (\tau_1 \circ \sigma) \circ \tau_2 = \mathrm{id}_A \circ \tau_2 = \tau_2.$$

This completes the proof. □

The **cardinality** (or **size**) of a set A describes "how many elements are in A", and will be denoted $|A|$. Specifically, two sets A and B have the same cardinality if there is a bijection $\sigma\colon A \to B$, and in this case we write $|A| = |B|$. If there is an injection $\sigma\colon A \to B$, then we write $|A| \leq |B|$ (or $|B| \geq |A|$), and if furthermore there is no bijection $A \to B$, then we write $|A| < |B|$ (also $|B| > |A|$). The Cantor–Bernstein–Schroeder Theorem from set theory states that $|A| \leq |B|$ and $|B| \leq |A|$ implies $|A| = |B|$. The following familiar fact is also quite useful.

Pigeonhole Principle. *If A and B are sets and $|A| > |B|$, then there is no injective function $\sigma\colon A \to B$.*

Recall that a set A is **finite**, written $|A| < \infty$, if every injective function $\sigma\colon A \to A$ is a surjection (and hence a bijection). Equivalently, a set A is finite if for any *proper* subset $B \subsetneq A$, we have $|B| < |A|$. The

pigeonhole principle for finite sets is fairly straightforward (see Exercise 1.1.4). The natural numbers are precisely the cardinalities of finite sets, with $0 = |\emptyset|$ and $n \in \mathbb{Z}_+$ being the cardinality $n = |\{1, 2, \ldots, n\}|$, for example. A set which is not finite is said to be **infinite**. We will be primarily interested only in the cardinalities of finite sets, and the distinction between finite sets and infinite sets.

Given sets A, B, we can consider the set of all functions from A to B, denoted B^A:

$$B^A = \{\sigma \colon A \to B \mid \sigma \text{ is a function } \}.$$

The notation is explained by the next example.

Example 1.1.9. Consider the case that $A = \{1, \ldots, n\}$ is the set of natural numbers between 1 and $n \in \mathbb{N}$, and B is any set. Then we define a function $f \colon B^{\{1,\ldots,n\}} \to B^n = B \times \ldots \times B$ by the equation $f(\sigma) = (\sigma(1), \ldots, \sigma(n))$. Exercise 1.1.5 asks you to show that this is a bijection.

Example 1.1.10. If $A = B = \mathbb{R}$, then the continuous functions form a subset of $\mathbb{R}^{\mathbb{R}}$:

$$C(\mathbb{R}, \mathbb{R}) = \{f \colon \mathbb{R} \to \mathbb{R} \mid f \text{ is continuous } \} \subset \mathbb{R}^{\mathbb{R}}.$$

Remark 1.1.11. When the range of a function is a set of functions, things can get a little confusing. We often adopt a special notation for functions in this setting. Specifically, if A, B, C are sets and $\sigma \colon A \to B^C$ is a function, we will often write σ_a for the value of $a \in A$, instead of writing $\sigma(a)$. Since σ_a is a function $\sigma_a \colon C \to B$, we can evaluate it on an element $c \in C$, and when we do so we write $\sigma_a(c)$.

Example 1.1.12. Define $L \colon \mathbb{Z} \to \mathbb{Z}^{\mathbb{Z}}$ by $L_a(x) = ax$. That is, L is the function that assigns to the integer a the function $L_a \colon \mathbb{Z} \to \mathbb{Z}$ which is multiplication by a.

1.1.4 Operations

Given a set A, a **binary operation** on A (also simply called an **operation**) is a function $* \colon A \times A \to A$. We denote the image of (a, b) by $a * b$ (as opposed to $*(a, b)$). The operation is **associative** if $a*(b*c) = (a*b)*c$, for all $a, b, c \in A$, and it is **commutative** if $a * b = b * a$ for all $a, b \in A$.

Example 1.1.13. If X is any of the sets $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, or $\mathbb{R}$, then addition $+ \colon X \times X \to X$ defines an operation $a + b$, for all $a, b \in X$, which is both commutative and associative. Subtraction is also an operation for $X = \mathbb{Z}$, $\mathbb{Q}$, or $\mathbb{R}$, but not $\mathbb{N}$. It is *neither* commutative nor associative since, for example,

$$3 - (2 - 1) = 3 - 1 = 2 \text{ while } (3 - 2) - 1 = 1 - 1 = 0 \neq 2.$$

Multiplication is a commutative, associative operation $\cdot \colon X \times X \to X$ for $X = \mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, or $\mathbb{R}$.

Example 1.1.14. From Proposition 1.1.3, for any set A, composition defines an associative binary operation on the set of all functions from A to itself:

$$\circ \colon A^A \times A^A \to A^A.$$

If A has at least three distinct elements $a, b, c \in A$ (so $a \neq b \neq c \neq a$), then composition is not commutative. To see this, define $\sigma(a) = b$, $\sigma(b) = a$, and $\sigma(d) = d$ for all $d \neq a, b$, and likewise $\tau(a) = c$, $\tau(c) = a$, and $\tau(d) = d$ for all $d \neq a, c$. Note that σ and τ are bijections. Exercise 1.1.6 asks you to show that $\sigma \circ \tau \neq \tau \circ \sigma$.

Suppose we have an associative operation $* \colon A \times A \to A$, and three elements $a, b, c \in A$. Then we can unambiguously write $a * b * c$, and whether we read this as $(a * b) * c$ or $a * (b * c)$, the result is the same. More generally, for any integer $n \geq 2$ and elements $a_1, \ldots, a_n \in A$, we can unambiguously defined $a_1 * a_2 * \cdots * a_n$ by inserting parentheses in any way we like, for example $(a_1 * (a_2 * \cdots * (a_{n-1} * a_n)) \cdots)$, so that we can evaluate on one pair of elements of A at a time. By the associativity of $*$, the result is independent of the choice of parentheses. A formal proof of this can be given via mathematical induction, but we will not bother with that here.

Exercises.

Exercise 1.1.1. Given functions $\sigma\colon A \to B$ and $\tau\colon B \to C$, prove that if $\tau \circ \sigma$ is injective, then so is σ.

Exercise 1.1.2. Given functions $\sigma\colon A \to B$ and $\tau\colon B \to C$, prove that if $\tau \circ \sigma$ is surjective, then so is τ.

Exercise 1.1.3. Prove that the two definitions of finiteness for a set A are equivalent. That is, prove that every injective function $\sigma\colon A \to A$ is surjective if and only if every proper subset $B \subsetneq A$ has $|B| < |A|$.

Exercise 1.1.4. Prove the Pigeonhole Principle for finite sets A and B.

Exercise 1.1.5. Suppose B is any set and $n \in \mathbb{Z}_+$. Prove that $f\colon B^{\{1,\dots,n\}} \to B^n$ defined by $f(\sigma) = (\sigma(1), \dots, \sigma(n))$ is a bijection.

Exercise 1.1.6. Prove that the functions σ and τ in Example 1.1.14 do not commute, $\sigma \circ \tau \neq \tau \circ \sigma$, and hence composition is not commutative for sets A with $|A| > 3$.

You should...

- Be able to do all the exercises from this section and know the proofs of statements marked by †.
- Know basic properties of functions: injective, surjective, bijective, and decide if a given function has these properties.
- Know what an operation is, what it means to be associative and commutative and decide if an operation has these properties.
- Be able to write proofs.

1.2 Equivalence relations

Given a set X, a **relation** on X is a subset of $R \subset X \times X$. We write $x \sim y$ to mean that $(x, y) \in R$, and refer to the relation as $\sim$ (dropping the reference to R altogether). A relation is said to be...

...**reflexive** if for all $x \in X$, we have $x \sim x$.

...**symmetric** if for all $x, y \in X$, if $x \sim y$, then $y \sim x$.

...**transitive** if for all $x, y, z \in X$, if $x \sim y$ and $y \sim z$, then $x \sim z$.

A relation $\sim$ is called an **equivalence relation** if it is reflexive, symmetric, and transitive.

Example 1.2.1. Equality, "=", is an equivalence relation on any set X. Another equivalence relation $\sim$ that makes sense on any set X is to declare $x \sim y$ for every $x, y \in X$. In Exercise 1.2.1 you are asked to show that this is indeed an equivalence relation. These are sometime called the **trivial** equivalence relations.

Example 1.2.2. Suppose U is the set of all students at the University of Illinois. Define a relation $\sim$ on U by declaring $a \sim b$ if $a, b \in U$ were the same age on January 1, 2016. We claim that this is an equivalence relation on the set U of students at the University of Illinois. First, note that every student a is the same age as himself (or herself), so $a \sim a$. Second, observe that if $a \sim b$, then a is the same age as b, but then b is the same age as a, and hence $b \sim a$. Finally, if $a \sim b$ and $b \sim c$, then a is the same age as b and b is the same age as c, so a is the same age as c, and hence $a \sim c$.

Example 1.2.3. Suppose again that U is the set of students at the University of Illinois. Define a different relation $\sim$ by declaring $a \sim b$ if the difference in ages on January 1, 2016 between a and b is at most 1 year. Reasoning in a similar fashion to the previous example shows that $\sim$ is reflexive and symmetric. However, it is not transitive since (surely) there are people $a, b, c \in U$ so that a about 10 months older than b and b is about 10 months older than c, so that $a \sim b$ and $b \sim c$. However, a is about 20 months older than c, and so $a \not\sim c$.

Example 1.2.4. Let $X = \{(a, b) \in \mathbb{Z}^2 \mid b \neq 0\}$. That is, X is the set of ordered pairs of integers, such that the second integer is nonzero. For $(a, b), (c, d) \in X$, define $(a, b) \sim (c, d)$ if $ad = bc$. We claim that $\sim$ is an equivalence relation.

To prove the claim, first observe that for any $(a, b) \in X$, since $ab = ba$, we have $(a, b) \sim (a, b)$, so $\sim$ is reflexive. Next, supposing $(a, b) \sim (c, d)$, we have $ad = bc$, but then $cb = da$ as well, and so $(c, d) \sim (a, b)$, and $\sim$ is symmetric. Finally, suppose $(a, b) \sim (c, d)$ and $(c, d) \sim (e, f)$. Then $ad = bc$ and $cf = de$. Multiplying the first equation by f and the second equation by b we have

$$fad = fbc \quad \text{and} \quad bcf = bde$$

but then $fad = fbc = bcf = bde = bed$. Because $d \neq 0$, this implies $fa = be$, or equivalently, $af = be$, so $(a, b) \sim (e, f)$, and $\sim$ is transitive.

Example 1.2.5. Define a relation on $\mathbb{Z}$ by declaring $x \sim y$ if $x - y$ is even. We claim that $\sim$ is an equivalence relation. First, observe that $x - x = 0$, which is even, and hence $x \sim x$, so $\sim$ is reflexive. Next, $x - y = -(y - x)$, and so $x - y$ is even if and only if $y - x$ is even, so $x \sim y$ implies $y \sim x$. Therefore, $\sim$ is symmetric. Finally, if $x \sim y$ and $y \sim z$, then $x - y$ and $y - z$ are both even. Since the difference of even integers is even, this implies $(x - y) - (y - z) = x - z$ is even, and hence $x \sim z$, so $\sim$ is transitive.

This relation is often written as $x \equiv y \bmod 2$. In §1.5 we will see a more general version of this example.

Example 1.2.6. Suppose $f\colon X \to Y$ is a function. Define a relation $\sim_f$ on X by $x \sim_f y$ if $f(x) = f(y)$. In Exercise 1.2.3 you are asked to prove that this defines an equivalence relation on X. Note that Example 1.2.2 is a special case of this where the function in question is $age\colon U \to \mathbb{N}$ which assigns to each student $a \in U$ their age on January 1, 2016.

A **partition** of a set X is a collections Ω of subsets of X such that the following holds:

1. For all $A, B \in \Omega$, either $A = B$ or $A \cap B = \emptyset$, and
2. $\bigcup_{A \in \Omega} A = X$.

The first condition is sometimes expressed as saying that the sets in Ω are **pairwise disjoint**. Thus a partition is a collection of pairwise disjoint subsets of X whose union is all of X.

A simple, but very important fact is:

♣ **Partitions and equivalence relations are essentially the same thing.**

To make this precise, we need a little more notation. For a set X with an equivalence relation and an element $x \in X$, define the **equivalence class** of x to be the subset $[x] \subset X$ defined by

$$[x] = \{y \in X \mid y \sim x\}.$$

Since $\sim$ is reflexive, $x \in [x]$ for every $x \in X$. We say that any $y \in [x]$ is a **representative** of the equivalence class $[x]$ (in particular, x is a representative of $[x]$).

We write $X/\!\sim$ to denote the set of equivalence classes:

$$X/\!\sim = \{[x] \mid x \in X\}.$$

The next theorem clarifies the meaning of the statement ♣.

Theorem 1.2.7. † *Suppose X is a nonempty set and $\sim$ is an equivalence relation on X. Then the set of equivalence classes $X/\!\sim$ is a partition of X. Conversely, given a partition Ω of X, there exists a unique equivalence relation $\sim_\Omega$ on X so that $X/\!\sim_\Omega = \Omega$.*

Proof. We begin by proving that $X/\sim$ is a partition. For this, assuming that $[x] \cap [y] \neq \emptyset$, and we must show that $[x] = [y]$. Let $z \in [x] \cap [y]$ and let $w \in [x]$. So, $z \sim x$, $z \sim y$, and $w \sim x$. By symmetry, $x \sim z$ and then by transitivity, $w \sim z$. Applying transitivity again, we see $w \sim y$, and hence $w \in [y]$. Therefore, $[x] \subset [y]$. Reversing the roles of $[x]$ and $[y]$ in this proof, we see that $[y] \subset [x]$, and hence $[x] = [y]$. This proves the first requirement for a partition. The second follows from the fact that for all $x \in X$, $x \in [x]$, because then for all $x \in X$, we have

$$x \in [x] \subset \bigcup_{[y] \in X/\sim} [y].$$

Therefore, $X/\sim$ is partition of X.

Now suppose that Ω is a partition of X, and define $x \sim_\Omega y$ if there exists $A \in \Omega$ so that $x, y \in A$. That is, $x \sim_\Omega y$ if x and y are in the same subset of the partition. By the second condition of a partition, for every $x \in X$, there exists $A \in \Omega$ with $x \in A$. But then x and x are obviously both in A, so $x \sim_\Omega x$, and hence $\sim_\Omega$ is reflexive. Next, if $x \sim_\Omega y$, let $A \in \Omega$ be such that $x, y \in A \in \Omega$. This also means that $y \sim_\Omega x$, so $\sim_\Omega$ is symmetric. Finally, if $x \sim_\Omega y$ and $y \sim_\Omega z$, then let $A, B \in \Omega$ be such that $x, y \in A$ and $y, z \in B$. Since $y \in A \cap B$, we see that $A \cap B \neq \emptyset$, and hence $A = B$. But then $z \in A$, ensuring that $x \sim_\Omega z$, and thus $\sim_\Omega$ is transitive. Therefore, $\sim_\Omega$ is an equivalence relation. Note if $x \in A$, then A consists of all the elements of X that are equivalent to x (by definition of $\sim_\Omega$), and hence $A = [x]$. It follows that the equivalence classes of $\sim_\Omega$ are precisely the subsets in Ω.

For the uniqueness, suppose that $\sim$ is any equivalence relation and that $X/\sim = \Omega$. We must see that $x \sim y$ if and only if $x \sim_\Omega y$. Since $X/\sim = \Omega$, for every $x \in X$, there exists $A \in \Omega$ so that $[x] = A$. Then for any $y \in X$, we have $x \sim y$ if and only if $y \in [x]$, which happens if and only if $y \in A$, which by definition of $\sim_\Omega$ happens if and only if $x \sim_\Omega y$. This proves uniqueness, and hence completes the proof of the theorem. □

Example 1.2.8. For the trivial equivalence relations of Example 1.2.1 the equivalence classes are very simple. For the equivalence relation "=" we have $[x] = \{x\}$, the *singleton set* containing only x. For the equivalence relation with $x \sim y$ for all $x, y \in X$, we have just one equivalence class $[x] = X$, for all $x \in X$

Example 1.2.9. For the equivalence relation on the set of students U at the University of Illinois in Example 1.2.2, the equivalence class of $a \in U$ is the set of students who were the same age as a on January 1, 2016.

Example 1.2.10. There are exactly two subsets in the partition for the equivalence relation in Example 1.2.5, namely the even integers and the odd integers.

Example 1.2.11. For the equivalence relation $\sim_f$ defined by a function $f\colon X \to Y$, the second part of Exercise 1.2.3 shows that the equivalence classes are precisely the fibers $[x] = f^{-1}(f(x))$.

It turns out that the construction of an equivalence relation from a function $f\colon X \to Y$ is not special at all.

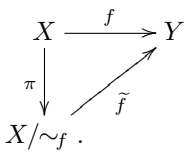
Proposition 1.2.12. *If $\sim$ is an equivalence relation on X, then the function $\pi\colon X \to X/\sim$ defined by $\pi(x) = [x]$ is a surjective map, and the equivalence relation $\sim_\pi$ determined by π is precisely $\sim$.*

Proof. Exercise 1.2.4. □

We prove one more technical fact which will be quite useful later in constructing maps.

Proposition 1.2.13.† *Given any function $f\colon X \to Y$ there exists a unique function $\widetilde{f}\colon X/\sim_f \to Y$ such that $\widetilde{f} \circ \pi = f$, where $\pi\colon X \to X/\sim_f$ is as in Proposition 1.2.12. Furthermore, $\widetilde{f}$ is a bijection onto the image $f(X)$.*

We can visualize the relation between the maps via the diagram of maps



Proof. We would like to define $\widetilde{f}([x])$ on the $\sim_f$–equivalence class $[x]$ by the expression

$$\widetilde{f}([x]) = f(x).$$

In order for this to be *well-defined* we must check that if $[x] = [y]$, then $f(x) = f(y)$. That is, we are attempting to define $\widetilde{f}$ in terms of a *choice of representative* of the equivalence class and we must show that the value is independent of this choice.

Observe that $[x] = [y]$ if and only if $x \sim_f y$, which happens if and only if $f(x) = f(y)$. Consequently $\widetilde{f}$ is well-defined. Since $\pi(x) = [x]$ for all $x \in X$ we have $\widetilde{f} \circ \pi(x) = \widetilde{f}(\pi(x)) = \widetilde{f}([x]) = f(x)$, as required.

To prove uniqueness, we note that if $\widetilde{f}'$ is any other map with $\widetilde{f}'(\pi(x)) = f(x)$, then since $\pi(x) = [x]$, this becomes $\widetilde{f}'([x]) = f(x)$. Since this is precisely the definition of $\widetilde{f}$, it follows that $\widetilde{f}' = \widetilde{f}$, and so $\widetilde{f}$ is unique.

Finally, we must show that $\widetilde{f}$ is a bijection onto the image $f(X)$. The fact that $\widetilde{f}$ maps onto $f(X)$, i.e. $\widetilde{f}(X/\sim_f) = f(X)$, is clear from the property that $\widetilde{f} \circ \pi = f$. To see that $\widetilde{f}$ is injective, we suppose $\widetilde{f}([x]) = \widetilde{f}([y])$. By definition of $\widetilde{f}$, we have $f(x) = \widetilde{f}([x]) = \widetilde{f}([y]) = f(y)$. But then $x \sim_f y$, and so $[x] = [y]$. □

We often refer to the map $\widetilde{f}\colon X/\sim_f$ as the **descent** to $X/\sim_f$ of the map f. The proof of the next fact is nearly identical to the proof of Proposition 1.2.13, and we leave the proof as an exercise.

Proposition 1.2.14. *Suppose $f\colon X \to Y$ is a function and $\sim$ is an equivalence relation on X such that if $x, y \in X$ and $x \sim y$, then $f(x) = f(y)$. Then there exists a unique function $\widetilde{f}\colon X/\sim \to Y$ such that $\widetilde{f} \circ \pi = f$.*

Proof. Exercise 1.2.5. □

Example 1.2.15. We return to Example 1.2.4. Recall that $X = \{(a, b) \in \mathbb{Z}^2 \mid b \neq 0\}$ and $(a, b) \sim (c, d)$ if $ad = bc$. In fact, the set of equivalence classes $X/\sim$ in this example is really the rational numbers, in disguise. More precisely, let $f\colon X \to \mathbb{Q}$ be the function defined by $f(a, b) = \frac{a}{b}$, and we claim that f descends to a bijection $\widetilde{f}\colon X/\sim \to \mathbb{Q}$. For this, we observe that

$$(a, b) \sim_f (c, d) \Leftrightarrow f(a, b) = f(c, d) \Leftrightarrow \frac{a}{b} = \frac{c}{d} \Leftrightarrow ad = bc \Leftrightarrow (a, b) \sim (c, d),$$

and so $\sim$ and $\sim_f$ are the same equivalence relations. Since f is surjective, Proposition 1.2.4 implies that f descends to a bijection $\widetilde{f}\colon X/\sim_f = X/\sim \to \mathbb{Q}$.

In fact, this example is really a construction of the rational numbers from the integers (and hence the natural numbers) as promised earlier. Indeed, if we denote the equivalence class of (a, b) by the symbol $\frac{a}{b}$ instead of $[(a, b)]$, then the set of equivalence classes is precisely the set of symbols $\frac{a}{b}$ where $a, b \in \mathbb{Z}$ and $b \neq 0$, where $\frac{a}{b} = \frac{c}{d}$ if and only if $ad = bc$ (which we often write in fractional form by obtaining a common denominator: $\frac{ad}{bd} = \frac{bc}{bd}$). The rational numbers are more than a set, and more work is required to define the operations of addition and multiplication, but this can be done in terms of the operations on $\mathbb{Z}$. This will be done later in greater generality, so we do not carry out the proof here.

Exercises.

Exercise 1.2.1. Let X be any set and define $x \sim y$ for every $x, y \in X$. Prove that $\sim$ is an equivalence relation.

Exercise 1.2.2. Decide which of the following are equivalence relations on the set of natural numbers $\mathbb{N}$. For those that are, prove it. For those that are not, explain why.

1. $x \sim y$ if $|x - y| \leq 3$.

2. $x \sim y$ if $|x - y| \geq 3$

3. $x \sim y$ if x and y have the same digit in the 1's place (expressed base 10).

4. $x \sim y$ if $x \geq y$.

Exercise 1.2.3. Prove that for any $f\colon X \to Y$, the relation $\sim_f$ defined in Example 1.2.6 is an equivalence relation. Show that for any $x \in X$, the equivalence class of x is precisely $[x] = f^{-1}(f(x))$.

Exercise 1.2.4. Prove Proposition 1.2.12.

Exercise 1.2.5. Prove Proposition 1.2.14.

You should...

- Be able to do all the exercises from this section and know the proofs of statements marked by †.
- Know what an equivalence relation is and be able to decide when a relation is or is not an equivalence relation. This includes justifying your answer (compare Exercise 1.2.2).
- Be able to construct maps using Propositions 1.2.13 and 1.2.14.

1.3 Permutations

Given any set X, we let $\mathcal{S}ym(X) \subset X^X$ denote the set of bijections from X to itself

$$\mathcal{S}ym(X) = \{\sigma\colon X \to X \mid \sigma \text{ is a bijection } \},$$

and call it the **symmetric group of** X or the **permutation group of** X. We call the elements in $\mathcal{S}ym(X)$ the **permutations of** X or the **symmetries of** X. Before we explain where this name comes from, we record the following basic facts about $\mathcal{S}ym(X)$.

Proposition 1.3.1. *For any nonempty set X, $\circ$ is an operation on $\mathcal{S}ym(X)$ with the following properties:*

(i) $\circ$ *is associative.*

(ii) $\mathrm{id}_X \in \mathcal{S}ym(X)$, *and for all* $\sigma \in \mathcal{S}ym(X)$, $\mathrm{id}_X \circ \sigma = \sigma \circ \mathrm{id}_X = \sigma$, *and*

(iii) *For all* $\sigma \in \mathcal{S}ym(X)$, $\sigma^{-1} \in \mathcal{S}ym(X)$ *and* $\sigma \circ \sigma^{-1} = \sigma^{-1} \circ \sigma = \mathrm{id}_X$.

Proof. According to Lemma 1.1.7, if σ, τ are bijections, then so is $\sigma \circ \tau$, and thus $\circ$ defines an operation on $\mathcal{S}ym(X)$. Associativity follows from Proposition 1.1.3, so (i) holds. Part (ii) comes from the definition of id_X (see Example 1.1.4). For part (iii), given $\sigma \in \mathcal{S}ym(X)$, according to Proposition 1.1.8 it suffices to show σ^{-1} has an inverse, but one easily checks that $(\sigma^{-1})^{-1} = \sigma$, so $\sigma^{-1} \in \mathcal{S}ym(X)$, and has the required properties for (iii) by definition and Proposition 1.1.8. □

Suppose X is finite, consisting of a collection of identical objects, say green marbles or something, arranged in a line. We can permute the marbles by picking them up and rearranging them into a (possibly) different order. The permutation can be recorded by a bijection $\sigma\colon X \to X$ defined as follows. Once-and-for-all we label the initial *locations* of the marbles, using the elements of X to label them, so that the marble x is initially in the location labeled x. Now, for any permutation of the marbles, we let $\sigma\colon X \to X$ be defined by $\sigma(x) = y$ if the marble in location labeled x is sent to location y.

Suppose Sam walks by the marbles and permutes them according to some bijection $\sigma\colon X \to X$ and an hour later, Taylor walks by and permutes them according to some bijection $\tau\colon X \to X$. How were the marbles permuted from before Sam arrived until after Taylor left (assuming no one else permuted them)?

Well, Sam took the marble in location x and set it down in location $y = \sigma(x)$, then Taylor picked it up from there, and moved it to location $\tau(y) = \tau(\sigma(x)) = \tau \circ \sigma(x)$. Therefore, the marbles were permuted according to the composition of the bijections $\tau \circ \sigma$, which is again a bijection by Lemma 1.1.7. If the marbles really are identical, then every permutation can be thought of as a "symmetry" of the arrangement. Whether you think of them as permutations of the marbles or symmetries of the arrangements, each determines, and is determined by, an element of $\mathcal{S}ym(X)$.

Given $\sigma \in \mathcal{S}ym(X)$, we let

$$\sigma^2 = \sigma \circ \sigma, \quad \sigma^3 = \sigma \circ \sigma^2 = \sigma \circ \sigma \circ \sigma, \quad \sigma^4 = \sigma \circ \sigma^3, \quad \text{etc.}$$

More precisely, we can define σ^r, for every $r \geq 1$ **recursively** by the formula $\sigma^1 = \sigma$, and $\sigma^r = \sigma \circ \sigma^{r-1}$ for every integer $r > 1$. Since σ is a bijection, $\sigma^{-1} \in \mathcal{S}ym(X)$ (see Proposition 1.1.8), and thus for $r > 0$, we can also write

$$\sigma^{-r} = (\sigma^{-1})^r.$$

Setting $\sigma^0 = \mathrm{id}_X \in \mathcal{S}ym(X)$, we have defined σ^r for every $r \in \mathbb{Z}$, and it is easy to check (though it requires multiple cases) that for all $r, s \in \mathbb{Z}$,

$$\sigma^r \circ \sigma^s = \sigma^{r+s}.$$

(We will prove a more general fact later, so do not prove this here.)

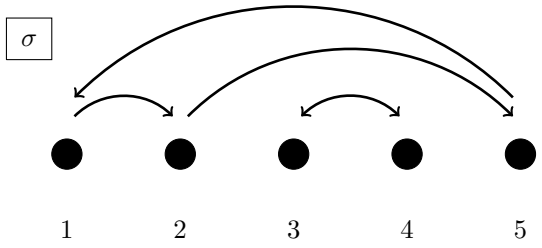
A special case of interest is when $X = \{1, \ldots, n\}$. Then $\mathcal{S}ym(X)$ is often denoted S_n, and is called the **symmetric group on n elements** or the **permutation group on n elements**. Any element $\sigma \in S_n$ is determined by its values $\sigma(1), \ldots, \sigma(n)$, so one way to denote σ is with a $2 \times n$ matrix:

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

Example 1.3.2. Suppose $\sigma \in S_5$ has $\sigma(1) = 2$, $\sigma(2) = 5$, $\sigma(3) = 4$, $\sigma(4) = 3$, $\sigma(5) = 1$. This can be represented as

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 5 & 4 & 3 & 1 \end{pmatrix}.$$

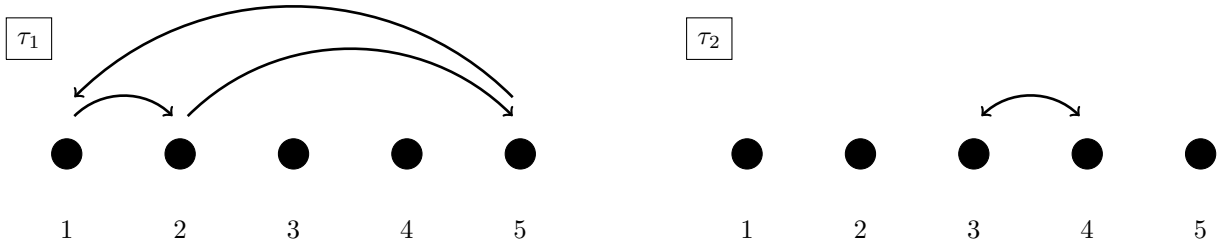
If we imagine five marbles arranged in a line, then this permutation can be represented pictorially as shown below.



From the "picture" of σ, we observe that σ is a composition $\sigma = \tau_1 \circ \tau_2$, where

$$\tau_1 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 5 & 3 & 4 & 1 \end{pmatrix} \quad \text{and} \quad \tau_2 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 4 & 3 & 5 \end{pmatrix}$$

which we can visualize as in the next picture.



Given $k \geq 2$, a k–**cycle** in S_n is a permutation σ with the property that $\{1, \ldots, n\}$ is the union of two disjoint subsets, $\{1, \ldots, n\} = Y \cup Z$ and $Y \cap Z = \emptyset$, such that

1. $\sigma(x) = x$ for every $x \in Z$, and
2. $|Y| = k$, and for any $x \in Y$, $Y = \{\sigma(x), \sigma^2(x), \sigma^3(x), \ldots, \sigma^{k-1}(x), \sigma^k(x) = x\}$.

We say that σ **cyclically permutes** the elements of Y and **fixes** the elements of Z. In the example above, τ_1 is a 3–cycle which cyclically permutes $\{1, 2, 5\}$, sending 1 to 2, 2 to 5, and 5 back to 1 and fixes $\{3, 4\}$. The permutation τ_2 is a 2–cycle that cyclically permutes $\{3, 4\}$ (interchanging the two) and fixes $\{1, 2, 5\}$. We use **cycle notation** to denote these cycles, writing

$$\tau_1 = (1\ 2\ 5) = (2\ 5\ 1) = (5\ 1\ 2) \qquad \text{and} \qquad \tau_2 = (3\ 4) = (4\ 3).$$

As is suggested, the choice of 1 as the starting point is arbitrary, and so there are 3 ways to denote τ_1 and two ways to denote τ_2. Each number is sent by the cycle to the next in order, and once we arrive at the final parenthesis, we return back to the first number.

More generally, if τ is a k–cycle and $Y = \{x, \sigma(x), \sigma^2(x), \ldots, \sigma^{k-1}(x)\}$ is the set cyclically permuted by τ, we denote this

$$(x\ \sigma(x)\ \sigma^2(x)\ \sigma^3(x)\ \ \ldots\ \ \sigma^{k-1}(x)).$$

In this notation, the numbers $x, \sigma(x), \sigma^2(x), \ldots, \sigma^{k-1}(x)$ are listed in the order in which they are cyclically permuted. Thus σ sends x to $\sigma(x)$, it sends $\sigma(x)$ to $\sigma^2(x)$, etc., and finally it sends $\sigma^{k-1}(x)$ to $\sigma^k(x) = x$. As above, the final parenthesis means that σ returns the element back to the beginning. The choice of x as the initial element is arbitrary, and in particular, there are k different ways to denote any k–cycle.

Example 1.3.3. Consider $\tau \in S_7$ given by

$$\tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 3 & 2 & 7 & 5 & 1 & 6 & 4 \end{pmatrix}.$$

This is a 5–cycle, which we can express in cycle notation in 5 ways:

$$\tau = (1\ 3\ 7\ 4\ 5) = (3\ 7\ 4\ 5\ 1) = (7\ 4\ 5\ 1\ 3) = (4\ 5\ 1\ 3\ 7) = (5\ 1\ 3\ 7\ 4).$$

Given a k–cycle $\sigma \in S_n$ and a k'–cycle $\sigma' \in S_n$, we say that σ and σ' are **disjoint cycles** if the sets cyclically permuted by σ and σ' are disjoint. Note that τ_1 and τ_2 are disjoint cycles in the example above.

Lemma 1.3.4. † *If $\sigma, \sigma' \in S_n$ are disjoint cycles, then σ and σ' commute. If x is cyclically permuted by σ and x' is cyclically permuted by σ', then $\sigma \circ \sigma'(x) = \sigma' \circ \sigma(x) = \sigma(x)$ and $\sigma \circ \sigma'(x') = \sigma' \circ \sigma(x') = \sigma'(x')$.*

Proof. Let $x \in \{1, \ldots, n\}$. If x is in the set cyclically permuted by σ, then so is $\sigma(x)$, and x and $\sigma(x)$ are fixed by σ' by definition of disjoint cycles. Thus

$$\sigma' \circ \sigma(x) = \sigma'(\sigma(x)) = \sigma(x) = \sigma(\sigma'(x)) = \sigma \circ \sigma'(x).$$

Similarly, if x' is in the set cyclically permuted by σ', it is fixed by σ and a similar computation shows

$$\sigma \circ \sigma'(x') = \sigma'(x') = \sigma' \circ \sigma(x').$$

Finally, if y is fixed by both σ and σ', then

$$\sigma \circ \sigma'(y) = \sigma(\sigma'(y)) = \sigma(y) = y \text{ and } \sigma' \circ \sigma(y) = \sigma'(\sigma(y)) = \sigma'(y) = y.$$

So, for all $z \in \{1, \ldots, n\}$, $\sigma \circ \sigma'(z) = \sigma' \circ \sigma(z)$, and so $\sigma \circ \sigma' = \sigma' \circ \sigma$. □

In Example 1.3.2, we have $\sigma = \tau_1 \circ \tau_2$ where $\tau_1 = (1\ 2\ 5)$ and $\tau_2 = (3,\ 4)$ are disjoint cycles. We express this as

$$\sigma = (1\ 2\ 5)(3\ 4).$$

Note that we have dropped the "$\circ$", which we will often do with composition. We could have equally well written $\sigma = (3\ 4)(5\ 1\ 2)$: since τ_1 and τ_2 commute the order is irrelevant, and we have multiple ways to denote each cycle. We call this expression for σ its **disjoint cycle notation**. It turns out that every permutation can be expressed in disjoint cycle notation.

Proposition 1.3.5. *Given $\sigma \in S_n$, there exists a unique (possibly empty) set of pairwise disjoint cycles $\tau_1, \ldots, \tau_k \in S_n$, so that $\sigma = \tau_1 \circ \cdots \circ \tau_k$.*

By Lemma 1.3.4, we can take the composition of $\tau_1, \ldots, \tau_k$ in the proposition in any order. The situation that the set of disjoint cycles is empty only occurs when σ is the identity. In this case, we view the composition of the empty set of disjoint cycles as the identity, by convention.

Example 1.3.6. The proof of Proposition 1.3.5 is somewhat involved, though the idea is quite simple. Suppose $\sigma \in S_9$ is given by

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 3 & 5 & 1 & 4 & 8 & 2 & 9 & 6 & 7 \end{pmatrix}.$$

To find the disjoint cycle representation of σ, we pick an element in $\{1, \ldots, 9\}$ to start with (usually 1), and "follow" its images as we iteratively apply the mapping σ:

$$1 \mapsto \sigma(1) = 3 \mapsto \sigma(3) = 1.$$

So, σ cyclically permutes $\{1, 3\}$, and we record the 2–cycle $(1\ 3)$. We repeat for the remaining numbers. Moving on to 2, we see

$$2 \mapsto \sigma(2) = 5 \mapsto \sigma(5) = 8 \mapsto \sigma(8) = 6 \mapsto \sigma(6) = 2.$$

So σ permutes $\{2, 5, 8, 6\}$ the same way the 4–cycle $(2\ 5\ 8\ 6)$ does. The next number not yet appearing is 4, which is fixed. After that, we come to 7, and we follow it

$$7 \mapsto \sigma(7) = 9 \mapsto \sigma(9) = 7.$$

This permutes $\{7, 9\}$ as the 2–cycle $(7\ 9)$. Every number has now appeared in our list. We simply observe that the cycles we listed are pairwise disjoint, and σ is a composition of these disjoint cycles:

$$\sigma = (1\ 3)(2\ 5\ 8\ 6)(7\ 9).$$

We could have equally well include "(4)", which we can think of informally as a "1–cycle" (i.e. the identity):

$$\sigma = (1\ 3)(2\ 5\ 8\ 6)(4)(7\ 9),$$

and as mentioned, the order the cycles appear in doesn't matter since they commute, so

$$\sigma = (9\ 7)(4)(8\ 6\ 2\ 5)(1\ 3).$$

Proof of Proposition 1.3.5. We define a relation on the set $X_n = \{1, \ldots, n\}$, declaring $x \sim y$ if there exists $r \in \mathbb{Z}$ so that $\sigma^r(x) = y$. We first prove that this is in fact an equivalence relation:

- Since $\sigma^0(x) = x$, we have $x \sim x$ for all $x \in X_n$.
- If $\sigma^r(x) = y$, then $\sigma^{-r}(y) = x$, and hence $x \sim y$ implies $y \sim x$.
- If $x \sim y$ and $y \sim z$, let $r, s \in \mathbb{Z}$ be such that $\sigma^r(x) = y$ and $\sigma^s(y) = z$. Then $\sigma^{r+s}(x) = \sigma^s(\sigma^r(x)) = \sigma^s(y) = z$, so $x \sim z$.

Note that $[x] = \{\sigma^s(x) \mid s \in \mathbb{Z}\}$. We wish to give a more precise description $[x]$.

First observe that for any $x \in X_n$, $\sigma^r(x) = x$ for some $r > 0$. To see this, observe that because $|X_n| = n < \infty$, the set $\{\sigma^r(x) \mid r > 0\} \subset X_n$ is also finite, and hence by the Pigeonhole Principle, for some $0 < t < s$ we must have $\sigma^t(x) = \sigma^s(x)$ (the function from $\mathbb{Z}_+ \to \{\sigma^r(x) \mid r > 0\}$ given by $r \mapsto \sigma^r(x)$ cannot be injective). But then setting $r = s - t > 0$, we have

$$\sigma^r(x) = \sigma^{s-t}(x) = \sigma^{-t}(\sigma^s(x)) = x.$$

Let $r_x > 0$ be the smallest positive integer such that $\sigma^{r_x}(x) = x$. From the calculation just given, we observe that if $0 < t < s$ and $\sigma^t(x) = \sigma^s(x)$, then $s - t \geq r_x$. That is, $\{\sigma(x), \ldots, \sigma^{r_x}(x)\}$ consists of r_x distinct elements.

We claim that

$$[x] = \{\sigma(x), \ldots, \sigma^{r_x}(x)\}.$$

To see this, note that $[x] = \{\sigma^s(x) \mid s \in \mathbb{Z}\}$, and hence $\{\sigma(x), \ldots, \sigma^{r_x}\} \subset [x]$. So, suppose $\sigma^s(x) \in [x]$, and we prove that $\sigma^s(x) = \sigma^t(x)$ for some $0 < t \leq r_x$. For this, let $m \in \mathbb{Z}$ be such that $mr_x < s \leq (m+1)r_x$ and set $t = s - mr_x$. Then $0 < t \leq r_x$ and since $\sigma^{r_x}(x) = x$, we have $\sigma^{-mr_x}(x) = x$ so that

$$\sigma^t(x) = \sigma^{s-mr_x}(x) = \sigma^s \circ \sigma^{-mr_x}(x) = \sigma^s(\sigma^{-mr_x}(x)) = \sigma^s(x),$$

as required.

Choose representatives of the equivalence classes $x_1, \ldots, x_k, x_{k+1}, \ldots, x_m \in X_n$, so that $|[x_j]| > 1$ if and only if $1 \leq j \leq k$. Observe that σ fixes x_j if and only if $k+1 \leq j \leq m$. Furthermore, any element x fixed by σ has $[x] = \{x\}$, and so is one of the representatives $x = x_j$ for some $k+1 \leq j \leq m$.

For each $1 \leq j \leq k$ let $\tau_j \in S_n$ be defined by

$$\tau_j(x) = \begin{cases} \sigma(x) & \text{for } x \in [x_j] \\ x & \text{otherwise} \end{cases}$$

So, τ_j cyclically permutes the elements in $[x_j]$ in the same way σ does, and fixes every other element of X_n. Since the equivalence classes form a partition, $\tau_1, \ldots, \tau_k$ are disjoint cycles. From the definition and Lemma 1.3.4 we have

$$\sigma(x) = \tau_1 \circ \cdots \circ \tau_k(x).$$

To prove the uniqueness, suppose $\sigma = \tau'_1 \circ \cdots \circ \tau'_{k'}$, a composition of disjoint cycles. Then from Lemma 1.3.4, either $\sigma(x) = x$ or else there exists a unique $1 \leq i \leq k'$ so that $\sigma(x) = \tau'_i(x)$. In the latter case, σ and τ'_i agree on the equivalence class of x, and hence both agree with some τ_j on that equivalence class. But τ'_i and τ_j will fix every element of X not in that equivalence class, and therefore $\tau'_i = \tau_j$. Since this is true for every $1 \leq i \leq k'$, this proves the uniqueness. □

Computing composition in terms of disjoint cycle notation is straightforward. We describe this with an example.

Example 1.3.7. Suppose $\sigma = (1\ 3)(2\ 5\ 8\ 6)(7\ 9)$ as above and $\rho = (1\ 4\ 5)(3\ 6\ 7\ 9)$ are elements in S_9. We compute $\sigma \circ \rho$ and $\rho \circ \sigma$, and express these in disjoint cycle notation. For $\sigma \circ \rho$, we look where 1 is sent when we iterate this permutation. We have

$$1 \mapsto \sigma(\rho(1)) = \sigma(4) = 4 \mapsto \sigma(\rho(4)) = \sigma(5) = 8 \mapsto \sigma(\rho(8)) = \sigma(8) = 6$$
$$6 \mapsto \sigma(\rho(6)) = \sigma(7) = 9 \mapsto \sigma(\rho(9)) \mapsto \sigma(3) = 1,$$

giving the cycle (1 4 8 6 9). Moving on to 2, we have

$$2 \mapsto \sigma(\rho(2)) = \sigma(2) = 5 \mapsto \sigma(\rho(5)) \mapsto \sigma(1) = 3 \mapsto \sigma(\rho(3)) = \sigma(6) = 2,$$

which gives the cycle (2 5 3). The only number remaining is 7, which should be fixed, since its the only number in $\{1, \ldots, 9\}$ not accounted for. We verify this:

$$7 \mapsto \sigma(\rho(7)) = \sigma(9) = 7.$$

Thus, we have

$$\sigma \circ \rho = (1\ 3)(2\ 5\ 8\ 6)(7\ 9)(1\ 4\ 5)(3\ 6\ 7\ 9) = (1\ 4\ 8\ 6\ 9)(2\ 5\ 3).$$

Looking at this, we can see another way to compute the composition: Starting at the right and working your way to the left, look for the first cycle containing 1, we see it is sent to 4. Then we continue to the left, looking for a cycle containing 4, and see where it is sent. There isn't another one, so we deduce that 1 is sent to 4. To see where 4 goes, we start over from the right, moving our way to the left, looking for a cycle containing 4. The first one we come to sends 4 to 5. Continuing to the left, we come upon a cycle telling us that 5 should be sent to 8. Moving on further to the left, 8 does not appear and so we see that 4 is sent to 8. This tells us what $\sigma \circ \rho$ does to any number (if a number never appears in any cycle, it is fixed), and so we can compute the composition this way. Try it for $\rho \circ \sigma$, then check your answer against ours here

$$\rho \circ \sigma = (1\ 4\ 5)(3\ 6\ 7\ 9)(1\ 3)(2\ 5\ 8\ 6)(7\ 9) = (1\ 6\ 2)(3\ 4\ 5\ 8\ 7).$$

We note that this procedure gives us a way of computing disjoint cycle notation for any permutation expressed as a composition of (not necessarily disjoint) cycles.

From the explanation in this example for computing composition of permutations in disjoint cycle notation, we see that the disjoint cycle notation for the inverse of a permutation given in disjoint cycle notation is obtained by "reversing all the cycles".

Example 1.3.8. Let $\sigma = (1\ 7\ 4\ 5\ 3)(8\ 6\ 9)$. This is the disjoint cycle notation for σ, so we get σ^{-1} by reversing each of the cycles:

$$\sigma^{-1} = (3\ 5\ 4\ 7\ 1)(9\ 6\ 8).$$

Proposition 1.3.9. *Given $n \geq 2$, any $\sigma \in S_n$ can be expressed as a composition of 2–cycles.*

Proof. We claim that for every $k \geq 2$, any k–cycle can be expressed as a product of 2–cycles. By Proposition 1.3.5, every permutation can be expressed as a composition of disjoint cycles, so proving this claim will suffice to prove the proposition. The special case that we have a composition of an empty set of disjoint cycles means we have the identity, which we can write, for example, as (1 2)(1 2). We prove the claim by mathematical induction.

The base case for the induction is $k = 2$. A 2–cycle is a composition of a single 2–cycle (itself!), so the claim holds in this case. Now suppose that for some $k \geq 2$, any k–cycle is a composition of 2–cycles, and we prove that a $(k+1)$–cycle is a composition of 2–cycles. Let $\sigma = (x_1\ x_2\ \ldots\ x_k\ x_{k+1})$ be a $(k+1)$–cycle in S_n. Then composing with the 2–cycle $(x_1\ x_{k+1})$ we get

$$(x_1\ x_{k+1})\sigma = (x_1\ x_{k+1})(x_1\ x_2\ \ldots\ x_k\ x_{k+1}) = (x_1\ x_2\ \ldots\ x_k).$$

So, $\tau = (x_1\ x_{k+1})\sigma$ is a k–cycle, and by induction, this can be expressed as a composition of 2–cycles. But since $\sigma = (x_1\ x_{k+1})(x_1\ x_{k+1})\sigma = (x_1\ x_{k+1})\tau$, we see that σ can also be expressed as the composition of 2–cycles. By the principle of mathematical induction, the claim holds for every integer $k \geq 2$, as required. This completes the proof. □

Exercises.

Exercise 1.3.1. Let $\sigma \in S_8$ be the permutation given by the 2×8 matrix

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 4 & 1 & 3 & 2 & 7 & 6 & 8 & 5 \end{pmatrix}.$$

Express σ, σ^2, σ^3, and σ^{-1} in disjoint cycle notation.

Exercise 1.3.2. consider $\sigma = (3\ 4\ 8)(5\ 7\ 6\ 9)$ and $\tau = (1\ 9\ 3\ 5)(2\ 7\ 4)$ in S_9 expressed in disjoint cycle notation. Compute $\sigma \circ \tau$ and $\tau\sigma$ expressing both in disjoint cycle notation.

Exercise 1.3.3. Prove that for any $k \geq 2$, a k–cycle can be expressed as a composition of exactly $k-1$ 2–cycles. Hint: Look at the proof of Proposition 1.3.9.

You should...

- Be able to do all the exercises from this section and know the proofs of statements marked by †.
- Know what $\mathcal{S}ym(X)$ and S_n are. Be able to represent an element of S_n in disjoint cycle notation. Be able to compute the composition of permutations given in disjoint cycle notation.
- Be able to write a proof by induction.

1.4 The integers.

We will assume the reader has a basic understanding of arithmetic in $\mathbb{Z}$. We will assume familiarity with addition $a+b$, subtraction $a-b$, multiplication $a \cdot b = ab$, absolute value $|a|$, and the meaning of the comparisons $a \leq b$, $a \geq b$, $a < b$, and $a > b$, for $a, b \in \mathbb{Z}$. While we have already mentioned some of the following properties in passing, we begin with a systematic study of $\mathbb{Z}$ here and so list them formally.

Proposition 1.4.1. *The following hold in the integers $\mathbb{Z}$:*

(i) *Addition and multiplication are commutative and associative operations in $\mathbb{Z}$.*

(ii) *$0 \in \mathbb{Z}$ is an identity element for addition; that is, for all $a \in \mathbb{Z}$, $0 + a = a$.*

(iii) *Every $a \in \mathbb{Z}$ has an additive inverse, denoted $-a$ and given by $-a = (-1)a$, satisfying $a + (-a) = 0$.*

(iv) *$1 \in \mathbb{Z}$ is an identity element for multiplication; that is, for all $a \in \mathbb{Z}$, $1a = a$.*

(v) *The distributive law holds: for all $a, b, c \in \mathbb{Z}$, $a(b+c) = ab + ac$.*

(vi) *Both $\mathbb{N} = \{x \in \mathbb{Z} \mid x \geq 0\}$ and $\mathbb{Z}_+ = \{x \in \mathbb{Z} \mid x > 0\}$ are closed under addition and multiplication. That is, if x and y are in one of these sets, then $x + y$ and xy are also in that set.*

(vii) *For any two nonzero integers $a, b \in \mathbb{Z}$, $|ab| \geq \max\{|a|, |b|\}$. Strict inequality holds if $|a| > 1$ and $|b| > 1$.*

Using these properties, we also get **cancellation** for multiplication, meaning that if $ca = ba$ and $a \neq 0$, then $b = c$. To see this, we may apply the properties from above to deduce $0 = ca - ba = (c-b)a$, and so since $a \neq 0$, we must have $c - b = 0$, and hence $c = c + 0 = c - b + b = b$.

Divisibility will play an important role, and we begin our detailed analysis with it. Given integers $a, b \in \mathbb{Z}$, with $b \neq 0$, we say that b **divides** a if there exists $m \in \mathbb{Z}$ so that $a = mb$. If b divides a, we will write $b|a$ (and in writing this, we always assume $b \neq 0$). If b does not divide a, then we write $b \not| a$. If $b|a$, then we say that b is a **divisor** of a or a **factor** of a.

Before proceeding, we make some elementary observations about divisibility.

Proposition 1.4.2. *Let $a, b, c \in \mathbb{Z}$. Then*

(i) *If* $a \neq 0$, *then* $a|0$.

(ii) *If* $a|1$, *then* $a = \pm 1$.

(iii) *If* $a|b$ *and* $b|a$, *then* $a = \pm b$.

(iv) *If* $a|b$ *and* $b|c$, *then* $a|c$.

(v) *If* $a|b$ *and* $a|c$, *then* $a|(mb + nc)$ *for all* $m, n \in \mathbb{Z}$.

Proof. For part (i), we have $a0 = 0$, so $a|0$. To prove part (ii), writing $na = 1$, we note that $n, a \neq 0$ and $1 \geq |a| > 0$ by Proposition 1.4.1. Therefore, $|a| = 1$, and hence $a = \pm 1$. For part (iii) we can find $u, v \in \mathbb{Z}$ so that $b = ua$ and $a = vb$. But then $a = uva$, so $uv = 1$. Again by Proposition 1.4.1 we see that $1 \geq |uv| > 0$, and so $u = v = \pm 1$. Parts (iv) and (v) are left as Exercises 1.4.1 and 1.4.2 □

A positive integer $p > 1$ is said to be **prime** if the only divisors are ± 1 and $\pm p$ (alternatively, we could say that the only positive divisors are 1 and p). For an integer $a \geq 2$, a **prime factorization** of a is an expression of a as a product of powers of primes:

$$a = p_1^{n_1} \cdots p_k^{n_k}$$

where $p_1, \ldots, p_k > 1$ are distinct prime integers and $n_1, \ldots, n_k \in \mathbb{Z}_+$ are positive integers.

Theorem 1.4.3 (Prime Factorization)**.** *Any integer* $a > 1$ *has a prime factorization* $a = p_1^{n_1} \cdots p_k^{n_k}$. *Moreover, this factorization is unique in the sense that if* $q_1^{m_1}, \ldots, q_j^{m_j}$ *is another prime factorization, then* $k = j$, *and after reindexing we have* $p_i = q_i$ *and* $n_i = m_i$, *for each* $i = 1, \ldots, k$.

The uniqueness will require some additional machinery. However, the existence is fairly straightforward, and we prove it first.

Lemma 1.4.4. *Any integer* $a > 1$ *has a prime factorization.*

Proof. The proof is by induction. The base case for the induction is $a = 2$. Since the only positive integer less than 2 is 1, there can be no other positive divisors other than 1 and 2, and hence 2 is prime. Thus $2 = 2^1$ is a prime factorization completing the proof of the base case.

Next, we suppose that for some $a > 2$, any integer $2 \leq b < a$ has a prime factorization, and we prove that a also has a prime factorization. To this end, we note that either a is itself prime, in which case we are done, or else it can be expressed as a product $a = uv$ for some positive integers u, v with $1 < u, v < a$. But then by the inductive hypothesis u and v have prime factorizations $u = p_1^{n_1} \cdots p_k^{n_k}$ and $v = q_1^{b_1} \cdots q_j^{b_j}$. Then

$$a = p_1^{n_1} \cdots p_k^{n_k} q_1^{m_1} \cdots q_j^{m_j}$$

and combining powers of common primes gives a prime factorization of a, as required. □

Remark 1.4.5. The version of "mathematical induction" used in this proof—specifically making the inductive hypothesis include the validity of the claim not just for the positive integer $a - 1$, but all integers from the base case up to $a - 1$—follows formally from the usual version of induction, but as illustrated here, can be more useful in certain situations. We will see this kind of proof again before this section has ended.

The uniqueness part of Theorem 1.4.3 requires a little better understanding of divisibility. Given $a, b \in \mathbb{Z}$, not both zero, the **greatest common divisor** of a and b is the positive integer c such that

1. $c|a$ and $c|b$, and
2. if $d|a$ and $d|b$, then $d|c$.

We say that nonzero integers $a, b \in \mathbb{Z}$ are **relatively prime** if the greatest common divisor is 1.

If the greatest common divisor of a and b exists, it is unique by Proposition 1.4.2 part (iii), and we denote it $\gcd(a, b) = \gcd(b, a)$. Note that by Proposition 1.4.1, if $a \neq 0$ and $\gcd(a, b)$ exists, then $\gcd(a, b) \leq |a|$, and similarly for b.

Remark 1.4.6. It is tempting to want to appeal to the prime factorization of a pair of integers a and b saying that the product of the minimum powers of the common prime factors of a and b is the greatest common divisor. This is obviously true, but this requires the uniqueness part of Theorem 1.4.3, and we have not yet proved that. Indeed, the proof of uniqueness in Theorem 1.4.3 will indirectly *use* the existence of the greatest common divisor.

The proof that greatest common divisors exist will use the **Euclidean algorithm** (which is essentially the result of long division). In fact, we obtain crucial information about the greatest common divisor from the proof. Before we can get to it, we first prove

Proposition 1.4.7 (Euclidean Algorithm)**.** *Given $a, b \in \mathbb{Z}$, $b \neq 0$, there exists unique $q, r \in \mathbb{Z}$ with $0 \leq r < |b|$ so that*

$$a = qb + r.$$

Proof. We prove the existence statement first in two cases, then give the uniqueness statement in general. Before beginning, observe that without loss of generality we may assume that $b > 0$, for if $b < 0$, we can prove the existence for $-b = |b|$, and multiply the resulting q by -1. We now proceed to the proof.

Case 1. $a \geq 0$.

The proof is by induction. The base case actually consists of multiple cases, namely all a with $0 \leq a < b$. When $0 \leq a < b$, we set $q = 0$ and $r = a$. Now, assume that $a \geq b$ and that the conclusion holds *for all* $0 \leq a' < a$, and we prove that it holds for a. Since $a \geq b$, we see that $0 \leq a - b < a$. Therefore, $a - b = q'b + r'$, for some $q', r' \in \mathbb{Z}$ with $0 \leq r' < b$. Then

$$a = q'b + b + r' = q'b + b + r' = (q' + 1)d + r'.$$

Thus, setting $q = q' + 1$ and $r = r'$ proves the statement.

Case 2. $a < 0$.

Since $-a > 0$, we can find $q', r' \in \mathbb{Z}$ with $0 \leq r' < b$, so that $-a = q'b + r'$. This implies $a = -q'b - r'$. If $r' = 0$, then settting $q = -q'$ and $r = 0$ completes the proof. Otherwise, let $q = -(q + 1)$ and $r = b - r'$. Since $0 < r' < b$, we have $b > b - r' = r > 0$, and furthermore

$$a = -q'b - r' = -q'b - b + b - r' = -(q' + 1)b + r = qb + r$$

as required.

Finally, to prove the uniqueness statement, suppose $a = qb + r = q'b + r'$, with $q, q', r, r' \in \mathbb{Z}$, and $0 \leq r, r' < |b|$. If $r = r'$, then $q = q'$ and we are done, so suppose $r \neq r'$. Without loss of generality, assume $r < r'$ (otherwise reverse the roles of r and r'). Then $0 < r' - r < |b|$ and $r' - r = (q - q')b$. By Proposition 1.4.1, $q - q'$ is nonzero, as is b, and so

$$r' - r = (q - q')b = |(q - q')b| \geq \max\{|q - q'|, |b|\} \geq |b|.$$

This contradicts the fact that $r' - r < |b|$. Therefore, we must have $r = r'$ and $q = q'$, which proves uniqueness. □

Proposition 1.4.8. † *For any $a, b \in \mathbb{Z}$, not both zero, $\gcd(a, b)$ exists. Furthermore, $\gcd(a, b)$ is the smallest positive integer of the form $ma + nb$, for $m, n \in \mathbb{Z}$. In particular, there exists $m_0, n_0 \in \mathbb{Z}$ so that $\gcd(a, b) = m_0 a + n_0 b$.*

Proof. Let $M = \{ma + nb \mid m, n \in \mathbb{Z}\}$. For all $x = ma + nb \in M$, we have $-x = (-m)a + (-n)b \in M$, and since both of a and b are nonzero, M contains a positive integer. Let $c = m_0 a + n_0 b$ be the smallest positive integer in M. We claim that $c = \gcd(a, b)$.

First, suppose $d|a$ and $d|b$. Then by Proposition 1.4.2 part (v), $d|x$ for all $x \in M$, and hence $d|c$. Next, let $x = ma + nb \in M$ be any element. By the Euclidean Algorithm (Proposition 1.4.7) there exists $q, r \in \mathbb{Z}$ with $0 \leq r < c$ so that $x = qc + r$, or equivalently

$$r = x - qc = ma + nb - (q(m_0a + n_0b)) = (m - qm_0)a + (n - qn_0)b.$$

Therefore, $r \in M$. Since $r < c$, we must have $r = 0$ (by minimality of c) and hence $x = qc$. Applying this to the cases that x is $a = 1a + 0b$ and $b = 0a + 1b$, it follows that $c|a$ and $c|b$, and hence $c = \gcd(a, b)$. □

Although this is an "existence" proof, and doesn't provide an obvious way to find $\gcd(a, b)$, it does give a hint. From the proof we see that $\gcd(a, b)$ is the unique positive element of M which is a divisor of both a and b. Therefore, we can "search" for divisors of a and b in M, and we can do so using the Euclidean Algorithm.

Suppose $a, b \in \mathbb{Z}$ and $a, b \neq 0$, and write $a = q_0b + r_0$ for $q_0, r_0 \in \mathbb{Z}$, $0 \leq r_0 < |b|$. If $r_0 = 0$, then $b|a$, and so $|b| = \gcd(a, b)$. If not, we can apply the Euclidean Algorithm *again*, this time to b and r_0, to find $q_1, r_1 \in \mathbb{Z}$ with $b = q_1r_0 + r_1$ and $0 \leq r_1 < r_0$. If $r_1 = 0$, then $r_0|b$ and since $a = q_0b + r_0$, we see $r_0|a$. On the other hand, $r_0 \in M$ and so $r_0 = \gcd(a, b)$. If $r_1 \neq 0$, we repeat again with r_0 and r_1, finding q_2, r_2, with $0 \leq r_2 < r_1$, and $r_0 = q_2r_1 + r_2$. If $r_2 = 0$, then r_1 divides r_0 and hence also a and b and substituting into previous equations, we see that $r_1 \in M$, hence $r_1 = \gcd(a, b)$. If $r_2 \neq 0$, we repeat again. The last nonzero remainder in this process is the greatest common divisor and substituting into previous equations finds m_0 and n_0.

Example 1.4.9. Find $\gcd(222, 42)$ and express it in the form $m_0222 + n_042$ for $m_0, n_0 \in \mathbb{Z}$. Iterating the Euclidean algorithm (i.e. long division), we find

$$\begin{array}{rcl} 222 & = & 5(42) + 12 \\ 42 & = & 3(12) + 6 \\ 12 & = & 2(6) \end{array}$$

So, $\gcd(222, 42) = 6$. To find m_0 and n_0, first solve for 6 in the second equation, to get $6 = 42 - 3(12)$. Then we can solve for 12 in the first equation to get $12 = 222 - 5(42)$ and substitute to get

$$6 = 42 - 3(12) = 42 - 3(222 - 5(42)) = 42 - 3(222) + 15(42) = -3(222) + 16(42).$$

So we can set $m_0 = -3$ and $n_0 = 16$.

An easy consequence of Proposition 1.4.8 is the following.

Proposition 1.4.10. † *Suppose $a, b, c \in \mathbb{Z}$. If b and c are relatively prime and $b|ac$, then $b|a$.*

Proof. Since $\gcd(b, c) = 1$, we may find $m, n \in \mathbb{Z}$ so that $1 = mb + nc$. Then $a = mba + nca$, and since $b|ac$ it follows that $b|nca$. Since we also have $b|mba$, Proposition 1.4.2 part (v) again implies $b|a$, as required. □

As an immediate consequence, we have the following.

Corollary 1.4.11. *Suppose $a, b, p \in \mathbb{Z}$ and $p > 1$ is prime. If $p|ab$, then $p|a$ or $p|b$.*

Proof. If $p|b$, we are done, so suppose $p \nmid b$. Since p is prime, $\gcd(b, p) = 1$. By Proposition 1.4.10, $p|a$. □

We are finally ready for the

Proof of Theorem 1.4.3. The existence part of the theorem was already established in Lemma 1.4.4, so it remains only to prove the uniqueness. The proof of this part is also by induction. The base case is $a = 2$, which being prime, is only divisible by 1 and 2 and so can only be factored as a product of the prime 2.

We again suppose that for some $a > 2$, the uniqueness statement is true for all integers $2 \leq b < a$, and prove it for a. Suppose we have two prime factorizations $a = p_1^{n_1} \cdots p_k^{n_k} = q_1^{m_1} \cdots q_j^{m_j}$. Without loss

of generality we assume the indices have been chosen so that $p_1 > \ldots > p_k > 1$ and $q_1 > \ldots > q_j > 1$ (otherwise, reindex so that this is true).

Note that $p_1|a$ and hence by Corollary 1.4.11, p_1 must divide one of the primes q_i for some i, and consequently, $p_1 \leq q_i$. Similarly, q_1 must divide some $p_{i'}$ for some i', and hence $q_1 \leq p_{i'}$. By our index choice we have

$$p_{i'} \leq p_1 \leq q_i \leq q_1 \leq p_{i'}.$$

The only way this can be true is if all the inequalities are equalities, and thus $p_1 = q_1$. Moreover, we must have $n_1 = m_1$, for otherwise, setting $N = \min\{n_1, m_1\}$ and dividing by $p_1^N = q_1^N$, we would have two prime factorizations of the quotient where the largest primes are different (contradicting what we just proved).

Therefore, $n_1 = m_1$, and dividing a by $p_1^{n_1} = q_1^{m_1}$ we get

$$b = \frac{a}{p_1^{n_1}} = p_2^{n_2} \cdots p_k^{n_k} = q_2^{n_2} \cdots q_j^{n_j}.$$

If $b = 1$, then $a = p_1^{n_1} = q_1^{m_1}$ with $p_1 = q_1$ and $n_1 = m_1$, and we are done. Otherwise, $2 \leq b < a$. In this case, note that the equation above gives two prime factorizations of b, and so by the inductive assumption, $k = j$, and further by our choice of indices, for each $i = 2, \ldots, k$, $p_i = q_i$ and $n_i = m_i$. Since we already proved this for $i = 1$ as well, this completes the proof of the inductive step. By induction we are done. □

Exercises.

Exercise 1.4.1. Prove that if $a, b, c \in \mathbb{Z}$, $a|b$, and $b|c$, then $a|c$.

Exercise 1.4.2. Prove that if $a, b, c, m, n \in \mathbb{Z}$, $a|b$, and $a|c$, then $a|(mb + nc)$.

Exercise 1.4.3. For each of the pairs $(a, b) = (130, 95), (1295, 406), (1351, 165)$, find $\gcd(a, b)$ using the Euclidean Algorithm and express it in the form $\gcd(a, b) = m_0 a + n_0 b$ for $m_0, n_0 \in \mathbb{Z}$.

Exercise 1.4.4. Suppose $a, b, c \in \mathbb{Z}$. Prove that if $\gcd(a, b) = 1$, $a|c$, $b|c$, then $ab|c$.

Exercise 1.4.5. The **least common multiple** of two nonzero integers a, b, denoted $\operatorname{lcm}(a, b)$ is the smallest integer positive integer n so that $a|n$ and $b|n$. Prove that $ab = \operatorname{lcm}(a, b) \gcd(a, b)$. Hint: express both in terms of the prime factorizations of a and b.

You should...

- Be able to do all the exercises from this section and know the proofs of statements marked by †.
- Know what the greatest common divisor of two positive integers a, b is, be able to find it using the Euclidean algorithm and express it in the form $ma + nb$.
- Know what the prime factorization of a positive integer is.

1.5 Modular arithmetic.

While the integers form the basis for the construction of the rational numbers, as mentioned at the end of §1.2, (and consequently, also the basis for the real numbers), they can also be used to define other "number systems" of interest. To describe these constructions, we start by defining a relation associated to every positive integer.

Given any $n \in \mathbb{Z}$, $n \geq 1$, say that two integers $a, b \in \mathbb{Z}$ are **congruent modulo** n, if $n|(a - b)$, and in this case we write "$a \equiv b \bmod n$" (read "a is congruent to b mod n"). This defines a relation on $\mathbb{Z}$ called **congruence modulo** n.

Proposition 1.5.1. *For any $n \geq 1$, congruence modulo n is an equivalence relation.*

Proof. We must verify that the relation of congruence modulo n is reflexive, symmetric, and transitive. For this, consider any three integers $a, b, c \in \mathbb{Z}$.

- By Proposition 1.4.2 part (i), $n|0 = a - a$, and thus $a \equiv a \bmod n$, so the relation is reflexive.
- Suppose $a \equiv b \bmod n$. Then $n|(a-b)$ and hence $n|(b-a)$. Therefore, $b \equiv a \bmod n$, proving symmetry.
- Assume $a \equiv b \bmod n$ and $b \equiv c \bmod n$. Then n divides both $a - b$ and $b - c$ and so also the sum, $a - b + b - c = a - c$. Therefore $a \equiv c \bmod n$, which proves that the relation is transitive.

□

Given $a, n \in \mathbb{Z}$, with $n \geq 1$, the equivalence class of a with respect to the equivalence relation of congruence modulo n will be denoted $[a]_n$, or simply $[a]$ if n is understood from the context. We call $[a]$ the **congruence class** of a modulo n. Note that since $b \equiv a \bmod n$ means there is some $k \in \mathbb{Z}$ so that $b - a = nk$, or equivalently, $b = a + nk$, we see that

$$[a] = \{a + nk \mid k \in \mathbb{Z}\}.$$

We let $\mathbb{Z}_n$ denote the set of congruence classes modulo n

$$\mathbb{Z}_n = \{[a]_n \mid a \in \mathbb{Z}\}.$$

Proposition 1.5.2. *For any $n \geq 1$ there are exactly n congruences classes modulo n, which we may write as*

$$\mathbb{Z}_n = \{[0], \ldots, [n-1]\}.$$

Proof. Let $a \in \mathbb{Z}$ and apply Proposition 1.4.7 to write $a = qn + r$ where $q, r \in \mathbb{Z}$ and $0 \leq r < n$. Clearly $a \equiv r \bmod n$, and hence $[a] = [r]$. It follows that

$$\mathbb{Z}_n = \{[0], \ldots, [n-1]\}.$$

On the other hand, if $0 \leq a < b < n$, then $0 < b-a < n$, and hence $n \nmid (b-a)$, so $a \not\equiv b \bmod n$. Consequently, $[a] \neq [b]$, and the n congruence classes listed are all distinct. Thus there are exactly n congruence classes modulo n. □

So, $\mathbb{Z}_n$ is a set with exactly n elements. It turns out that the arithmetic in $\mathbb{Z}$ can be used to define arithmetic in $\mathbb{Z}_n$ (that is, a notion of addition and multiplication). This requires the following lemma.

Lemma 1.5.3. † *Suppose $a, b, c, d, n \in \mathbb{Z}$ with $n \geq 1$, $a \equiv c \bmod n$, and $b \equiv d \bmod n$. Then*

(i) *$a + b \equiv c + d \bmod n$, and*

(ii) *$ab \equiv cd \bmod n$.*

Proof. Since $a - c$ and $b - d$ are divisible by n, so is their sum, $a - c + b - d = (a + b) - (c + d)$. Therefore, $a + b \equiv c + d \bmod n$, proving part (i) of the lemma. For part (ii), we let $k, j \in \mathbb{Z}$ be such that $a = c + nk$ and $b = d + nj$. Then

$$ab = (c + nk)(d + nj) = cd + ndk + ncj + jkn^2 = cd + n(dk + cj + jkn).$$

Consequently, $ab \equiv cd \bmod n$, proving part (ii). □

We now fix $n \geq 1$ and attempt to define the operations of addition and multiplication on $\mathbb{Z}_n$ by the formulas:

$$[a] + [b] = [a + b] \qquad \text{and} \qquad [a][b] = [ab]$$

for all $[a], [b] \in \mathbb{Z}_n$. We are attempting to define $[a] + [b]$ and $[a][b]$ by choosing representatives, namely a and b, of the equivalence classes $[a]$ and $[b]$, respectively. This is partially masked by the notation: writing $[a]$ does not uniquely determine a. For example, $[1] = [7]$ in $\mathbb{Z}_6$. So, to see that addition and multiplication are well-defined, we write $[a] = [c]$ and $[b] = [d]$ and must verify that

$$[a + b] = [c + d] \qquad \text{and} \qquad [ab] = [cd].$$

This is precisely the content of Lemma 1.5.3, however, so these operations are indeed well-defined.

Example 1.5.4. We can write down the entire addition and multiplication table for $\mathbb{Z}_n$, for any n, since this is a finite set. For $\mathbb{Z}_3$, we have

$+$	$[0]$	$[1]$	$[2]$
$[0]$	$[0]$	$[1]$	$[2]$
$[1]$	$[1]$	$[2]$	$[0]$
$[2]$	$[2]$	$[0]$	$[1]$

$\cdot$	$[0]$	$[1]$	$[2]$
$[0]$	$[0]$	$[0]$	$[0]$
$[1]$	$[0]$	$[1]$	$[2]$
$[2]$	$[0]$	$[2]$	$[1]$

We have the following consequence of this definition of addition and multiplication in $\mathbb{Z}_n$ and Proposition 1.4.1.

Proposition 1.5.5. *Let $a, b, c, n \in \mathbb{Z}$ with $n \geq 1$. Then*

(i) *Addition and multiplication are commutative and associative operations in $\mathbb{Z}_n$.*

(ii) $[a] + [0] = [a]$.

(iii) $[-a] + [a] = [0]$.

(iv) $[1][a] = [a]$.

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

Proof. Exercise 1.5.1 □

An element $[a] \in \mathbb{Z}_n$ is said to be **invertible** (with respect to multiplication), or we say that $[a]$ is a **unit** in $\mathbb{Z}_n$, if there exists $[b] \in \mathbb{Z}_n$ so that $[a][b] = [1]$. Observe that if $[b], [b'] \in \mathbb{Z}_n$ and $[a][b] = [1] = [a][b']$, then since multiplication is commutative, we have

$$[b] = [b][1] = [b][a][b'] = [a][b][b'] = [1][b'] = [b'].$$

Thus, $[b] = [b']$. We sometimes write $[b] = [a]^{-1}$ for this unique element with $[a][a]^{-1} = [1]$. The set of invertible elements in $\mathbb{Z}_n$ will be denoted $\mathbb{Z}_n^\times$.

Proposition 1.5.6. *For all $n \geq 1$, we have $\mathbb{Z}_n^\times = \{[a] \in \mathbb{Z}_n \mid \gcd(a, n) = 1\}$.*

Proof. Suppose $\gcd(a, n) = 1$, and let $b, k \in \mathbb{Z}$ be such that $ab + nk = 1$. Then $[a][b] = [ab] = [1]$, proving $\{[z] \in \mathbb{Z}_n \mid \gcd(a, n) = 1\} \subset \mathbb{Z}_n^\times$. To prove the other containment, suppose $[a] \in \mathbb{Z}_n^\times$, and let $b \in \mathbb{Z}$ be such that $[a][b] = [1]$. Then there exists $k \in \mathbb{Z}$ so that $ab - 1 = nk$, or equivalently $ab + (-k)n = 1$. Since 1 is the smallest positive integer, Proposition 1.4.8 guarantees that $\gcd(a, n) = 1$. □

This immediately implies

Corollary 1.5.7. *If $p \geq 2$ is prime, then $\mathbb{Z}_p^\times = \mathbb{Z}_p - \{[0]\}$.*

So for p a prime, there are exactly $p-1$ units in $\mathbb{Z}_p$. Deciding how many units there are in $\mathbb{Z}_n$ is slightly more complicated. To set some notation, we let $\varphi(n) = |\mathbb{Z}_n^\times|$ denote the number of units in $\mathbb{Z}_n$. The function $\varphi(n)$ is called the **Euler phi function**, or the **Euler totient function**, and according to Proposition 1.5.6 it counts the number of positive integers, less than n, which are relatively prime to n. Below (Proposition 1.5.11), we will see that we can compute $\varphi(n)$ in terms of the prime factorization of n. This computation requires the so-called **Chinese Remainder Theorem**, which we now explain.

If $m|n$, we can define $\pi_{m,n}\colon \mathbb{Z}_n \to \mathbb{Z}_m$ by $\pi_{m,n}([a]_n) = [a]_m$. In Exercise 1.5.4 you are asked to prove that this is well-defined.

Theorem 1.5.8 (Chinese Remainder Theorem)**.** *Suppose that $n = mk$ with $m, n, k > 0$, and that m, k are relatively prime. Then the function*

$$F\colon \mathbb{Z}_n \to \mathbb{Z}_m \times \mathbb{Z}_k$$

given by $F([a]_n) = ([a]_m, [a]_k) = (\pi_{m,n}([a]_n), \pi_{k,n}([a]_n))$ is a bijection.

Proof. From Exercise 1.5.4, the function F is well-defined. Suppose first that $F([a]_n) = F([b]_n)$. Then $a \equiv b \bmod m$ and $a \equiv b \bmod k$, and so $a - b$ is divisible by both m and k. According to Exercise 1.4.4, it follows that $a - b$ is divisible by $mk = n$, and hence $a \equiv b \bmod n$. Therefore, $[a]_n = [b]_n$, and F is injective.

To prove that F is also surjective, suppose $u, v \in \mathbb{Z}$ are any two integers. We must find $a \in \mathbb{Z}$ so that $[a]_m = [u]_m$ and $[a]_k = [v]_k$. For then, any element $([u]_m, [v]_k)$ is the F–image of some element (namely $[a]_n$).

By Proposition 1.4.8, there exists $s, t \in \mathbb{Z}$ so that $1 = sm + tk$. Now let $a = u(1 - tk) + v(1 - sm)$, and observe that since $1 - tk = sm$ and $1 - sm = tk$, we have

$$[a]_m = [usm + v - vsm]_m = [v]_m \quad \text{and} \quad [a]_k = [u - utk + vtk]_k = [u]_k,$$

as required. Therefore, F is also surjective, completing the proof. □

Supposing $n = mk$ with $\gcd(m, k) = 1$, let $F\colon \mathbb{Z}_n \to \mathbb{Z}_m \times \mathbb{Z}_k$ be the bijection from Theorem 1.5.8. Let $[a]_n, [b]_n \in \mathbb{Z}_n$, and note that

$$F([a]_n[b]_n) = F([ab]_n) = ([ab]_m, [ab]_k) = ([a]_m[b]_m, [a]_k[b]_k).$$

Because F is a bijection $[ab]_n = [1]_n$ if and only if $[ab]_m = [1]_m$ and $[ab]_k = [1]_k$. From this it follows that $[a]_n$ is a unit in $\mathbb{Z}_n$ (with inverse $[b]_n$) if and only if $[a]_m$ and $[a]_k$ are units in $\mathbb{Z}_m$ and $\mathbb{Z}_k$, respectively (with respective inverse $[b]_m$ and $[b]_k$). This proves

Proposition 1.5.9. *Suppose that $n = mk$ with $m, n, k > 0$, that $\gcd(m, k) = 1$, and $F\colon \mathbb{Z}_n \to \mathbb{Z}_m \times \mathbb{Z}_k$ is the bijection from Theorem 1.5.8. Then $F(\mathbb{Z}_n^\times) = \mathbb{Z}_m^\times \times \mathbb{Z}_k^\times$.*

Corollary 1.5.10. *Suppose that $n = mk$ with $m, n, k > 0$, and $\gcd(m, k) = 1$. Then $\varphi(n) = \varphi(m)\varphi(k)$.*

Proposition 1.5.11. *If $n \in \mathbb{Z}$ is positive integer with prime factorization $n = p_1^{r_1} \cdots p_k^{r_k}$, then*

$$\varphi(n) = (p_1 - 1)p_1^{r_1 - 1} \cdots (p_k - 1)p_k^{r_k - 1}.$$

Proof. From Corollary 1.5.10 it follows that

$$\varphi(n) = \varphi(p_1^{r_1}) \cdots \varphi(p_k^{r_k}).$$

Therefore, it suffices to prove that for a prime p and integer $r > 0$, we have $\varphi(p^r) = (p - 1)p^{r-1}$. There are p^r elements of $\mathbb{Z}_{p^r}$, given by $[a]$ with $0 \le a \le p^r - 1$. By Proposition 1.5.6, $[a]$ is a unit if and only if a relatively prime to p^r. However, a is relatively prime to p^r if and only if it is relatively prime to p, meaning it is *not* a multiple of p. Exactly one out of every p consecutive integers is a multiple of p, and so the number of multiples of p from 0 to $p^r - 1$ is exactly $p^r/p = p^{r-1}$. The rest are relatively prime, and hence

$$\varphi(p^r) = p^r - p^{r-1} = (p - 1)p^{r-1}.$$

□

Exercises.

Exercise 1.5.1. Prove Proposition 1.5.5.

Exercise 1.5.2. Write down the addition and multiplication tables for $\mathbb{Z}_5$.

Exercise 1.5.3. List all elements of $\mathbb{Z}_5^\times$, $\mathbb{Z}_6^\times$, $\mathbb{Z}_8^\times$, and $\mathbb{Z}_{20}^\times$.

Exercise 1.5.4. Prove that if $m|n$, then $\pi_{m,n}\colon \mathbb{Z}_n \to \mathbb{Z}_m$ is well-defined.

Exercise 1.5.5. Compute $\varphi(12), \varphi(15), \varphi(16), \varphi(100)$.

You should...

- Be able to do all the exercises from this section and know the proofs of statements marked by †.
- Be able to carry out simple modular arithmetic calculations.
- Know what the Chinese Remainder Theorem says, and be able to use it.
- Be able to compute $\varphi(n)$ for any integer n.

Chapter 2

Abstract algebra

We begin this chapter by first recalling the arithmetic and geometry of the *complex numbers*, which are constructed from the real numbers. This includes a discussion of the *Fundamental Theorem of Algebra* (Theorem 2.1.1), concerning the existence of roots of polynomials with complex coefficients. With this new addition to our growing library of "number systems", we take our first steps into abstract algebra, defining the notion of a *field* in §2.2. This serves to unify not only the rational, real, and complex numbers, but many others, including $\mathbb{Z}_p$, for p a prime, which we encountered in §1.5.

Next we turn to a more abstract view on polynomials where the coefficients are allowed to be any elements of a field. By considering the set of all polynomials (with coefficients in any fixed field), we see that this set has many striking similarities with the integers, and we will take some time to explore these similarities. Next we turn to another abstract algebraic object defined with respect to a field, namely *vector spaces*. We recall basic definitions and facts from linear algebra—the study of vector spaces and the maps between them. Linear algebra provides multiple ties to abstract algebra: from their very definitions, to connections with fields and polynomials, as well as deeper connections we will see later. We also use linear algebra over the real numbers as a natural language for studying Euclidean geometry.

The final section of this chapter introduces one of the most primitive and fundamental objects in abstract algebra called a *group*. This extracts the key ideas and ingredients from essentially all the topics discussed in the first two chapters, tying them together in a variety of ways. Groups will serve as our primary object of study moving forward, but in this section, we just give the definitions, and list a number of examples.

Note: The level of formalism varies a bit in this chapter. Some topics we will study serve primarily as motivation. Assuming familiarity with these, we do not always develop everything "from the ground up".

2.1 Complex numbers and the Fundamental Theorem of Algebra.

We will not describe the construction of the real numbers—they are formally defined from the rational numbers (which are, in turn, defined from the integers; see §1.2). Instead, as with the integers, we will assume the reader is familiar with basic properties of the rational and real numbers. Specifically, we assume the reader understands containments

$$\mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}$$

and that the associative, commutative operations of addition and multiplication on $\mathbb{Z}$ extend to $\mathbb{Q}$ and $\mathbb{R}$. The number 0 is again the additive identity, and 1 the multiplicative identity:

$$a + 0 = a \text{ and } 1 \cdot a = a.$$

Every real number has an additive inverse, $-a = -1(a)$ and every *nonzero* real number has a multiplicative inverse $a^{-1} = \frac{1}{a}$. From this subtraction and division are defined by $a - b = a + (-b)$ and $a \div b = a \cdot \frac{1}{b}$. The ordering of $\mathbb{R}$, specifically the meaning of the symbols $a < b$, $a > b$, $a \leq b$, and $a \geq b$, are assumed to be understood, as are the usual rules for manipulating equations and inequalities. The ordering allows us to

visualize the real numbers as a line, which we often refer to as the **real line**. We recall that the absolute value of a real number is defined as $|x| = x$ if $x \geq 0$ and $|x| = -x$, and so $|x| \geq 0$ with equality if and only if $x = 0$. Finally, we assume the reader understands the completeness property of the real numbers to the extent that they are unbothered by existence of square roots of positive real numbers, and have working knowledge of exponential and trigonometric functions.

The **complex numbers** are defined, *as a set*, to be formal sums of pairs of real numbers:

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

where i is, for the time being, just a symbol. We also write $a + bi = a + ib$, when convenient. We can visualize the complex numbers as we do $\mathbb{R}^2$, identifying $a + bi \in \mathbb{C}$ with $(a, b) \in \mathbb{R}^2$ (this "identification" is really a bijection between $\mathbb{C}$ and $\mathbb{R}^2$ which we suppress explicit reference to in general). Addition is defined in a way that respects this identification, and thus

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

and it is straightforward to check that it is associative and commutative.

To define multiplication, we now give the symbol i meaning, declaring its square to be -1. That is,

$$i^2 = -1.$$

Equipped with this, we can multiply two complex numbers $z = a+bi$ and $w = c+di$, treating i as a variable and multiplying them like polynomials, but replacing i^2 with -1. Thus we have

$$zw = (a + bi)(c + di) = ac + bci + adi + bdi^2 = ac + bci + adi - bd = (ac - bd) + (bc + ad)i.$$

Observe that $zw = wz$ (since multiplication of real numbers is commutative). In Exercise 2.1.1 you are asked to show that multiplication is also associative.

Let $z = a + bi$. We define the **conjugate** of z by

$$\bar{z} = a - bi,$$

and the **absolute value** of z by

$$|z| = \sqrt{a^2 + b^2}.$$

With our identification of $\mathbb{C}$ with $\mathbb{R}^2$, we see that the absolute value is simply the distance to the origin. The absolute value and conjugation are related by the equation:

$$|z|^2 = a^2 + b^2 = (a + bi)(a - bi) = z\bar{z}.$$

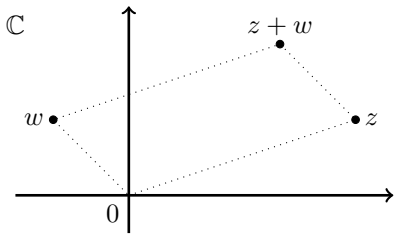
We will consider the real numbers as a subset of the complex numbers, identifying $a \in \mathbb{R}$ with $a + 0i$. With this inclusion, observe that $\mathbb{R}$ is precisely the set of elements of $\mathbb{C}$ *fixed* by conjugation: $\bar{z} = z$ if and only if $z \in \mathbb{R}$. We also note that if $c \in \mathbb{R}$ and $z = a + bi$, then multiplying $c \in \mathbb{R}$ with $z \in \mathbb{C}$ agrees with scalar multiplication (again with our identification of $\mathbb{C}$ and $\mathbb{R}^2$):

$$cz = (ca) + (cb)i = (ca - 0b) + (cb + 0a)i = (c + 0i)(a + bi).$$

The real numbers 0 and 1 serve as additive and multiplicative identities. Every complex number $z = a+bi$ has an additive inverse $-z = (-1)z = -a - bi$. Every nonzero element has a multiplicative inverse, denoted $z^{-1} = \frac{1}{z}$, which from the calculations above is given by

$$\frac{1}{z} = \frac{1}{|z|^2}\bar{z}.$$

It thus makes sense to write $z \div w = \frac{z}{w} = z\frac{1}{w} = \frac{1}{|w|^2}z\bar{w}$. Complementary to the real numbers $\mathbb{R} \subset \mathbb{C}$ we have the **imaginary numbers** $\{bi = 0 + bi \in \mathbb{C} \mid b \in \mathbb{R}\}$.



In this example:
$z = 3 + i$
$w = -1 + i$
$z + w = 2 + 2i$

Since addition in $\mathbb{C}$ is essentially the same as addition in $\mathbb{R}^2$ (via our identification), we understand geometrically that addition is given by the "parallelogram law", as the figure indicates.

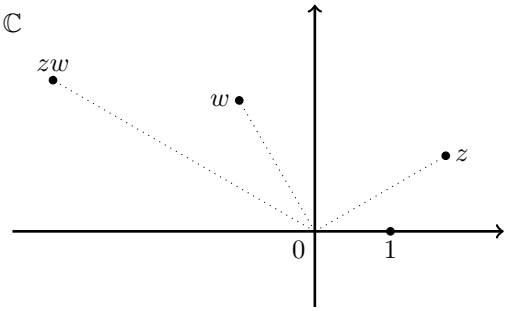
It turns out that multiplication also has a nice geometric interpretation if one uses polar coordinates (and our identification with $\mathbb{R}^2$). Specifically, given two nonzero complex numbers, we may write

$$z = r(\cos(\theta) + i\sin(\theta)) \quad \text{and} \quad w = s(\cos(\psi) + i\sin(\psi)),$$

where $r = |z| > 0$ and $s = |w| > 0$. From the definition of multiplication and standard trigonometric identities, we have

$$\begin{aligned} zw &= sw(\cos(\theta)\cos(\psi) - \sin(\theta)\sin(\psi) + i(\cos(\theta)\sin(\psi) + \cos(\psi)\sin(\theta))) \\ &= sw(\cos(\theta + \psi) + i\sin(\theta + \psi)). \end{aligned}$$

Therefore, when multiplying complex numbers, the absolute value of the product is the product of the absolute values, and the **angle** (also called the **argument**) of the product is the *sum* of the angles.



In this example:
$z = 2\cos(\pi/6) + i2\sin(\pi/6)$
$w = 2\cos(4\pi/6) + i2\sin(4\pi/6)$
$zw = 4\cos(5\pi/6) + i4\cos(5\pi/6)$

Writing $\cos(\theta) + i\sin(\theta)$ is a bit cumbersome, and as shorthand, we write

$$e^{i\theta} = \cos(\theta) + i\sin(\theta).$$

We will not dwell on the rationale for this expression, but the interested reader can formally manipulate power series for e^x, $\cos(x)$, and $\sin(x)$ to arrive at this equation (in fact, this formalism can be justified with a little complex analysis). Then multiplication of $z = re^{i\theta}$ and $w = se^{i\psi}$ takes the simple form

$$zw = rse^{i(\theta+\psi)}$$

showing that these imaginary exponents behave the same as their real counterparts.

One very important property of the complex numbers is given by the following theorem. As this is really a theorem from analysis (or topology), we will not prove it here.

Theorem 2.1.1 (Fundamental Theorem of Algebra)**.** *Suppose $f(x) = a_0 + a_1x + \cdots + a_nx^n$ is a nonconstant polynomial with complex coefficients, $a_0, \ldots, a_n \in \mathbb{C}$. Then $f(x)$ has a* **root** *in $\mathbb{C}$: that is, there exists $\alpha \in \mathbb{C}$ such that $f(\alpha) = 0$.*

Corollary 2.1.2. *Every nonconstant polynomial $f(x) = a_0 + a_1x + \cdots + a_nx^n$ with coefficients $a_0, \ldots, a_n$ in $\mathbb{C}$ can be factored as a product of linear polynomials:*

$$f(x) = a_n \prod_{j=1}^{n} (x - \alpha_j) = a_n(x - \alpha_1)(x - \alpha_2) \cdots (x - \alpha_n),$$

where $\alpha_1, \ldots, \alpha_n$ are roots of $f(x)$.

Proof. The corollary follows from the theorem by "long division" of polynomials: if α is a root of $f(x)$, then there exists a polynomial $q(x)$ of degree $n-1$ so that $f(x) = (x - \alpha)q(x)$, now induct on the degree. See Corollary 2.3.16 below. □

As another consequence of Theorem 2.1.1 we have the following, possibly familiar fact.

Corollary 2.1.3. *If $f(x) = a_0 + a_1x + \cdots + a_nx^n$ is a nonconstant polynomial with coefficients $a_0, \ldots, a_n \in \mathbb{R}$, then f factors as a product of linear and quadratic polynomials with coefficients in $\mathbb{R}$.*

Proof. The proof is by induction on the degree of $f(x)$. The base case is $\deg(f) = 1$ or 2. Then f is itself linear or quadratic, and the corollary follows. So, suppose $\deg(f) = n \geq 3$, and that the corollary holds for all polynomials of degree less than n and at least 1.

By Theorem 2.1.1 there exists $\alpha \in \mathbb{C}$ which is a root of $f(x)$. If $\alpha \in \mathbb{R}$, then appealing to long division again (Corollary 2.3.16) as in the proof of Corollary 2.1.2, we can write $f(x) = q(x)(x - \alpha)$, where $\deg(q(x)) = n - 1$. By induction, $q(x)$ can be factored as a product of linear and quadratic polynomials, and therefore $f(x)$ can as well. Therefore, we assume $\alpha \in \mathbb{C} - \mathbb{R}$.

By Exercise 2.1.2, $\overline{zw} = \bar{z}\bar{w}$. Then, since $f(\alpha) = 0$, and since $0, a_0, \ldots, a_n \in \mathbb{R}$, we have

$$0 = \bar{0} = \overline{f(\alpha)} = \bar{a}_0 + \bar{a}_1\bar{\alpha} + \bar{a}_2\bar{\alpha}^2 + \cdots + \bar{a}_n\bar{\alpha}^n = a_0 + a_1\bar{\alpha} + a_2\bar{\alpha}^2 + \cdots + a_n\bar{\alpha}^n = f(\bar{\alpha}).$$

Thus, $\bar{\alpha}$ is also a root of f. Furthermore, since $\alpha \notin \mathbb{R}$, it follows that $\alpha \neq \bar{\alpha}$. Therefore, by applying long division *over the complex numbers* twice, we can write $f(x) = q(x)(x - \alpha)(x - \bar{\alpha})$ where $q(x)$ is a polynomial of degree $n - 2$ with coefficient in $\mathbb{C}$. On the other hand, observe that

$$(x - \alpha)(x - \bar{\alpha}) = x^2 - (\alpha + \bar{\alpha})x + |\alpha|^2,$$

and this is a polynomial with coefficients in $\mathbb{R}$ (again by Exercise 2.1.2), and so the long division used to write

$$f(x) = q(x)(x^2 - (\alpha + \bar{\alpha})x + |\alpha|^2)$$

actually takes place over the *real numbers*, and so in fact, $q(x)$ has coefficients in $\mathbb{R}$. Since it's degree is $1 \leq n - 2 < n$, by induction, $q(x)$ can be factors into linear and quadratic polynomials, and so the corollary follows. □

Remark 2.1.4. Justification for some of the claims in this proof will be clear after reading §2.3.

Exercises.

Exercise 2.1.1. Prove that multiplication of complex numbers is associative. More precisely, let $z = a + bi$, $w = c + di$, and $v = g + hi$, and prove that $z(wv) = (zw)v$.

Exercise 2.1.2. Let $z = a + bi, w = c + di \in \mathbb{C}$ and prove each of the following statements.

(i) $z + \bar{z}$ is real and $z - \bar{z}$ is imaginary.

(ii) $\overline{z + w} = \bar{z} + \bar{w}$.

(iii) $\overline{zw} = \bar{z}\bar{w}$.

You should...

- Be able to do all the exercises from this section.
- Understand both the arithmetic and geometry of the complex numbers.
- Know the statement of the Fundamental Theorem of Algebra (Theorem 2.1.1)

2.2 Fields

The rational, real, and complex numbers are all examples of our first abstract algebraic object called a *field*.

Definition 2.2.1. *A nonempty set $\mathbb{F}$, together with two operations called addition (denoted $a + b$) and multiplication (denoted $a \cdot b$ or ab), is a* **field** *if it satisfies the following axioms.*

(i) *Addition and multiplication are commutative, associative operations.*

(ii) *Multiplication distributes over addition: $a(b + c) = ab + ac$ for all $a, b, c \in \mathbb{F}$.*

(iii) *There exists an additive identity denoted $0 \in \mathbb{F}$: $a + 0 = a$ for all $a \in \mathbb{F}$.*

(iv) *For all $a \in \mathbb{F}$, there exists an additive inverse denoted $-a$, such that $-a + a = 0$.*

(v) *There exists a multiplicative identity denoted $1 \in \mathbb{F}$ with $1 \neq 0$: $1a = a$ for all $a \in \mathbb{F}$.*

(vi) *Every nonzero element $a \in \mathbb{F} - \{0\}$ has a multiplicative inverse a^{-1}, such that $aa^{-1} = 1$.*

A field $\mathbb{F}$ consists of all the data of $(\mathbb{F}, +, \cdot)$, since the set $\mathbb{F}$ alone does not determine the operations. However, we often suppress specific mention of the operations $+$ and $\cdot$ in the notation. When we do this abstractly, saying for example "*Let $\mathbb{F}$ be a field*", we are implying the existence of addition and multiplication on $\mathbb{F}$. When we do this for a specific example, such as $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$, we will assume that the operations are the "usual ones" (which implies that the example has already been described, and the operations introduced).

We use the additive and multiplicative inverses to make sense of subtraction and division in a field: $a - b = a + (-b)$ and $a \div b = \frac{a}{b} = ab^{-1}$ for $b \neq 0$.

Before we proceed, we point out a few familiar facts from $\mathbb{Q}, \mathbb{R}, \mathbb{C}$ which we may be tempted to take for granted, but which we should be careful to verify first.

Proposition 2.2.2. † *Suppose that $\mathbb{F}$ is a field and $a, b \in \mathbb{F}$.*

(i) *If $a + b = b$, then $a = 0$.*

(ii) *If $ab = b$ and $b \neq 0$, then $a = 1$.*

(iii) *$0a = 0$ for all $a \in \mathbb{F}$.*

(iv) *If $a + b = 0$, then $b = -a$.*

(v) *If $a \neq 0$ and $ab = 1$, then $b = a^{-1}$.*

(i), (ii), (iv), and (v) state that identities and inverses (additive and multiplicative) are unique.

Proof. For part (i), using the defining properties of a field we have

$$0 = b + (-b) = (a + b) + (-b) = a + (b + (-b)) = a + 0 = a.$$

For (ii), since $b \neq 0$, we have b^{-1} and so

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

Verifying (iii), we have

$$0a + a = 0a + 1a = (0 + 1)a = 1a = a$$

and so by part (i) just proved, $0a = 0$.

To prove (iv), observe that

$$b = a - a + b = a + b + (-a) = 0 - a = -a.$$

Similarly, to prove (v) we note that

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

□

If $\mathbb{F}$ is a field, then a subset $\mathbb{K} \subset \mathbb{F}$ is called a **subfield** if the elements $0, 1 \in \mathbb{F}$ are also in $\mathbb{K}$, and for every $a, b \in \mathbb{K}$, we have

$$a + b, ab, -a \in \mathbb{K} \quad \text{and if } a \neq 0 \text{ then } a^{-1} \in \mathbb{K}.$$

Proposition 2.2.3. *If $\mathbb{K} \subset \mathbb{F}$ is a subfield of a field $\mathbb{F}$, then the operations from $\mathbb{F}$ make $\mathbb{K}$ into a field.*

Proof. Since $\mathbb{K}$ is closed under addition and multiplication (meaning that $a+b$ and ab are in $\mathbb{K}$ when a, b are), it follows that addition and multiplication define operations on $\mathbb{K}$. Now we need to verify that they satisfy the axioms for a field. Since the operations are commutative and associative on $\mathbb{F}$, and since multiplication distributes over addition, this is true when we restrict to $\mathbb{K}$. Therefore (i) and (ii) hold. Since $0, 1 \in \mathbb{K}$, and the operations in $\mathbb{K}$ come from $\mathbb{F}$ (and since $\mathbb{F}$ is a field), conditions (iii) and (v) are verified. Because $\mathbb{K}$ is closed under taking additive and multiplicative inverses (by assumption), (iv) and (vi) also hold. □

If $\mathbb{K} \subset \mathbb{F}$ is a subfield, we also sometimes say that $\mathbb{F}$ is an **extension field** of $\mathbb{K}$. Note that the containments $\mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$ are all subfield containments.

Example 2.2.4. Given a positive real number $d > 0$, let $\sqrt{d} \in \mathbb{R}$ denote the positive square root of d.

Before providing a new example of a field, we observe that $\sqrt{2} \notin \mathbb{Q}$. To see this, suppose it were. Then we could write $\sqrt{2} = \frac{r}{s}$, where $r, s \in \mathbb{Z}$ are relatively prime. But then $r^2 = 2s^2$, so 2 would have to be a prime factor of r, and hence 4 a factor of r^2. Then $\frac{r^2}{4}$ would be an integer, and $\frac{r^2}{4}2 = s^2$, and thus 2 would be a prime factor of s^2, contradicting the fact that $\gcd(r, s) = 1$.

Let

$$\mathbb{Q}(\sqrt{2}) = \{a + b\sqrt{2} \in \mathbb{R} \mid a, b \in \mathbb{Q}\} \subset \mathbb{R}.$$

In Exercise 2.2.2 you are to prove that $\mathbb{Q}(\sqrt{2})$ is a subfield of $\mathbb{R}$. Note that $\mathbb{Q} \subset \mathbb{Q}(\sqrt{2})$, but $\mathbb{Q} \neq \mathbb{Q}(\sqrt{2})$ since $\sqrt{2} \notin \mathbb{Q}$ by the argument above.

Example 2.2.5. Let

$$\mathbb{Q}(i) = \{a + bi \mid a, b \in \mathbb{Q}\} \subset \mathbb{C}.$$

We claim that $\mathbb{Q}(i)$ is a subfield of $\mathbb{C}$. Note that $0, 1 \in \mathbb{Q}(i)$. Let $a + bi, c + di \in \mathbb{Q}(i)$. Then since $\mathbb{Q}$ is a field $a + c, b + d, ac - bd, ad + bc, -a, -b \in \mathbb{Q}$ and therefore

$$(a + bi) + (c + di) = (a + c) + (b + d)i \in \mathbb{Q}(i),$$

$$(a + bi)(c + di) = (ac - bd) + (ad + bc)i \in \mathbb{Q}(i), \text{ and}$$

$$-a - bi \in \mathbb{Q}(i).$$

All that remains is to verify that if $a+bi \neq 0$, then $(a+bi)^{-1} \in \mathbb{Q}(i)$. For this, observe that since $a+bi \neq 0$, $a^2+b^2 \neq 0$

$$(a+bi)^{-1} = \frac{a-bi}{a^2+b^2} = \frac{a}{a^2+b^2} + \frac{-b}{a^2+b^2}i$$

and this is in $\mathbb{Q}(i)$ since $\frac{a}{a^2+b^2}$ and $\frac{-b}{a^2+b^2}$ are in $\mathbb{Q}$.

Observe that since any field has both 0 and $1 \neq 0$, it has at least two elements. In fact, there are fields with exactly 2 elements, as the next fact shows.

Proposition 2.2.6. *For any prime integer $p > 1$, addition and multiplication of congruence classes makes $\mathbb{Z}_p$ into a field.*

Proof. According to Proposition 1.5.5, addition and multiplication are commutative, associative operations. Moreover, [0] and [1] act as the additive and multiplicative identities, respectively, every element has an additive inverse, and multiplication distributes over addition. By Corollary 1.5.7, any $[a] \neq [0]$ has a multiplicative inverse, which completes the proof. □

Exercises.

Exercise 2.2.1. Prove that if $\mathbb{F}$ is a field and $a, b \in \mathbb{F}$ with $ab = 0$, then either $a = 0$ or $b = 0$.

Exercise 2.2.2. Prove that $\mathbb{Q}(\sqrt{2})$ is a field. Hint: you should use the fact that $\sqrt{2} \notin \mathbb{Q}$.

You should...

- Be able to do all the exercises from this section and know the proofs of statements marked by †.
- Understand the *abstract nature* of what a field is, and that $\mathbb{Q}, \mathbb{R}, \mathbb{C}, \mathbb{Q}(\sqrt{2}), \mathbb{Q}(i), \mathbb{Z}_p$ for p a prime are all *examples* of fields.
- Be able to decide when a set with two operations is and isn't a field, and justify your conclusion (by proving that the axioms hold, or by illustrating the failure of one of the axioms). This means you need to know what the axioms are.
- Be able to decide when a subset of a field is a subfield.

2.3 Polynomials

Throughout this section, we let $\mathbb{F}$ be any field. A **polynomial** in the variable x with coefficients in $\mathbb{F}$ is a formal sum:

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n = \sum_{k=0}^{n} a_k x^k$$

where $x^0 = 1$. We often write $f = f(x)$ with the dependence on the variable x implicit.

Formally, a polynomial is an infinite sequence $\{a_k\}_{k=0}^{\infty}$ such that for some $n \geq 0$ we have $a_k = 0$ for all $k > n$. The terms of the sequence are just the coefficients, but the "sequential definition" is something we can make more precise than "a formal sum". In particular, coefficients that are 0 can be included or not included in the formal sum, potentially causing there to be ambiguity as to whether this is different than a formal sum in which all such zeros are omitted. Formally defining polynomials in terms of the sequence of coefficients alleviates any such confusion.

With the formalism to keep things straight, we continue to write our polynomials as formal sums as initially suggested (which will aid in our intuition). So, if

$$f = \sum_{k=0}^{n} a_k x^k,$$

and $m > n$, we can define $a_k = 0$ for all $n < k \leq m$, and then

$$\sum_{k=0}^{n} a_k x^k = \sum_{k=0}^{m} a_k x^k.$$

Because of this, we will always allow ourselves to add additional zero terms in this way, whenever convenient.

Remark 2.3.1. One may be tempted to define polynomials as certain types of functions $\mathbb{F} \to \mathbb{F}$. As we will see in Example 2.3.15, there are problems with this approach.

Let $\mathbb{F}[x]$ denote the set of all polynomials with coefficients in the field $\mathbb{F}$. We can define addition and multiplication on $\mathbb{F}[x]$ in the usual way. Specifically, given $f, g \in \mathbb{F}[x]$, if we conveniently add zero terms as necessary and write

$$f = \sum_{k=0}^{n} a_k x^k \text{ and } g = \sum_{j=0}^{n} b_j x^j,$$

then

$$f + g = \sum_{k=0}^{n} (a_k + b_k) x^k$$

and

$$fg = \sum_{k=0}^{2n} \Big(\sum_{j=0}^{k} a_j b_{k-j} \Big) x^k.$$

The expression in this last equation may look a little daunting, but it may be clarified by observing that the inner sum is the sum of all products of coefficients $a_j b_i$, where $i + j = k$.

A **constant** polynomial is a polynomial of the form $f = a_0$; that is, a constant polynomial is one in which all coefficients are zero, except possibly the initial one. We may then view $\mathbb{F}$ as a subset of $\mathbb{F}[x]$, identifying $a \in \mathbb{F}$ with the constant polynomial $f = a$. Because addition and multiplication in $\mathbb{F}[x]$ is defined in terms of addition and multiplication in $\mathbb{F}$, the following should not be surprising.

Proposition 2.3.2. *Suppose $\mathbb{F}$ is a field.*

(i) *Addition and multiplication are commutative, associative operations on $\mathbb{F}[x]$ which restrict to the operations of addition and multiplication on $\mathbb{F} \subset \mathbb{F}[x]$.*

(ii) *Multiplication distributes over addition: $f(g + h) = fg + fh$ for all $f, g, h \in \mathbb{F}[x]$.*

(iii) *$0 \in \mathbb{F}$ is an additive identity in $\mathbb{F}[x]$: $f + 0 = f$ for all $f \in \mathbb{F}[x]$.*

(iv) *Every $f \in \mathbb{F}[x]$ has an additive inverse given by $-f = (-1)f$ with $f + (-f) = 0$.*

(v) *$1 \in \mathbb{F}$ is the multiplicative identity in $\mathbb{F}[x]$: $1f = f$ for all $f \in \mathbb{F}[x]$.*

Proof. Let

$$f = \sum_{k=0}^{n} a_k x^k, \qquad g(x) = \sum_{k=0}^{n} b_k x^k, \qquad h(x) = \sum_{k=0}^{n} c_k x^k.$$

Then since addition and multiplication in $\mathbb{F}$ is commutative, we have

$$f+g=\sum_{k=0}^{n}(a_k+b_k)x^k=\sum_{k=0}^{n}(b_k+a_k)x^k=g+f,$$

proving that addition in $\mathbb{F}[x]$ is commutative. Next we calculate

$$fg=\sum_{k=0}^{2n}\Big(\sum_{j=0}^{k}a_jb_{k-j}\Big)x^k=\sum_{k=0}^{2n}\Big(\sum_{j=0}^{k}b_ja_{k-j}\Big)x^k=gf,$$

where the middle equality comes from the fact that the interior sum is the sum of all the products a_jb_i where $i+j=k$. Therefore, multiplication is also commutative.

Next we turn to associativity. For this, we note that since addition in $\mathbb{F}$ is associative, we have

$$\begin{aligned}(f+g)+h &= \Big(\sum_{k=0}^{n}(a_k+b_k)x^k\Big)+\sum_{k=0}^{n}c_kx^k=\sum_{k=0}^{n}((a_k+b_k)+c_k)x^k\\ &= \sum_{k=0}^{n}(a_k+(b_k+c_k))x^k\\ &= \Big(\sum_{k=0}^{n}a_kx^k\Big)+\Big(\sum_{k=0}^{n}(b_k+c_k)x^k\Big)=f+(g+h),\end{aligned}$$

and so addition is associative. For associativity of multiplication it is convenient to write

$$U_k=\sum_{i=0}^{k}a_ib_{k-i}\quad\text{and}\quad V_k=\sum_{j=0}^{k}b_jc_{k-j}.$$

Using the fact that in $\mathbb{F}$, multiplication is associative, addition is commutative and associative, and that multiplication distributes over addition in $\mathbb{F}$, we have

$$\begin{aligned}(fg)h &= \Big(\sum_{k=0}^{2n}\Big(\sum_{j=0}^{k}a_jb_{k-j}\Big)x^k\Big)\Big(\sum_{k=0}^{n}c_kx^k\Big)=\Big(\sum_{k=0}^{2n}U_kx^k\Big)\Big(\sum_{k=0}^{2n}c_kx^k\Big)\\ &= \sum_{k=0}^{4n}\Big(\sum_{j=0}^{k}U_jc_{k-j}\Big)x^k=\sum_{k=0}^{4n}\Big(\sum_{j=0}^{k}\Big(\sum_{i=0}^{j}a_ib_{j-i}\Big)c_{k-j}\Big)x^k\\ &= \sum_{k=0}^{4n}\Big(\sum_{j=0}^{k}\sum_{i=0}^{j}(a_ib_{j-i})c_{k-j}\Big)x^k=\sum_{k=0}^{4n}\Big(\sum_{j=0}^{k}\sum_{i=0}^{j}a_i(b_{j-i}c_{k-j})\Big)x^k\\ &= \sum_{k=0}^{4n}\Big(\sum_{i=0}^{k}a_i\sum_{j=0}^{k-i}b_jc_{k-i-j}\Big)x^k=\sum_{k=0}^{4n}\Big(\sum_{j=0}^{k}a_iV_{k-i}\Big)x^k\\ &= \Big(\sum_{k=0}^{2n}a_kx^k\Big)\Big(\sum_{k=0}^{2n}V_kx^k\Big)=\Big(\sum_{k=0}^{2n}a_kx^k\Big)\Big(\sum_{k=0}^{2n}\Big(\sum_{j=0}^{k}b_jc_{k-j}\Big)x^k\Big)=f(gh).\end{aligned}$$

The first equality in line 4 of this string of equations may seem slightly mysterious, but it is obtained from the preceding expression by observing that for each $k=0,\dots,4n$, the inner double sum is a sum of terms $a_i(b_jc_\ell)$, where $i+j+\ell=k$, together with a clever change of indices. Therefore multiplication is associative. The fact that multiplication distributes over addition is a similarly messy calculation which we leave to the interested reader.

The proofs of (iii)–(v) are in Exercise 2.3.1. □

There are many striking similarities between the polynomials over a field $\mathbb{F}$ on one hand, and the integers on the other. This is made possible by the notion of the **degree** of a polynomial

$$\deg\colon \mathbb{F}[x] \to \mathbb{Z} \cup \{-\infty\}$$

defined as follows. For any nonconstant polynomial f, declare $\deg(f) = n$ if $f = a_0 + a_1x + \cdots + a_nx^n$ and $a_n \neq 0$. The degree of a nonzero, constant polynomial $f = a_0 \neq 0$ is defined to be $\deg(f) = 0$, while the degree of the zero polynomial 0 is defined to be $\deg(0) = -\infty$. For a nonzero polynomial of degree n, $f = a_0 + a_1x + \cdots + a_nx^n$, the coefficient a_n of x^n is called the **leading coefficient**. A nonzero polynomial f is called **monic** if the leading coefficient is 1.

If we define $-\infty + a = a + (-\infty) = -\infty$ for any $a \in \mathbb{Z} \cup \{-\infty\}$, then we have the following

Lemma 2.3.3. *For any field $\mathbb{F}$ and $f, g \in \mathbb{F}[x]$, we have*

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

and

$$\deg(f + g) \leq \max\{\deg(f), \deg(g)\}.$$

The second inequality may indeed be strict since $\deg(f) = \deg(-f)$, but $f + (-f) = 0$ with $\deg(0) = -\infty$.

Proof. If $f = 0$, then $fg = 0$, and hence $\deg(fg) = -\infty = -\infty + \deg(g) = \deg(f) + \deg(g)$ by our convention. Also $f + g = g$ and so $\deg(f + g) = \deg(g) = \max\{-\infty, \deg(g)\} = \max\{\deg(f), \deg(g)\}$. So, we may assume neither f nor g is zero.

Let $n = \deg(f)$ and $m = \deg(g)$, and without loss of generality, suppose $m \leq n$. Then if

$$f = \sum_{k=0}^{n} a_k x^k \text{ and } g = \sum_{k=0}^{m} b_k x^k = \sum_{k=0}^{n} b_k x^k$$

we have

$$fg = \sum_{k=0}^{2n} \Big(\sum_{j=0}^{k} a_j b_{k-j} \Big) x^k.$$

The coefficient of x^k is

$$\sum_{j=0}^{k} a_j b_{k-j}.$$

If $k > m + n$, then for each $0 \leq j \leq k$, either $k - j > m$ or else $m \geq k - j > m + n - j$ and $j > n$. In either case, $a_j b_{k-j} = 0$, and hence the coefficient of x^k is zero. Therefore, $\deg(fg) \leq m + n$. On the other hand, a similar calculation shows that if $k = m + n$, then there is exactly one nonzero term in the sum defining the coefficient of x^k, namely $a_n b_m$ (by Exercise 2.2.1, the product of two nonzero elements of a field is nonzero). Consequently, $\deg(fg) = m + n = \deg(f) + \deg(g)$, proving the first statement.

For the second, we note that

$$f + g = \sum_{k=0}^{n} (a_k + b_k) x^k$$

and so $\deg(f + g) \leq n = \max\{n, m\} = \max\{\deg(f), \deg(g)\}$. □

Corollary 2.3.4. *If $\mathbb{F}$ is a field and $f, g \in \mathbb{F}[x]$ with $fg = 0$, then either $f = 0$ or $g = 0$.*

Proof. Since $-\infty = \deg(0) = \deg(fg) = \deg(f) + \deg(g)$ we must have either $\deg(f)$ or $\deg(g) = -\infty$, and consequently f or g is 0. □

In the integers, the only elements with a multiplicative inverse are 1 and -1. We call these the **units** in $\mathbb{Z}$. An element $f \in \mathbb{F}[x]$ is called a **unit** in $\mathbb{F}[x]$ if there exists a multiplicative inverse, that is if there exists an element $g \in \mathbb{F}[x]$ such that $fg = 1$. Note that every nonzero $a \in \mathbb{F} \subset \mathbb{F}[x]$ is a unit. It turns out that these are the only units in $\mathbb{F}[x]$.

Corollary 2.3.5. *Let $\mathbb{F}$ be a field and $f \in \mathbb{F}[x]$. Then f is a unit in $\mathbb{F}[x]$ if and only if $\deg(f) = 0$.*

Proof. Note that $\deg(f) = 0$ if and only if $f \in \mathbb{F}$ and $f \neq 0$ (that is, f is a nonzero constant). As we have already noted, f has a unit.

Conversely, suppose $f \in \mathbb{F}[x]$ has a multiplicative inverse $g \in \mathbb{F}[x]$, so that $fg = 1$. Then $0 = \deg(1) = \deg(fg) = \deg(f) + \deg(g)$. Note that neither $\deg(f)$ nor $\deg(g)$ can be $-\infty$ for otherwise their sum would be $-\infty$ instead of 0. So, neither f nor g is zero, and so $\deg(f), \deg(g) \geq 0$. Since their sum is also zero, both must equal zero, and hence f and g are constant polynomials. □

A nonconstant polynomial $f \in \mathbb{F}[x]$ is said to be **irreducible** (or more precisely, **irreducible in** $\mathbb{F}[x]$) if whenever we write $f = uv$ with $u, v \in \mathbb{F}[x]$, then either u or v is a unit (i.e. a nonzero constant by Corollary 2.3.5). Comparing $\mathbb{F}[x]$ with $\mathbb{Z}$, we see that irreducible polynomials are defined similar to prime numbers. Indeed, analogous to the Prime Factorization Theorem for $\mathbb{Z}$, Theorem 1.4.3, we have the following.

Theorem 2.3.6. *Suppose $\mathbb{F}$ is a field and $f \in \mathbb{F}[x]$ is any nonconstant polynomial. Then $f = ap_1 \cdots p_k$, where $a \in \mathbb{F}$ and $p_1, \ldots, p_k \in \mathbb{F}[x]$ are monic irreducible polynomials. If $f = bq_1 \cdots q_r$ with $b \in \mathbb{F}$ and $q_1, \ldots, q_r \in \mathbb{F}[x]$ monic irreducible, then $a = b$, $k = r$, and after reindexing*

$$p_i = q_i \text{ for all } i = 1, \ldots, k.$$

We note that if $f = u_1 \cdots u_k$ with $u_1, \ldots, u_k \in \mathbb{F}[x]$ all nonzero, then the leading coefficent of f is the product of the leading coefficients of $u_1, \ldots, u_k$. Consequently, if $f \in \mathbb{F}[x]$ is nonconstant, then the constant a from Theorem 2.3.6 is simply the leading coefficient of f. Therefore, an equivalent, cleaner statement of the theorem states: *any nonconstant, monic polynomial $f \in \mathbb{F}[x]$ can be factored as $f = p_1 \cdots p_k$ with $p_1, \ldots, p_k \in \mathbb{F}[x]$ monic irreducible polynomials, unique up to ordering.* You can view this simplification as being analogous to our assumption that the integers are positive in Theorem 1.4.3 (instead of just nonzero).

As with the prime factorization of the integers, the existence of the factorization in Theorem 2.3.6 is easier, and we dispense with that part first.

Lemma 2.3.7. *Suppose $\mathbb{F}$ is a field and $f \in \mathbb{F}[x]$ is any nonconstant, monic polynomial. Then $f = p_1 \cdots p_k$, where $p_1, \ldots, p_k$ are monic, irreducible polynomials.*

Proof. We induct on degree. For the base case, suppose f is a monic polynomial with $\deg(f) = 1$. Then if $f = uv$, with $u, v \in \mathbb{F}[x]$, we must have $1 = \deg(f) = \deg(u) + \deg(v)$, and since $\deg(u), \deg(v) \geq 0$, it follows that either $\deg(u) = 0$ or $\deg(v) = 0$, in which case either u or v is constant. Consequently, f is irreducible, completing the proof of the base case.

Now suppose that for some $n \geq 2$, any monic polynomial $g \in \mathbb{F}[x]$ with $0 < \deg(g) < n$ can be written as a product of monic, irreducible polynomials. Suppose $f \in \mathbb{F}[x]$ is a monic polynomial with $\deg(f) = n$, and we prove that it can be written as a product of monic, irreducible polynomials. If f is itself irreducible, we are done, so suppose it is not. In this case, we can find nonconstant polynomials $u, v \in \mathbb{F}[x]$ so that $f = uv$. Since the leading coefficient of f is 1, we may assume u and v are monic (why?). Since $\deg(f) = \deg(u) + \deg(v)$ and $\deg(u), \deg(v) > 0$, we must have $\deg(u), \deg(v) < \deg(f) = n$. Consequently, u and v can be written as products of monic, irreducible polynomials. But then, being a product of u and v, f is also a product of monic, irreducible polynomials, as required. By induction, we are done. □

Given polynomials $f, g \in \mathbb{F}[x]$ with $f \neq 0$, we say that f **divides** g, writing $f|g$, if we have $g = fu$ for some $u \in \mathbb{F}[x]$. The following observations are completely analogous to those in Proposition 1.4.2 for $\mathbb{Z}$.

Proposition 2.3.8. *Let $f, g, h \in \mathbb{F}[x]$. Then*

(i) *If $f \neq 0$, then $f|0$.*

(ii) *If $f|1$, then $f \in \mathbb{F}$ (that is, f is constant).*

(iii) *If $f|g$ and $g|f$, then $f = bg$ for some $b \in \mathbb{F}$.*

(iv) *If $f|g$ and $g|h$, then $f|h$.*

(v) *If $f|g$ and $f|h$, then $f|(ug+vh)$ for all $u, v \in \mathbb{F}[x]$.*

Proof. See Exercise 2.3.2. □

If $f, g \in \mathbb{F}[x]$ are nonzero polynomials, a **greatest common divisor** of f and g is a polynomial $h \in \mathbb{F}[x]$ such that

$$\text{(i) } h|f \text{ and } h|g, \text{ and} \qquad \text{(ii) if } k \in \mathbb{F}[x] \text{ and } k|f \text{ and } k|g, \text{ then } k|h.$$

Greatest common divisors (if they exist) are not quite unique, but only differ by constants. That is, suppose $h, h' \in \mathbb{F}[x]$ are greatest common divisors of nonzero polynomials $f, g \in \mathbb{F}[x]$. Then $h|h'$ and $h'|h$, and so by Proposition 2.3.8 (iii), we have $h = bh'$ for some $b \in \mathbb{F}$. If a greatest common divisor of a pair of nonzero polynomials $f, g \in \mathbb{F}[x]$ exists, then there is one which is monic (by multiplying any greatest common divisor by the inverse of the leading coefficient) and hence unique. We will refer to this as **the greatest common divisor** (of if clarification is needed, **the monic greatest common divisor**), and denote it $\gcd(f, g)$.

As with the integers, greatest common divisors do indeed exist for all pairs of nonzero polynomials, and as with the integers, the key to finding them is the following **Euclidean algorithm for polynomials**.

Proposition 2.3.9. *Given $f, g \in \mathbb{F}[x]$ with $g \neq 0$, there exist unique $q, r \in \mathbb{F}[x]$ so that $f = gq + r$ with $\deg(r) < \deg(g)$.*

Sketch of proof. If $\deg(g) > \deg(f)$, we simply let $q = 0$ and $r = f$. Therefore, suppose that $\deg(g) \leq \deg(f)$. The proof is basically long division of polynomials, which we can think of as being carried out recursively. Consequently, the proof is most easily described by induction on $N = \deg(f) - \deg(g)$.

Let $f = a_0 + a_1 x + \cdots + a_n x^n$, where $n = \deg(f)$. Let $m = \deg(g) \leq \deg(f) \leq n$ and write

$$g = b_0 + b_1 x + \cdots + b_m x^m.$$

Setting $c = \frac{a_n}{b_m}$, we observe that $cx^{n-m} \in \mathbb{F}[x]$ and $cx^{n-m}g$ has degree n with leading term $a_n x^n$. Thus,

$$\deg(f - cx^{n-m}g) \leq n - 1.$$

We now begin the induction. For the base case we suppose $n - m = \deg(f) - \deg(g) = 0$. Then $\deg(f - cx^{n-m}g) \leq n - 1 < n = m = \deg(g)$, and we can set $q = cx^{n-m}$ and $r = f - cx^{n-m}g$, implying $f = qg + r$, completing the proof of the base case.

Suppose that for some $N > 0$ we know that the proposition is true for any pair of polynomials for which the difference in degrees is less than N. Let $f, g \in \mathbb{F}[x]$ $n = \deg(f)$, $m = \deg(g)$ and $n - m = \deg(f) - \deg(g) = N$. Just as above, setting $c = \frac{a_n}{b_m}$ we have $\deg(f - cx^{n-m}g) < n$. If $\deg(f - cx^{n-m}g) < m = \deg(g)$, we can set $q = cx^{n-m}$ and $r = f - cx^{n-m}g$, and we are done. So suppose $\deg(f - cx^{n-m}g) \geq m$ and set $f' = f - cx^{n-m}g$. Since $\deg(f') - \deg(g) < \deg(f) - \deg(g) = N$, the inductive assumption implies that there exists $q', r' \in \mathbb{F}[x]$ with $\deg(r') < \deg(g)$ so that $f' = q'g + r'$. But then

$$f = f' + cx^{n-m}g = q'g + r' + cx^{n-m}g = (q' + cx^{n-m})g + r'.$$

Setting $q = q' + cx^{n-m}$ and $r = r'$, we have $f = qg + r$ and $\deg(r) < \deg(g)$. By induction, we are done with the existence statement.

For the uniqueness, suppose $f = qg + r = q'g + r'$, with $\deg(r), \deg(r') < \deg(g)$. Then $g(q' - q) = r - r'$, and so

$$\deg(g) + \deg(q' - q) = \deg(g(q' - q)) = \deg(r - r') < \deg(g).$$

This is only possible if $\deg(q' - q) = -\infty = \deg(r - r')$, and hence $r = r'$ and $q = q'$, as required. □

Proposition 2.3.10. † *Any two nonzero polynomials $f, g \in \mathbb{F}[x]$ have a greatest common divisor in $\mathbb{F}[x]$. In fact, among all polynomials in the set*

$$\mathcal{M} = \{uf + vg \mid u, v \in \mathbb{F}[x]\},$$

if $d \geq 0$ is the smallest degree of any nonzero element of $\mathcal{M}$, then any element of $\mathcal{M}$ of degree d is a greatest common divisor of f and g.

The reader should compare the next proof to the proof of Proposition 1.4.8.

Proof. Let $h \in \mathcal{M}$ be any element with $\deg(h) = d$. Suppose $k \in \mathbb{F}[x]$ and $k|f$ and $k|g$. Then by Proposition 2.3.8, $k|h$, which shows that h satisfies the second of the required properties for a greatest common divisor.

Now suppose $h' \in \mathcal{M}$ is any nonzero element. By assumption, $\deg(h') \geq \deg(h)$. Proposition 2.3.9 provides a pair of elements $q, r \in \mathbb{F}[x]$, with $\deg(r) < \deg(h)$ so that $h' = qh + r$. Since $h, h' \in \mathcal{M}$, we have $r = h' - qh \in \mathcal{M}$. Because of the minimality in $\deg(h)$, it follows that $r = 0$, and so $h' = qh$. Applying this to the two cases $h' = 1f + 0g = f$ and $h' = 0f + 1g = g$, we see that $h|f$ and $h|g$. Therefore, h is a greatest common divisor of f and g. □

Example 2.3.11. We find $\gcd(f, g)$, where $f = 3x^3 - 5x^2 - 3x + 5$ and $g = x^3 - 2x^2 + 1$ and express it as $\gcd(f, g) = uf + vg$. This is the same kind of calculation as for the integers. We do repeated long division until we end up with remainder 0. The last nonzero remainder is a greatest common divisor, and we can repeatedly substitute back in to find u and v (then multiply by an appropriate constant, if necessary, to get a monic polynomial). The first long division is

$$\begin{array}{r} 3 \\ x^3 - 2x^2 + 1 \,\overline{)\, 3x^3 - 5x^2 - 3x + 5} \\ -\quad 3x^3 - 6x^2 \qquad + 3 \\ \hline x^2 - 3x + 2 \end{array}$$

which gives

$$3x^3 - 5x^2 - 3x + 5 = 3(x^3 - 2x^2 + 1) + x^2 - 3x + 2.$$

We do long division again with g and the remainder, $x^2 - 3x + 2$:

$$\begin{array}{r} x \\ x^2 - 3x + 2 \,\overline{)\, x^3 - 2x^2 \qquad + 1} \\ -\quad x^3 - 3x^2 + 2x \\ \hline x^2 - 2x + 1 \end{array} \qquad \begin{array}{r} x + 1 \\ x^2 - 3x + 2 \,\overline{)\, x^3 - 2x^2 \qquad + 1} \\ -\quad x^3 - 3x^2 + 2x \\ \hline x^2 - 2x + 1 \\ -\quad x^2 - 3x + 2 \\ \hline x - 1 \end{array}$$

which gives

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

Repeating long division again with the first remainder and the second remainder, we get

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

Since this last long division has no remainder, the second remainder is a greatest common divisor, and so $\gcd(f, g) = x - 1$. We may solve for this in the second equation, solve for the first remainder in the first equation, and substitute to arrive at:

$$\gcd(f, g) = x - 1 = x^3 - 2x^2 + 1 - (x + 1)(x^2 - 3x + 2) = g - (x + 1)(f - 3g) = -(x + 1)f + (3x + 4)g.$$

Proposition 2.3.12. *If $f, g, h \in \mathbb{F}[x]$, $\gcd(f, g) = 1$, and $f|gh$, then $f|h$.*

Proof. This is Exercise 2.3.4. □

Corollary 2.3.13. *If $f \in \mathbb{F}[x]$ is irreducible and $f|gh$, then $f|g$ or $f|h$.*

Proof. Since f is irreducible, the only divisors are constant multiples of itself, and constants. Thus either $f|g$ and we are done, or else $\gcd(f,g)=1$. But then by the previous proposition $f|h$, and we are done. □

Sketch of proof of Theorem 2.3.6. As mentioned earlier, it suffices to prove the theorem for monic polynomials. The existence of the factorization of a monic polynomial $f \in \mathbb{F}[x]$ into monic, irreducible polynomials is precisely Lemma 2.3.7. The uniqueness will follow from Corollary 2.3.13, in much the same way as the proof of Theorem 1.4.3 followed from Corollary 1.4.11, so we just sketch the idea.

The proof is by induction on the degree of f. The base case is when $\deg(f)=1$ in which case f is irreducible, and the theorem follows. Suppose that for some $n \geq 2$, the theorem is true for all nonconstant, monic polynomials with degree less than n, and let $f \in \mathbb{F}[x]$ be any monic polynomial with $\deg(f)=n$. If f is itself irreducible, then the theorem follows, so we suppose it has at least two irreducible factors in any factorization.

Write $f = p_1 \cdots p_k = q_1 \cdots q_r$, with all p_i, q_j monic irreducible polynomials, $k, r \geq 2$, and we assume the factors have been ordered so that $\deg(p_1) \geq \deg(p_2) \geq \cdots \geq \deg(p_k)$ and $\deg(q_1) \geq \deg(q_2) \geq \cdots \geq \deg(q_r)$. Then by Corollary 2.3.13, $p_1|q_i$ for some i and $q_1|p_j$ for some j. By our degree assumptions, reordering among polynomials of the same degree if necessary, we may assume $p_1|q_1$ and $\deg(p_1)=\deg(q_1)$. Since these polynomials are monic, it follows that $p_1 = q_1$. Then $\frac{f}{p_1} \in \mathbb{F}[x]$, and

$$\frac{f}{p_1} = p_2 p_3 \cdots p_k = q_2 q_3 \cdots q_r.$$

Since $1 \leq \deg(\frac{f}{p_1}) < n$, the inductive assumption implies $r=k$, and after reordering, $p_i = q_i$ for all $2 \leq i \leq k$. Since $p_1 = q_1$, the uniqueness is proved for f. By induction, this completes the proof. □

We now return to the notion of roots of a polynomial, first visited in §2.1. First, we must allow ourselves to think about polynomials as functions.

Proposition 2.3.14. *Suppose $f = a_0 + a_1 x + \cdots + a_n x^n \in \mathbb{F}[x]$ is a polynomial. Then there is a well-defined function, also denoted $f\colon \mathbb{F} \to \mathbb{F}$, defined for all $a \in \mathbb{F}$ by*

$$f(a) = a_0 + a_1 a + \cdots + a_n a^n.$$

Proof. The formula clearly defines a function, and this depends only on the coefficients (addition of zero coefficients does not affect the value by Proposition 2.2.2 (iii)). □

One might be tempted to define polynomials as functions given by a formula as in the proposition, but this runs into difficulty as the following example illustrates.

Example 2.3.15. Suppose $\mathbb{F}$ is a *finite* field, for example $\mathbb{Z}_p$ for some prime $p > 1$. Then there are only finitely many functions $\mathbb{F} \to \mathbb{F}$. On the other hand, $\{x^n\}_{n=0}^{\infty}$ gives us an infinite family of distinct polynomials. Therefore, multiple polynomials (in fact, infinitely many) define the same functions.

Given a polynomial $f \in \mathbb{F}[x]$, a **root** of f in $\mathbb{F}$ is an element $\alpha \in \mathbb{F}$ such that $f(\alpha)=0$. The Euclidean Algorithm for polynomials has the following useful corollary about roots.

Corollary 2.3.16. *For every $f \in \mathbb{F}[x]$ and $\alpha \in \mathbb{F}$, there exists a polynomial $q \in \mathbb{F}[x]$ so that*

$$f = (x-\alpha)q + f(\alpha).$$

In particular, α is a root of f if and only if $(x-\alpha)|f$.

Proof. By Proposition 2.3.9, there exists $q, r \in \mathbb{F}[x]$ so that $\deg(r) < \deg(x-\alpha) = 1$ and $f = (x-\alpha)q + r$. Since $\deg(r) < 1$, r is constant and its value is given by $r(\alpha) = (\alpha-\alpha)q(\alpha) + r(\alpha) = f(\alpha)$, proving $f = (x-\alpha)q + f(\alpha)$. The last statement follows from the definition of a root and of divisibility. □

Degree 1 polynomials (also called **linear** polynomials) are necessarily irreducible. According to Corollary 2.3.16, every root of $f \in \mathbb{F}[x]$ gives rise to a linear factor in the factorization from Theorem 2.3.6. For any root α of f, we say that the **multiplicity** of α is the number of times $(x-\alpha)$ appears in the factorization from Theorem 2.3.6.

Proposition 2.3.17. *Given a nonconstant polynomial $f \in \mathbb{F}[x]$, the number of roots of f, counted with multiplicity, is at most* $\deg(f)$.

Proof. According to Theorem 2.3.6 we may write $f = ap_1 \cdots p_k(x-\alpha_1)\cdots(x-\alpha_r)$ where $a \in \mathbb{F}$, $p_1, \ldots, p_k$ are the irreducible factors of degree at least 2, and $\alpha_1, \ldots, \alpha_r \in \mathbb{F}$ are the roots. Then r is the number of roots, counted with multiplicity, and

$$\deg(f) = \deg(p_1) + \cdots + \deg(p_k) + r \geq r.$$

□

If $\mathbb{K} \subset \mathbb{F}$ is a subfield, then by Proposition 2.2.3, $\mathbb{K}$ is a field, and we get a natural inclusion $\mathbb{K}[x] \subset \mathbb{F}[x]$. That is, any polynomial with coefficients in $\mathbb{K}$ can be thought of as a polynomial with coefficients in $\mathbb{F}$. A polynomial $f \in \mathbb{K}[x]$ may have no roots in $\mathbb{K}$, but could have roots in $\mathbb{F}$.

Example 2.3.18. For any positive integer n, the polynomial $x^n - 1 \in \mathbb{Q}[x]$ has $1 \in \mathbb{Q}$ as a root, as well as $-1 \in \mathbb{Q}$ if n is even, but no other roots in $\mathbb{Q}$. However, $\mathbb{Q} \subset \mathbb{C}$, and by Corollary 2.1.2, $x^n - 1$ has exactly n roots in $\mathbb{C}$. Let $\zeta_n = e^{2\pi i/n}$, then the n roots of $x^n - 1$ in $\mathbb{C}$ are precisely

$$C_n = \{\zeta_n^k \mid k \in \{0, \ldots, n-1\}\} = \{e^{2\pi k i/n} \mid k \in \{0, \ldots, n-1\}\}.$$

The set C_n is called the set of n^{th} **roots of unity** in $\mathbb{C}$. If, for any positive rational number $d > 0$, we let $\sqrt[n]{d} \in \mathbb{R}$ denote the positive n^{th} root of d, then the set of roots of $x^n - d \in \mathbb{Q}[x] \subset \mathbb{C}[x]$ in $\mathbb{C}$ is given by

$$\{\zeta_n^k \sqrt[n]{d} \mid k \in \{0, \ldots, n-1\}\}.$$

More generally, if for any $d \in \mathbb{C}$, if we let $\sqrt[n]{d}$ denote some root of $x^k - d$ in $\mathbb{C}$, then the set above again describes the set of all roots of $x^k - d$.

For $n = 2$ and $d \in \mathbb{Q}$, $d > 0$, this simplifies to the familiar set of roots $\{\pm\sqrt{d}\}$. Turning to the general quadratic (i.e. degree two) polynomial $ax^2 + bx + c \in \mathbb{Q}[x]$, the *quadratic formula* tells us that the roots are

$$\left\{\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}\right\}.$$

where $\sqrt{b^2 - 4ac}$ is positive real if $b^2 - 4ac$ is positive, and is the imaginary number $i\sqrt{|b^2 - 4ac|}$, otherwise.

There are formulas for roots of general cubic and quartic (degree 3 and 4, respectively), which are more complicated, but again are given in terms of *radicals*, $\sqrt[m]{\ }$, and the roots of unity C_n, for $n, m \geq 1$. It turns out that degree 4 is the largest integer for which such general formulas exist. For example, the degree 5 polynomial $f = x^5 - x - 1 \in \mathbb{Q}[x]$, has 5 roots in $\mathbb{C}$ by Corollary 2.1.2, but one can show that these roots **cannot** be obtained from the rational numbers and the roots of unity using the operations of addition, subtraction, multiplication, division, and taking radicals. One of the crowing achievements of abstract algebra is to provide an elegant explanation for this. This will require quite a bit more development of the theory, but we hope to be able to explain this eventually.

Exercises.

Exercise 2.3.1. Prove parts (iii)–(v) of Proposition 2.3.2.

Exercise 2.3.2. Prove Proposition 2.3.8.

Exercise 2.3.3. Let $f = x^5 + 2x^4 + 2x^3 - x^2 - 2x - 2$ and $g = 4x^4 + 16$. Find $\gcd(f, g)$ and express it as $uf + vg$.

Exercise 2.3.4. Prove Proposition 2.3.12. Hint: Look at the proof of Proposition 1.4.10.

Exercise 2.3.5. Prove that a polynomial $f \in \mathbb{F}[x]$ of degree 3 is irreducible in $\mathbb{F}[x]$ if it does not have a root in $\mathbb{F}$.

Exercise 2.3.6. Consider the polynomial $f = x^3 - x + 2 \in \mathbb{Z}_5[x]$ (more precisely, $f = [1]x^3 - [1]x + [2]$). Prove that f is irreducible in $\mathbb{Z}_5[x]$. *Hint: Use Exercise 2.3.5.*

You should...

- Be able to do all the exercises from this section and know the proofs of statements marked by †.
- Understand what the Euclidean algorithm for $\mathbb{F}[x]$ says.
- Be able to compute the greatest common divisor $\gcd(f, g)$ for $f, g \in \mathbb{F}[x]$, and express it in the form $uf + vg$, for $u, v \in \mathbb{F}[x]$.
- Know what an irreducible polynomial is and what the factorization into irreducible factors is.
- Understand the relationship between linear factors and roots (see Corollary 2.3.16).
- Understand the similarities between $\mathbb{Z}$ and $\mathbb{F}[x]$.

2.4 A little linear algebra

Linear algebra formally belongs in the realm of abstract algebra, but it is usually taught separately, often with a focus on applications to calculus and differential equations. Here we recall some of the basic facts from linear algebra. Although we will discuss this topic in a slightly more general situation than you may have learned it, the proofs you should have seen before carry over, essentially verbatim, and so we will not develop them from the ground up. Furthermore, our interest in linear algebra is primarily motivational, providing us with useful example to analyze, rather than being one of the primary players in our discussion. Consequently, we will sketch some proofs, while for others we refer the reader to essentially any text on linear algebra.

Definition 2.4.1. *Let $\mathbb{F}$ be any field. A* **vector space** *over $\mathbb{F}$ is a nonempty set V together with an operation called* **addition***, written $\mathbf{v} + \mathbf{w}$ for $\mathbf{v}, \mathbf{w} \in V$, as well as* **scalar multiplication**

$$\mathbb{F} \times V \to V$$

written $(a, \mathbf{v}) \mapsto a\mathbf{v}$, satisfying the following set of axioms*:*

(i) *Addition is associative and commutative.*

(ii) *There is an additive identity element, $\mathbf{0} \in V$, with the property that $\mathbf{v} + \mathbf{0} = \mathbf{v}$ for all $\mathbf{v} \in V$.*

(iii) *Every element $\mathbf{v} \in V$ has an additive inverse, given by $-\mathbf{v} = -1\mathbf{v}$, so that $\mathbf{v} + (-\mathbf{v}) = \mathbf{0}$.*

(iv) *For every $\mathbf{v} \in V$, $1\mathbf{v} = \mathbf{v}$.*

(v) *For all $a, b \in \mathbb{F}$ and $\mathbf{v} \in V$, we have $a(b\mathbf{v}) = (ab)\mathbf{v}$.*

(vi) *For all $a \in \mathbb{F}$ and $\mathbf{v}, \mathbf{w} \in V$, we have $a(\mathbf{v} + \mathbf{w}) = a\mathbf{v} + a\mathbf{w}$.*

(vii) *For all* $a, b \in \mathbb{F}$ *and* $\mathbf{v} \in V$, $(a+b)\mathbf{v} = a\mathbf{v} + b\mathbf{v}$.

We call the readers attention to the following similarity in the definition of a field: Both fields and vector spaces are defined as sets with operations and additional structure. These are both examples of *abstract algebraic objects*. In each of the following examples, we invite the reader to verify that each satisfies all the axioms of a vector space.

Example 2.4.2. The set $V = \{\mathbf{0}\}$ consisting of a single element called $\mathbf{0}$ is a vector space over any field $\mathbb{F}$ where addition is defined (the only way it can be) as $\mathbf{0} + \mathbf{0} = \mathbf{0}$ and scalar multiplication (also the only way possible) as $a\mathbf{0} = \mathbf{0}$, for all $a \in \mathbb{F}$.

Example 2.4.3. For any integer $n \geq 1$, $\mathbb{F}^n$ with the usual addition and scalar multiplication

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n) \quad \text{and} \quad a(x_1, \ldots, x_n) = (ax_1, \ldots, ax_n),$$

for $a \in \mathbb{F}$, and $(x_1, \ldots, x_n), (y_1, \ldots, y_n) \in \mathbb{F}^n$, make $\mathbb{F}^n$ into a vector space over $\mathbb{F}$. The case of $\mathbb{F} = \mathbb{R}$ gives the familiar n–**dimensional space**, $\mathbb{R}^n$.

Example 2.4.4. For any set X, the set of functions $\mathbb{F}^X = \{f \colon X \to \mathbb{F}\}$ has the structure of a vector space over $\mathbb{F}$ where, for $f, g \in \mathbb{F}^X$ and $a \in \mathbb{F}$, we define $f + g$ and af by the following formulas

$$(f+g)(x) = f(x) + g(x) \quad \text{and} \quad (af)(x) = a(f(x)),$$

for all $x \in X$.

Example 2.4.5. Since we view $\mathbb{F} \subset \mathbb{F}[x]$, the polynomials with coefficients in $\mathbb{F}$, we can restrict the multiplication in $\mathbb{F}[x]$ to multiplication of polynomials by constants, $\mathbb{F} \times \mathbb{F}[x] \to \mathbb{F}[x]$. Then $\mathbb{F}[x]$ is a vector space over $\mathbb{F}$.

Example 2.4.6. Suppose $\mathbb{K} \subset \mathbb{F}$ is a subfield of a field $\mathbb{F}$. Then $\mathbb{F}$ is a vector space over $\mathbb{K}$. For example, $\mathbb{C}$ is a vector space over $\mathbb{R}$ and over $\mathbb{Q}$, $\mathbb{Q}(i)$ and $\mathbb{Q}(\sqrt{2})$ are a vector spaces over $\mathbb{Q}$.

Suppose that V is a vector space over $\mathbb{F}$. A **vector subspace** (or just **subspace**) is a nonempty subset $W \subset V$ **closed** under addition and scalar multiplication. This means that for all $\mathbf{v}, \mathbf{w} \in W$ and $\mathbf{a} \in \mathbb{F}$, we have

$$\mathbf{v} + \mathbf{w} \in W \quad \text{and} \quad a\mathbf{v} \in W.$$

Exercise 2.4.2 asks you to prove that if W is a subspace, then addition and scalar multiplication make it into a vector space over $\mathbb{F}$.

Example 2.4.7. The polynomials $\mathbb{F}[x]$ have many interesting subspaces. For example, for every integer $n \geq 0$, we can consider $\mathbb{F}[x]_n \subset \mathbb{F}[x]$ consisting of polynomials of degree at most n. According to Lemma 2.3.3 this is a subspace.

Example 2.4.8. The vector space of all functions $\mathbb{R}^{\mathbb{R}}$ has many interesting subspaces. One of these is the subspace of continuous functions $C(\mathbb{R}, \mathbb{R}) \subset \mathbb{R}^{\mathbb{R}}$. Every polynomial $f \in \mathbb{R}[x]$ determines a function (see Proposition 2.3.14), and unlike the case of $\mathbb{F}[x]$, where $\mathbb{F}$ is a finite field, no two distinct polynomials in $\mathbb{R}[x]$ define the same function (see Example 2.3.15). Consequently, we may view $\mathbb{R}[x]$ as a subspace of $\mathbb{C}(\mathbb{R}, \mathbb{R})$, and hence also a subspace of $\mathbb{R}^{\mathbb{R}}$.

Example 2.4.9. If $\mathbb{K} \subset \mathbb{F}$ is a subfield, then any intermediate field $\mathbb{K} \subset \mathbb{L} \subset \mathbb{F}$ is a subspace of $\mathbb{F}$, thought of as a vector space over $\mathbb{K}$ (see Example 2.4.6).

A set of elements $\mathbf{v}_1, \ldots, \mathbf{v}_n \in V$ is said to be **linearly independent** if whenever $a_1, \ldots, a_n \in \mathbb{F}$ have

$$a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n = \mathbf{0}$$

then $a_1 = \cdots = a_n = 0$. We call the expression on the left a **linear combination** of the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$, and $a_1, \ldots, a_n$ are called the **coefficients** of the linear combination. Linear independence can then be

expressed briefly by saying that no nontrivial linear combination of the vectors is zero (here "nontrivial" means, "not all coefficients equal to 0"). A set of elements $\mathbf{v}_1, \ldots, \mathbf{v}_n \in V$ is said to **span** V if every vector $\mathbf{v} \in V$ can be expressed as a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_n$. If $\mathbf{v}_1, \ldots, \mathbf{v}_n$ spans and is linearly independent, then we call the set a **basis** for V.

The basic relation between linear independence and spanning is the following.

Proposition 2.4.10. *Suppose V is a vector space over a field $\mathbb{F}$ having a basis $\{v_1, \ldots, v_n\}$ with $n \geq 1$.*

(i) *For all $\mathbf{v} \in V$, $\mathbf{v} = a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n$ for exactly one $(a_1, \ldots, a_n) \in \mathbb{F}^n$.*

(ii) *If $\mathbf{w}_1, \ldots, \mathbf{w}_n$ span V, then they are linearly independent.*

(iii) *If $\mathbf{w}_1, \ldots, \mathbf{w}_n$ are linearly independent, then they span V.*

Consequently, any two bases have the same number of vectors.

We refer the reader to any text on linear algebra for a proof.

If a vector space V over $\mathbb{F}$ has a basis with n vectors, then V is said to be n–**dimensional** (over $\mathbb{F}$) or is said to have **dimension** n.

Example 2.4.11. For any field $\mathbb{F}$, the set of vectors

$$\mathbf{e}_1 = (1, 0, \ldots, 0), \mathbf{e}_2 = (0, 1, 0, \ldots, 0), \ldots, \mathbf{e}_n = (0, 0, \ldots, 0, 1) \in \mathbb{F}^n$$

are a basis for $\mathbb{F}^n$ called the **standard basis vectors**. Thus $\mathbb{F}^n$ is n–dimensional.

Example 2.4.12. If X is any finite set $X = \{x_1, \ldots, x_n\}$, then the function $f_1, \ldots, f_n$ defined by

$$f_i(x) = \begin{cases} 1 & \text{if } x = x_i \\ 0 & \text{otherwise} \end{cases}$$

form a basis for the vector space of functions $\mathbb{F}^X$. For $X = \{1, \ldots, n\}$ this is really the same as the previous example (see Exercise 1.1.5).

Example 2.4.13. The vector space of polynomials $\mathbb{F}[x]_n$ of degree at most n has as basis $1, x, x^2, \ldots, x^n$, and thus has dimension $n + 1$. You are asked to prove this in Exercise 2.4.3.

Example 2.4.14. The field $\mathbb{Q}(\sqrt{2})$ is a vector space over $\mathbb{Q}$, and has as basis $1, \sqrt{2}$, so is 2–dimensional over $\mathbb{Q}$. Similarly, $\mathbb{Q}(i)$ is 2–dimensional over $\mathbb{Q}$ with basis $1, i$ and $\mathbb{C}$ is 2–dimensional over $\mathbb{R}$ with the same basis, $1, i$.

Given two vector spaces V and W over $\mathbb{F}$ a **linear transformation** is a function $T\colon V \to W$ such that for all $a \in \mathbb{F}$ and $\mathbf{v}, \mathbf{w} \in V$, we have

$$T(a\mathbf{v}) = aT(\mathbf{v}) \quad \text{and} \quad T(\mathbf{v} + \mathbf{w}) = T(\mathbf{v}) + T(\mathbf{w}).$$

In Exercise 2.4.4 you are asked to prove that the composition of two linear transformation is a linear transformation.

Proposition 2.4.15. *If V and W are vector spaces and $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is a basis for V then any function from $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \to W$ extends uniquely to a linear transformation $V \to W$. In particular, any linear transformation $T\colon V \to W$ is determined by $T(\mathbf{v}_1), \ldots, T(\mathbf{v}_n)$.*

Sketch of the proof. Given a linear transformation $T\colon V \to W$ and an arbitrary vector $\mathbf{v} = a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n$, we have

$$T(\mathbf{v}) = T(a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n) = a_1T(\mathbf{v}_1) + \cdots + a_nT(\mathbf{v}_n)$$

so T is uniquely determined by $T(\mathbf{v}_1), \ldots, T(\mathbf{v}_n)$. Conversely, given any function $T\colon \{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \to W$, we can use the equation above to define a function $T\colon V \to W$, and one easily checks that T is a linear transformation. □

From this, we easily deduce the following.

Corollary 2.4.16. *If $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is a basis for a vector space V and $\mathbf{w}_1, \ldots, \mathbf{w}_m$ is a basis for a vector space W (both over $\mathbb{F}$), then any linear transformation $T\colon V \to W$ determines (and is determined by) the $m \times n$ matrix*

$$A = A(T) = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix},$$

where the entries A_{ij} are defined by

$$T(\mathbf{v}_j) = A_{1j}\mathbf{w}_1 + A_{2j}\mathbf{w}_2 + \cdots + A_{mj}\mathbf{w}_m.$$

This corollary states that if we let $\mathcal{L}(V, W)$ denote the set of all linear transformations from V to W and $M_{m\times n}(\mathbb{F})$ the set of $m\times n$ matrices with entries in $\mathbb{F}$, then $T \mapsto A(T)$ defines a bijection $\mathcal{L}(V, W) \to M_{m\times n}(\mathbb{F})$. We often say that the matrix $A(T)$ **represents** the linear transformation T.

Remark 2.4.17. Bases for V and W are not unique, and different choices of bases will determine different bijections between $\mathcal{L}(V, W)$ and $M_{m\times n}(\mathbb{F})$. Thus when saying that a matrix determines a linear transformation, this will implicitly assume chosen bases on the domain and range.

If $A \in M_{m\times n}(\mathbb{F})$ and $B \in M_{k\times m}(\mathbb{F})$, then we can define the **matrix product** of B and A as the $k \times n$ matrix BA with ij-entry

$$(BA)_{ij} = B_{i1}A_{1j} + B_{i2}A_{2j} + \cdots + B_{im}A_{mj} = \sum_{l=1}^{m} B_{il}A_{lj}.$$

Because matrix multiplication is defined in terms of multiplication in $\mathbb{F}$, it follows that multiplication of matrices is associative, whenever it is defined: $C(BA) = (CB)A$ if $A \in M_{m\times n}(\mathbb{F})$, $B \in M_{k\times m}(\mathbb{F})$, and $C \in M_{r\times k}(\mathbb{F})$.

Example 2.4.18. We often represent vectors in $\mathbb{F}^n$ as $n \times 1$ matrices, also called **column vectors**:

$$\mathbf{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix},$$

where $v_1, \ldots, v_n \in \mathbb{F}$. Exercise 2.4.5 asks you to prove that the linear transformation $T\colon \mathbb{F}^n \to \mathbb{F}^n$ represented by a matrix $A \in M_{n\times n}(\mathbb{F})$ with respect to the standard basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$, is given simply by the matrix product

$$T(\mathbf{v}) = A\mathbf{v}.$$

(Note that this matrix product makes sense as A is $n \times n$ and $\mathbf{v}$ is $n \times 1$.) We will denote this linear transformation by $T = T_A\colon \mathbb{F}^n \to \mathbb{F}^n$.

Under the bijection between linear transformation and matrices, composition corresponds to matrix multiplication. More precisely, a calculation proves the following.

Proposition 2.4.19. *Suppose that V, W, and U are vector spaces over $\mathbb{F}$, with fixed chosen bases. If $T\colon V \to W$ and $S\colon W \to U$ are linear transformations represented by matrices $A = A(T)$ and $B = B(S)$, then $ST = S \circ T\colon V \to U$ is a linear transformation represented by the matrix $BA = B(S)A(T)$.* □

Given a vector space V over $\mathbb{F}$, we let $\mathrm{GL}(V) \subset \mathcal{L}(V, V)$ denote the subset of **invertible linear transformations**

$$\mathrm{GL}(V) = \{T \in \mathcal{L}(V, V) \mid T \text{ is a bijection } \} = \mathcal{L}(V, V) \cap \mathcal{S}ym(V).$$

If we choose a basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ for V, then the image of $GL(V)$ in $M_{n\times n}(\mathbb{F})$ is denoted $\mathrm{GL}(n, \mathbb{F})$. This can be alternatively defined in terms of the *determinant*, which we now recall.

Given a matrix $A \in M_{n\times n}(\mathbb{F})$, for each $1 \leq i, j \leq n$, the ij–**minor** of A is the $(n-1)\times(n-1)$ matrix obtained by deleting the i^{th} row and j^{th} column from A. The **determinant** of A, denoted $\det(A)$, is defined recursively as

$$\det(A) = \sum_{j=1}^{n} (-1)^{j+1} \det(M_{1j}(A)) = \sum_{i=1}^{n} (-1)^{i+1} \det(M_{i1}(A))$$

where the determinant of 1×1 matrix $A = (A_{11})$ is just the entry: $\det(A) = A_{11}$. The two formulas for the determinant here are sometimes called **cofactor expansions** (more precisely, the formulas are the cofactor expansions over the first row and first column, respectively).

Example 2.4.20. For a 2×2 matrix $A = (A_{ij})$, the determinant is given by

$$\det(A) = A_{11}A_{22} - A_{12}A_{21}.$$

Example 2.4.21. For a 3×3 matrix $A = (A_{ij})$, the determinant is given by

$$\det(A) = A_{11}(A_{22}A_{33} - A_{23}A_{32}) - A_{12}(A_{21}A_{33} - A_{23}A_{31}) + A_{13}(A_{21}A_{32} - A_{22}A_{31}).$$

The relationship between invertibility and the determinant is provided by the following theorem. We refer the reader to any text on linear algebra for a proof.

Theorem 2.4.22. *For $n \geq 1$,*

$$\mathrm{GL}(n, \mathbb{F}) = \{A \in M_{n\times n}(\mathbb{F}) \mid \det(A) \neq 0\}.$$

Concretely, this says that the linear transformation $T_A \colon \mathbb{F}^n \to \mathbb{F}^n$ defined by $T_A(\mathbf{v}) = A\mathbf{v}$ (see Example 2.4.18) is invertible if and only if $\det(A) \neq 0$.

We end with some other useful properties of the determinant (again, see any text on linear algebra). Let A^t be the **transpose** of the matrix A, obtained by reversing the roles of the rows and columns. That is, the ij^{th} entry of A^t is the ji^{th} entry of A: $A^t_{ij} = A_{ji}$.

Proposition 2.4.23. *For any matrices $A, B \in M_{n\times n}(\mathbb{F})$, we have*

$$\det(A^t) = \det(A) \quad \text{and} \quad \det(AB) = \det(A)\det(B).$$

An invertible linear transformation $T_A \colon \mathbb{R}^n \to \mathbb{R}^n$ represented by a matrix $A \in \mathrm{GL}(n, \mathbb{R})$ is said to be **orientation preserving** if $\det(A) > 0$, and **orientation reversing** if $\det(A) < 0$. In this case, $\det(A)$ represents the *signed changed in n–dimensional volume* of the transformation T_A (with negative sign for orientation reversing transformations). Here, 2–dimensional volume is area and 3–dimensional volume is the usual notion of volume in 3–space.

Exercises.

Exercise 2.4.1. Prove that the polynomials $\mathbb{F}[x]$ form a vector space over $\mathbb{F}$ as claimed in Example 2.4.5.

Exercise 2.4.2. Prove that if $W \subset V$ is a subspace of a vector space over $\mathbb{F}$, then addition and scalar multiplication make W into a vector space.

Exercise 2.4.3. Prove that $1, x, x^2, \ldots, x^n$ form a basis for the vector space $\mathbb{F}[x]_n$ of polynomials over $\mathbb{F}$ of degree at most n.

Exercise 2.4.4. Suppose that $T \colon V \to W$ and $S \colon W \to U$ are linear transformations (of vector spaces over $\mathbb{F}$). Prove that $ST = S \circ T \colon V \to U$ is a linear transformation.

Exercise 2.4.5. Prove that if $A \in M_{n\times n}$, then $T_A\colon \mathbb{F}^n \to \mathbb{F}^n$ defined by matrix multipication $T_A(\mathbf{v}) = A\mathbf{v}$ (viewing $\mathbf{v}$ as a column vector) defines a linear transformation. Further prove that the j^{th} column of A is the column vector $T_A(\mathbf{e}_j)$, and that with respect to the standard basis vectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$, the matrix representing T_A is precisely A.

Exercise 2.4.6. Suppose V is a vector space over a field $\mathbb{F}$ and that $\mathbb{K} \subset \mathbb{F}$ is a subfield. Prove that by restricting scalar multiplication to $\mathbb{K}$, V is a vector space over $\mathbb{K}$.

You should...

- Be able to do all the exercises from this section.
- Know the examples of this section, and have a general comfortability with linear algebra.

2.5 Euclidean geometry basics.

Linear algebra (especially over $\mathbb{R}$) provides the perfect framework for studying geometry. As we will see in the next section, there are deeper connections between abstract algebra and geometry as well.

We take the perspective that Euclidean geometry is the study of the **Euclidean distance** on $\mathbb{R}^n$, defined as usual by

$$|\mathbf{x}-\mathbf{y}| = \sqrt{(x_1-y_1)^2+\cdots+(x_n-y_n)^2}, \quad \text{where} \quad \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n.$$

The Euclidean distance is nothing but the **Euclidean norm** of the difference of two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, where the norm of a vector $\mathbf{v}$ is

$$|\mathbf{v}| = \sqrt{v_1^2+\cdots+v_n^2}.$$

The norm is closely related to the **Euclidean inner product**, defined by

$$\mathbf{x}\cdot\mathbf{y} = x_1y_1+\cdots+x_ny_n = \sum_{i=1}^n x_iy_i = \mathbf{x}^t\mathbf{y}.$$

Note that transpose of the column vector $\mathbf{x}$ is the **row vector** $\mathbf{x}^t = (x_1, x_2, \ldots, x_n)$, and $\mathbf{x}^t\mathbf{y}$ is the matrix product (strictly speaking, the matrix product is a 1×1 matrix, and we are abusing notation and simply referring to this as a real number). The norm is given by $|\mathbf{v}| = \sqrt{\mathbf{v}\cdot\mathbf{v}}$.

A few straightforward properties of the Euclidean inner product are listed here.

Proposition 2.5.1. *For all $\mathbf{v}, \mathbf{u}, \mathbf{w} \in \mathbb{R}^n$ and $\mathbf{a} \in \mathbb{R}$, we have*

(i) $\mathbf{v}\cdot\mathbf{v} \geq 0$ *with equality if and only if* $\mathbf{v} = \mathbf{0}$.

(ii) $\mathbf{v}\cdot\mathbf{w} = \mathbf{w}\cdot\mathbf{v}$, *and*

(iii) $(a\mathbf{v}+\mathbf{u})\cdot\mathbf{w} = a(\mathbf{v}\cdot\mathbf{w}) + (\mathbf{u}\cdot\mathbf{w})$.

We leave the proof as an easy exercise for the reader to verify. The three properties of the inner product in the proposition are sometimes called **positive definiteness**, **symmetry**, and **(bi)linearity**, respectively.

An **isometry** of $\mathbb{R}^n$ (or a **rigid motion** of $\mathbb{R}^n$) is a bijection $\Phi\colon \mathbb{R}^n \to \mathbb{R}^n$ that *preserves distance*, meaning that

$$|\Phi(\mathbf{x}) - \Phi(\mathbf{y})| = |\mathbf{x}-\mathbf{y}|,$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. We let

$$\mathrm{Isom}(\mathbb{R}^n) = \{\Phi\colon \mathbb{R}^n \to \mathbb{R}^n \mid \Phi \text{ is an isometry } \}$$

denote the set of all isometries of $\mathbb{R}^n$.

Proposition 2.5.2. *For any $n \geq 1$, the identity is an isometry* $\mathrm{id} \in \mathrm{Isom}(\mathbb{R}^n)$. *For all* $\Phi, \Psi \in \mathrm{Isom}(\mathbb{R}^n)$, *we have*

$$\Phi \circ \Psi \in \mathrm{Isom}(\mathbb{R}^n) \quad \text{and} \quad \Phi^{-1} \in \mathrm{Isom}(\mathbb{R}^n).$$

Proof. The fact that $\mathrm{id} \in \mathrm{Isom}(\mathbb{R}^n)$ is immediate from the definition of an isometry. Let $\Phi, \Psi \in \mathrm{Isom}(\mathbb{R}^n)$. First, since Φ and Ψ are bijections, so is $\Phi \circ \Psi$ (see Lemma 1.1.7). Moreover, since each are isometries we compute

$$|\Phi \circ \Psi(\mathbf{x}) - \Phi \circ \Psi(\mathbf{y})| = |\Phi(\Psi(\mathbf{x})) - \Phi(\Psi(\mathbf{y}))| = |\Psi(\mathbf{x}) - \Psi(\mathbf{y})| = |\mathbf{x} - \mathbf{y}|,$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Therefore, $\Phi \circ \Psi \in \mathrm{Isom}(\mathbb{R}^n)$. Next, since $\mathrm{id} \in \mathrm{Isom}(\mathbb{R}^n)$, we have

$$|\mathbf{x} - \mathbf{y}| = |\,\mathrm{id}(\mathbf{x}) - \mathrm{id}(\mathbf{y})| = |\Phi \circ \Phi^{-1}(\mathbf{x}) - \Phi \circ \Phi^{-1}(\mathbf{y})| = |\Phi(\Phi^{-1}(\mathbf{x})) - \Phi(\Phi^{-1}(\mathbf{y}))| = |\Phi^{-1}(\mathbf{x}) - \Phi^{-1}(\mathbf{y})|,$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Since Φ^{-1} is also a bijection we have $\Phi^{-1} \in \mathrm{Isom}(\mathbb{R}^n)$. □

In Exercise 2.5.1 you are asked to show that a linear transformation $T\colon \mathbb{R}^n \to \mathbb{R}^n$ is an isometry if and only if it preserve the inner product: $T(\mathbf{v}) \cdot T(\mathbf{w}) = \mathbf{v} \cdot \mathbf{w}$. According to Corollary 2.4.16 and Exercise 2.4.5, there is a matrix $A \in \mathrm{GL}(n, \mathbb{R})$ so that $T = T_A\colon \mathbb{R}^n \to \mathbb{R}^n$ is given by matrix multiplication $T_A(\mathbf{v}) = A\mathbf{v}$. The condition that T_A preserves the inner product is succinctly described in terms of the matrix A. For this, observe that

$$T_A(\mathbf{v}) \cdot T_A(\mathbf{w}) = (A\mathbf{v}) \cdot (A\mathbf{w}) = (A\mathbf{v})^t A\mathbf{w} = \mathbf{v}^t A^t A\mathbf{w}.$$

Thus, if $A^t A = I$, the $n \times n$ identity matrix, then T_A preserves the inner product. Conversely, we note that for the standard basis vectors

$$\mathbf{e}_i \cdot \mathbf{e}_j = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

while

$$\mathbf{e}_i^t A^t A \mathbf{e}_j = (A^t A)_{ij}.$$

Consequently, if T_A preserves the inner product, then $T_A(\mathbf{e}_i) \cdot T_A(\mathbf{e}_j) = \mathbf{e}_i \cdot \mathbf{e}_j$, and $A^t A = I$.

Therefore, $T_A(\mathbf{v}) = A\mathbf{v}$ is an isometry if and only if it preserves the inner product, which happens if and only if $A^t A = I$. We define the **orthogonal group**

$$\mathrm{O}(n) = \{A \in \mathrm{GL}(n, \mathbb{R}) \mid A^t A = I\} \subset \mathrm{GL}(n, \mathbb{R}).$$

Thus $\mathrm{O}(n)$ are the matrices representing linear isometries of $\mathbb{R}^n$.

By Proposition 2.4.23, $1 = \det(I) = \det(A^t A) = \det(A^t)\det(A) = \det(A)^2$, and consequently $\det(A) = 1$ or $\det(A) = -1$. The **special orthogonal group**, is defined to be

$$\mathrm{SO}(n) = \{A \in \mathrm{O}(n) \mid \det(A) = 1\}$$

These are the matrices representing orientation preserving linear isometries of $\mathbb{R}^n$.

Another source of isometries of $\mathbb{R}^n$ are *translations*. Given $\mathbf{v} \in \mathbb{R}^n$, we define the **translation by $\mathbf{v}$** to be the map

$$\tau_{\mathbf{v}}\colon \mathbb{R}^n \to \mathbb{R}^n,$$

defined by $\tau_{\mathbf{v}}(\mathbf{x}) = \mathbf{x} + \mathbf{v}$. In Exercise 2.5.3 you are to prove that for all $\mathbf{v} \in \mathbb{R}^n$, $\tau_{\mathbf{v}}$ is an isometry.

Since the composition of isometries is an isometry, it follows that for all $A \in \mathrm{O}(n)$ and $\mathbf{v} \in \mathbb{R}^n$, the composition

$$\Phi_{A,\mathbf{v}}(\mathbf{x}) = \tau_{\mathbf{v}}(T_A(\mathbf{x})) = A\mathbf{x} + \mathbf{v}$$

is an isometry. In fact, this accounts for all isometries.

Theorem 2.5.3.

$$\mathrm{Isom}(\mathbb{R}^n) = \{\Phi_{A,\mathbf{v}} \mid A \in \mathrm{O}(n),\ \mathbf{v} \in \mathbb{R}^n\}.$$

The proof would take us a bit far afield, so we leave it as an exercise for the interested reader (or perhaps it might eventually make it into an appendix).

Exercises.

Exercise 2.5.1. Suppose $T\colon \mathbb{R}^n \to \mathbb{R}^n$ is a linear transformation. Prove that T is an isometry if and only if $T(\mathbf{v}) \cdot T(\mathbf{w}) = \mathbf{v} \cdot \mathbf{w}$. Recall that an isometry is a *bijection* that preserves distance.

Exercise 2.5.2. Given a matrix $A \in \mathrm{GL}(n, \mathbb{R})$, prove that $A \in \mathrm{O}(n)$ if and only if the columns vectors of A form an **orthonormal basis**. That is, if we write $A = (\mathbf{v}_1 \,|\, \mathbf{v}_2 \,|\, \cdots \,|\, \mathbf{v}_n)$, were $\mathbf{v}_j$ is the j^{th} column, then $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is a basis and $\mathbf{v}_i \cdot \mathbf{v}_j = \delta_{ij}$ for all $1 \leq i, j \leq n$.

Exercise 2.5.3. For any $\mathbf{v} \in \mathbb{R}^n$, the translation by $\mathbf{v}$, $\tau_{\mathbf{v}}(\mathbf{x}) = \mathbf{x} + \mathbf{v}$ is an isometry.

Exercise 2.5.4. For any $A, B \in \mathrm{O}(n)$ and $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$, $\Phi_{A,\mathbf{v}} \circ \Phi_{B,\mathbf{w}}, \Phi_{A,\mathbf{v}}^{-1} \in \mathrm{Isom}(\mathbb{R}^n)$ by Proposition 2.5.2. According to Theorem 2.5.3 there exists $C, C' \in \mathrm{O}(n)$ and $\mathbf{u}, \mathbf{u}' \in \mathbb{R}^n$ so that

$$\Phi_{A,\mathbf{v}} \circ \Phi_{B,\mathbf{w}} = \Phi_{C,\mathbf{u}} \qquad \text{and} \qquad \Phi_{A,\mathbf{v}}^{-1} = \Phi_{C',\mathbf{u}'}.$$

Find $C, C', \mathbf{u}, \mathbf{u}'$ in terms of $A, B, \mathbf{v}, \mathbf{w}$.

You should...

- Be able to do all the exercises from this section.
- Know the examples of this section, and have a general comfortability with isometries.

2.6 Groups and rings

2.6.1 Groups

It may seem that the preceding sections have very little in common. It so happens that essentially everything that has been discussed so far can be connected (in multiple ways, quite often) to the following fundamental object in abstract algebra.

Definition 2.6.1. *A* **group** *is a nonempty set with an operation $(G, *)$ such that*

(i) *The operation $*$ is associative.*

(ii) *There is an* **identity** *$e \in G$, with the property that $e * g = g * e = g$, for all $g \in G$.*

(iii) *For all $g \in G$, there exists an* **inverse** *$g^{-1} \in G$, with the property that $g * g^{-1} = g^{-1} * g = e$.*

Remark 2.6.2. We often leave the operation out of the notation, for example saying "*G with the operation $*$ is a group*" or sometimes just "*G is a group*", with the operation implicitly understood. The latter case will typically only occur after the group has been introduced, so then the operation is assumed to be the one already defined. For $a, b \in G$, we may often shorten $a * b$ to simply ab, when no confusion can arise.

Groups have become the "poster child" for abstract algebra. On the one hand they are incredibly simple algebraic objects, being defined by only three axioms (compare this with the definition of a field, Definition 2.2.1, and a vector space, Definition 2.4.1). This simplicity allows one to find groups everywhere. On the other hand, **group theory** (the study of groups) is remarkably rich with an incredible amount of structure being deduced from the three simple axioms.

In this section, we list numerous examples of groups, tying together much of the discussion so far. In the next chapter we take up a systematic study of groups, providing a glimpse of their beauty and hopefully inspiring further investigation.

Example 2.6.3. Let X be any set. Then according to Proposition 1.3.1 the set $\mathcal{S}ym(X)$, bijections from X to itself, is a group with the operation of composition $\circ$.

Example 2.6.4. According to Proposition 1.4.1, the integers $\mathbb{Z}$ with addition forms a group.

Example 2.6.5. If $\mathbb{F}$ is a field, then from Definition 2.2.1 $(\mathbb{F}, +)$ is a group, as is the set of nonzero elements with multiplication $(\mathbb{F}^\times, \cdot)$.

Example 2.6.6. The previous example includes the case $\mathbb{F} = \mathbb{Z}_p$ and $\mathbb{F}^\times = \mathbb{Z}_p^\times$, for a prime $p > 1$. In fact, for any integer $n \geq 1$, $(\mathbb{Z}_n, +)$ is a group, as are the invertible elements with multiplication $(\mathbb{Z}_n^\times, \cdot)$. See Exercise 2.6.1 and Exercise 2.6.2.

Example 2.6.7. If $\mathbb{F}$ is a field, then by Proposition 2.3.2, the set of polynomials over $\mathbb{F}$ with addition $(\mathbb{F}[x], +)$ is a group. More generally from Definition 2.4.1, if V is any vector space over $\mathbb{F}$, then $(V, +)$ is a group.

Example 2.6.3 is different than the rest of the examples we have seen so far. Specifically, as we have seen, composition fails to be commutative on $\mathcal{S}ym(X)$ (unless X has only one or two elements). Any group in which the operation is commutative is called an *abelian group*, and otherwise it is called *nonabelian*. Thus, $\mathcal{S}ym(X)$ is a nonabelian group if $|X| > 2$. On the other hand, the groups $\mathbb{Z}$, $\mathbb{Z}_n$, $\mathbb{Z}_n^\times$, $\mathbb{F}$, $\mathbb{F}^\times$, and V are all abelian groups. This may suggest that "most" groups are abelian, though this is actually quite far from true.

As with subfields and subspaces, we obtain many more interesting examples by looking at subsets of a group, which are themselves groups. If $(G, *)$ is a group, then a nonempty subset $H \subset G$ is a **subgroup** if it is closed under $*$ and inversion. More precisely, H is a subgroup if

1. for all $g, h \in H$, $g * h \in H$, and
2. for all $g \in H$, $g^{-1} \in H$.

If $H \subset G$ is a subgroup, we will write $H < G$ noting that this does not exclude the possibility that $H = G$.

Proposition 2.6.8. *If $(G, *)$ is a group and $H \subset G$ is a subgroup, then the group operation on G restricts to an operation on H making it into a group.*

Proof. We leave the proof of this as Exercise 2.6.3. □

Example 2.6.9. If G is a group, then G is clearly a subgroup. If $e \in G$ is the identity, then $\{e\}$ is a subgroup. To see this, observe that $e * e = e \in \{e\}$ and $e^{-1} = e \in \{e\}$. These are sometimes called the **trivial subgroups**.

Example 2.6.10. If $\mathbb{K} \subset \mathbb{F}$ is a subfield, then $\mathbb{K}$ is a subgroup of $\mathbb{F}$ under addition. Similarly, $\mathbb{K}^\times$ is a subgroup of $\mathbb{F}^\times$ under multiplication. These both follow immediately from the definition of subfield.

Example 2.6.11. Suppose $\mathbb{F}$ is a field. Although $\mathbb{F}^\times \subset \mathbb{F}$, and each of $\mathbb{F}^\times$ and $\mathbb{F}$ are groups, the operations making them into groups are different, and consequently, $\mathbb{F}^\times$ is **not** a subgroup of $\mathbb{F}$. Similarly, for $\mathbb{Z}_n^\times$ is not a subgroup of $\mathbb{Z}_n$.

Example 2.6.12. The subset

$$S^1 = \{z \in \mathbb{C} \mid |z| = 1\} \subset \mathbb{C}^\times$$

of complex numbers with absolute value 1 is geometrically a circle in the complex plane. Given $z, w \in S^1$, since $|zw| = |z||w| = 1$, it follows that $zw \in S^1$. In addition, $\frac{1}{z} = \frac{\bar{z}}{|z|^2} = \frac{\bar{z}}{1} = \bar{z}$ and $|\bar{z}| = |z| = 1$, so that $\frac{1}{z} \in S^1$. Therefore, $S^1 \subset \mathbb{C}^\times$ (being nonempty) is a subgroup of $\mathbb{C}^\times$ with multiplication.

Since any complex number $z \in \mathbb{C}$ can be expressed as $z = re^{i\theta}$, where $r = |z| \geq 0$ and $\theta \in \mathbb{R}$, we have

$$S^1 = \{e^{i\theta} \mid \theta \in \mathbb{R}\}.$$

In Exercise 2.6.4, you are asked to prove that the roots of unity $C_n \subset S^1$ (defined in Example 2.3.18) form a subgroup.

Example 2.6.13. If $V \subset W$ is a subspace of the vector space W over a field $\mathbb{F}$, then V is a subgroup of W under addition.

Example 2.6.14. By Proposition 2.5.2, $\mathrm{Isom}(\mathbb{R}^n) \subset \mathcal{S}ym(\mathbb{R}^n)$ is a subgroup.

Example 2.6.15. Suppose V is a vector space. Recall that $\mathrm{GL}(V)$ is the set of invertible linear transformations from V to itself. We claim that $\mathrm{GL}(V) \subset \mathcal{S}ym(V)$ is a subgroup. Suppose that $T, S \in \mathrm{GL}(V)$, then we must show that $TS, T^{-1} \in \mathrm{GL}(V)$. First, we show TS is linear, and hence in $\mathrm{GL}(V)$. For this, suppose $\mathbf{u}, \mathbf{v} \in V$ and $a \in \mathbb{F}$. Then since S and T are both linear, we have

$$TS(\mathbf{u}+\mathbf{v}) = T(S(\mathbf{u}+\mathbf{v})) = T(S(\mathbf{u}) + S(\mathbf{v})) = TS(\mathbf{u}) + TS(\mathbf{v}),$$

and

$$TS(a\mathbf{u}) = T(S(a\mathbf{u})) = T(aS(\mathbf{u})) = aTS(\mathbf{u}).$$

So, TS is linear.

To see that T^{-1} is linear, let $\mathbf{u}, \mathbf{v} \in V$ and $a \in \mathbb{F}$. Then since T is linear, we have

$$T^{-1}(\mathbf{u}+\mathbf{v}) = T^{-1}(TT^{-1}(\mathbf{u}) + TT^{-1}(\mathbf{v})) = T^{-1}(T(T^{-1}(\mathbf{u}) + T^{-1}(\mathbf{v}))) = T^{-1}(\mathbf{u}) + T^{-1}(\mathbf{v}),$$

and

$$T^{-1}(a\mathbf{u}) = T^{-1}(aTT^{-1}(\mathbf{u})) = T^{-1}(TaT^{-1}(\mathbf{u})) = aT^{-1}(\mathbf{u}).$$

So, T^{-1} is linear, and hence also in $\mathrm{GL}(V)$. The identity is in $\mathrm{GL}(V)$, so it is nonempty, and hence a subgroup.

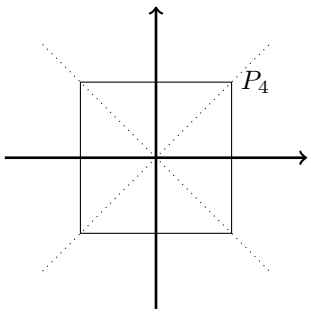
Example 2.6.16. If $\mathbb{F}$ is a field, an **automorphism** of $\mathbb{F}$ is a bijection $\sigma\colon \mathbb{F} \to \mathbb{F}$ such that $\sigma(a+b) = \sigma(a) + \sigma(b)$ and $\sigma(ab) = \sigma(a)\sigma(b)$. Let $\mathrm{Aut}(\mathbb{F}) = \{\sigma\colon \mathbb{F} \to \mathbb{F} \mid \sigma \text{ is an automorphism}\} \subset \mathcal{S}ym(\mathbb{F})$ denote the set of all automorphisms of $\mathbb{F}$. In Exercise 2.6.5 you are asked to prove that $\mathrm{Aut}(\mathbb{F})$ is a subgroup.

The three previous examples are illustrations of subgroups $G < \mathcal{S}ym(X)$ where X is a set with some additional structure. In Example 2.6.14 the additional structure on $\mathbb{R}^n$ is the Euclidean metric. In Example 2.6.15, the additional structure on V is the vector space structure. In Example 2.6.16, the structure on $\mathbb{F}$ is that of a field. Another common type of subgroup of $\mathcal{S}ym(X)$ is the following.

Proposition 2.6.17. *Suppose $Y \subset X$ is a nonempty subset and $G < \mathcal{S}ym(X)$ is a subgroup. Then $H = \{\sigma \in G \mid \sigma(Y) = Y\} < G$ is a subgroup.*

Proof. Let $\sigma, \tau \in H$. Then $\sigma \circ \tau(Y) = \sigma(\tau(Y)) = \sigma(Y) = Y$, hence $\sigma \circ \tau \in H$. In addition, since $\sigma(Y) = Y$, we have $\sigma^{-1}(Y) = \sigma^{-1}(\sigma(Y)) = Y$, hence $\sigma^{-1} \in H$. Therefore, $H < G$ is a subgroup. □

Example 2.6.18. Suppose $n \geq 3$ and $P_n \subset \mathbb{R}^2$ is a regular n–gon (below is a regular 4–gon, i.e. a square). Then we let $D_n < \mathrm{Isom}(\mathbb{R}^2)$ denote the subgroup consisting of isometries that preserve P_n (as in Proposition 2.6.17). This is called the **dihedral group** of the regular n–gon.



For $n = 4$, this group, D_4, contains the identity, three rotations through angles $\pi/2$, π, and $3\pi/2$, together with the reflections in the four lines shown (the horizontal and vertical axes together with the lines through opposite vertices). In Exercise 2.6.6 you are asked to show that these are all the elements of D_4, and thus it has exactly 8 elements.

The following illustrates a similar, but slightly different type of example to those coming from Proposition 2.6.17.

Example 2.6.19. Suppose $\mathbb{K} \subset \mathbb{F}$ is a subfield. Recall the group $\mathrm{Aut}(\mathbb{F})$ from Example 2.6.16. We define

$$\mathrm{Aut}(\mathbb{F}, \mathbb{K}) = \{\sigma \in \mathrm{Aut}(\mathbb{F}) \mid \sigma(a) = a \text{ for all } a \in \mathbb{K}\}.$$

This is clearly a subgroup, for if $\sigma, \tau \in \mathrm{Aut}(\mathbb{F}, \mathbb{K})$ and $a \in \mathbb{K}$, then $\sigma \circ \tau(a) = \sigma(\tau(a)) = \sigma(a) = a$ and $\sigma^{-1}(a) = \sigma^{-1}(\sigma(a)) = a$. We have pointed out that $\mathbb{F}$ is a vector space over $\mathbb{K}$, and we leave it to the reader to verify that the elements in $\mathrm{Aut}(\mathbb{F}, \mathbb{K})$ are also invertible linear transformations $\mathbb{F} \to \mathbb{F}$.

As a concrete example, observe that $\sigma(z) = \bar{z}$ defines an element of $\mathrm{Aut}(\mathbb{C}, \mathbb{R})$.

2.6.2 Rings

While groups are the simplest abstract algebraic object, and their study will dominate the remainder of the course, the following is another abstract algebraic object that generalizes several of the objects we have seen.

Definition 2.6.20. *A* **ring** *is a nonempty set with two operations, called addition and multiplication, $(R, +, \cdot)$ such that*

(i) *$(R, +)$ is an abelian group. The additive identity is denoted 0 (zero), and the additive inverse of an element $a \in R$ is denoted $-a$.*

(ii) *Multiplication is associative.*

(iii) *Multiplication distributes over addition: $a \cdot (b + c) = a \cdot b + a \cdot c$ and $(b + c) \cdot a = b \cdot a + c \cdot a$.*

We typically write multiplication as juxtaposition $ab = a \cdot b$.

Rings come in a variety of types. Two special types of rings are the following.

- If multiplication is commutative, we call $(R, +, \cdot)$ a **commutative ring**.
- If there exists an element $1 \in R \setminus \{0\}$ such that $a1 = 1a = a$ for all $a \in R$, then we say that R is a **ring with** 1.

Example 2.6.21. By Proposition 1.4.1, $(\mathbb{Z}, +, \cdot)$ is a commutative ring with 1.

As with groups and fields, we will often suppress the symbols $+$ and $\cdot$ when denoting a ring, if those operations are clear from the context. Thus, we simply say that $\mathbb{Z}$ is a ring (or that addition and multiplication make $\mathbb{Z}$ into a ring).

Example 2.6.22. We leave it as Exericse 2.6.8 to verify that modular arithmetic makes $\mathbb{Z}_n$ into a commutative ring with 1 for any $n \geq 1$.

Example 2.6.23. By Definition 2.2.1, any field is a commutative ring with 1. Indeed, a field is simply a commutative ring with 1 in which every nonzero element has a multiplicative inverse.

Example 2.6.24. Suppose $\mathbb{F}$ is any field. Proposition 2.3.2 shows that $\mathbb{F}[x]$, the polynomials over $\mathbb{F}$, form a ring with respect to the usual addition and multiplication.

Example 2.6.25. If R is a ring and X is any nonempty set, consider the set $R^X = \{f\colon X \to R\}$ of all functions from X to R. We claim that pointwise addition and multiplication of functions:

$$(f+g)(x) = f(x) + g(x) \text{ and } (fg)(x) = f(x)g(x)$$

for $x \in X$ and $f, g \in R^X$ makes R^X into a ring. Because the operations are defined pointwise, this follows almost immediately from the fact that R is itself a ring. For example, to see that the operations are associative, we let $f, g, h \in R^X$, then since the operations in R are associative, for all $x \in X$, we have

$$((f+g)+h)(x) = (f+g)(x)+h(x) = (f(x)+g(x))+h(x) = f(x)+(g(x)+h(x)) = f(x)+(g+h)(x) = (f+(g+h))(x),$$

and

$$((fg)h)(x) = (fg)(x)h(x) = (f(x)g(x))(h(x)) = f(x)(g(x)h(x)) = f(x)((gh)(x)) = (f(gh))(x).$$

The zero element of R^X is the zero function $0(x) = 0$ for all $x \in X$, and the additive inverse $-f$ of $f \in R^X$ is given by $(-f)(x) = -f(x)$ for all $x \in X$. We leave the rest of the verification that R^X is a ring to Exercise 2.6.9.

Example 2.6.26. Suppose V is a vector space over a field $\mathbb{F}$. We claim that pointwise addition and composition make the set of linear transformations $\mathcal{L}(V, V)$ into a ring with 1. First, just as in Example 2.6.25, we can easily check that pointwise addition makes $\mathcal{L}(V, V)$ into a group; see Exercise 2.6.10. Since $\mathcal{L}(V, V)$ consists of functions and "multiplication" is composition, Proposition 1.1.3 shows that multiplication (i.e. composition) is associative. All that remains is to prove that composition distributes over addition. This is where the linearity comes in. Indeed, suppose $S, T, U \in \mathcal{L}(V, V)$ and $v \in V$. Then we have

$$(S(T+U))(v) = S(T(v) + U(v)) = S(T(v)) + S(U(v)) = (ST + SU)(v),$$

and

$$((T+U)S)(v) = (T+U)(S(v)) = T(S(v)) + U(S(v)) = (TS + US)(v).$$

It is interesting to note that, in fact, only the first of these equations used linearity of the functions.

Note that $\mathcal{L}(V, V)$ is a ring with 1: indeed, $1 = \mathrm{id}_V\colon V \to V$, the identify function since $\mathrm{id}_V\, S = S\, \mathrm{id}_V = S$ for all $S \in \mathcal{L}(V, V)$. In general, this is not a commutative ring; see Exercise 2.6.11.

If $n \geq 1$ and $\mathbb{F}$ is a field, then $M_{n\times n}(\mathbb{F})$ is also a ring with matrix addition and matrix multiplication. Indeed, as in Section 2.4, with respect to the standard basis for $\mathbb{F}^n$, linear transformations uniquely determine (and are determined by) $n \times n$ matrices so that pointwise addition corresponds to matrix addition and composition corresponds to matrix multiplication (see Proposition 2.4.19).

If R is a ring, then a nonempty subset $S \subset R$ is called a **subring** if

1. S is closed under both operations: $a + b \in S$ and $ab \in S$, for all $a, b \in S$, and
2. the additive inverse of every element of S is in S: $-a \in S$, for all $a \in S$.

Note that a subring is a subgroup under addition by Proposition 2.6.8. By assumption, multiplication is an operation on S, and since the other properties of a ring are satisfied in R, hence also in S, we have the following.

Proposition 2.6.27. *If $S \subset R$ is a subring, then the operations on R make S into a ring. If R is commutative, so is S.*

We could in fact use Proposition 2.6.27 as the definition of a subring (just as we could use Proposition 2.6.8 as the definition of a subgroup).

Example 2.6.28. Subfields of a field are subrings. More interestingly, since $\mathbb{Z}$ is a ring, we see that $\mathbb{Z} \subset \mathbb{Q}$ is a subring: so the subring of a field need not be a field. The even integers $2\mathbb{Z}$ also form a subring of $\mathbb{Z}$, since the sum and product of even integers is even, as is the negative of an even integer. This is an interesting example as is shows that a subring of a ring with 1 need not have a 1. (On the other hand, the subset $S = \{0\}$ in any ring is always a subring, so this is perhaps not surprising.)

Example 2.6.29. The even degree polynomials (those for which the coefficients of odd powers of x are zero) form a subring of all polynomials $\mathbb{F}[x]$ over any field $\mathbb{F}$. To see this, we note that the sum and product of even degree polynomials is again even, and the negative of an even degree polynomial is also even degree. The odd degree polynomials do not: for example, the product of x and x is x^2, so two odd degree polynomials can be an even degree polynomial.

Example 2.6.30. Consider the continuous functions $C(\mathbb{R},\mathbb{R}) \subset \mathbb{R}^{\mathbb{R}}$ from the real numbers to itself. Since sums, products, and scalar multiples of continuous functions are continuous (in particular $-f$ is continuous if f is), we see that $C(\mathbb{R},\mathbb{R})$ is a subring of $\mathbb{R}^{\mathbb{R}}$. Similarly, continuously differentiable functions form a subring $C^1(\mathbb{R},\mathbb{R})$ of $C(\mathbb{R},\mathbb{R})$ (and hence also of $\mathbb{R}^{\mathbb{R}}$), as do twice differentiable, k-times differentiable, and infinitely differentiable functions:

$$\mathbb{R}^{\mathbb{R}} \supset C(\mathbb{R},\mathbb{R}) \supset C^1(\mathbb{R},\mathbb{R}) \supset C^2(\mathbb{R},\mathbb{R}) \supset \cdots \supset C^\infty(\mathbb{R},\mathbb{R}).$$

This also generalizes to subrings of continuously differentiable, and continuous functions on Euclidean space: $C(\mathbb{R}^n,\mathbb{R}) \supset C^1(\mathbb{R}^n,\mathbb{R}) \supset \cdots$.

Exercises.

Exercise 2.6.1. Prove that for all $n \geq 1$, $(\mathbb{Z}_n, +)$ is a group.

Exercise 2.6.2. Prove that for all $n \geq 1$, $(\mathbb{Z}_n^\times, \cdot)$ is a group.

Exercise 2.6.3. Prove Proposition 2.6.8.

Exercise 2.6.4. Prove that the roots of unity C_n defined in Example 2.3.18 form a subgroup of the group S^1 from Example 2.6.12.

Exercise 2.6.5. Prove that if $\mathbb{F}$ is a field, then $\mathrm{Aut}(\mathbb{F})$, defined in Example 2.6.16, is a subgroup of $\mathcal{S}ym(\mathbb{F})$.

Exercise 2.6.6. Prove that D_4 consists of exactly the eight elements listed in Example 2.6.18. For this, you may assume that the "center" of the square (the intersection of the diagonals) is the origin as illustrated in the figure, and that it is fixed by all elements of D_4 (though this is not that difficult to prove).

Exercise 2.6.7. Suppose $H, K < G$ are subgroups. Prove that $H \cap K$ is also a subgroup of G.

Exercise 2.6.8. Prove that $(\mathbb{Z}_n, +, \cdot)$ is a commutative ring with 1.

Exercise 2.6.9. Suppose R is a ring and X is a nonempty set. Complete the proof that R^X forms a ring by proving (a) that pointwise addition on R^X is commutative, (b) 0 is an additive identity, (c) $-f$ is the additive inverse of any $f \in R^X$ (and so $(R^X, +)$ is an abelian group), and (d) multiplication distributes over addition. If R is a commutative ring, prove that R^X is a commutative ring. If R has 1, prove that the function $1(x) = 1$ is a 1 for R^X.

Exercise 2.6.10. Suppose V is a vector space over a field $\mathbb{F}$. Prove that pointwise addition makes $\mathcal{L}(V,V)$ into an abelian group.

Exercise 2.6.11. Suppose $\mathbb{F}$ is any field. Find a pair of linear transformations $S, T \in \mathcal{L}(\mathbb{F}^2, \mathbb{F}^2)$ such that $ST \neq TS$.

You should...

- Be able to do all the exercises from this section.
- Know the examples of groups and subgroups from this section—they will be referred to often later.

- Be able to decide when a set with an operation is a group, and when a subset is a subgroup.
- Know what it means for a group to be abelian, or for it to be nonabelian.
- Be able to decide when a set with two operations is a ring, and when a subset is a subring.
- Know what a commutative ring is, and what a ring with 1 is.

Chapter 3

Group Theory

Section 2.6 provides the definition of a group and a subgroup, as well as numerous examples. Here we begin a systematic study of groups.

3.1 Basics

In this section, we investigate some of the "nuts-and-bolts" of group theory. We will provide some of the proofs in this section (and the next few), but leave the others as exercises for the reader. Doing these exercises is an important part of becoming comfortable with the abstract concepts involved in group theory.

Proposition 3.1.1. † *Let $(G, *)$ be a group with identity e.*

(i) *If $g, h \in G$ and either $g * h = h$ or $h * g = h$, then $g = e$.*

(ii) *If $g, h \in G$ and $g * h = e$ then $g = h^{-1}$ and $h = g^{-1}$.*

The proposition says that the identity in a group is unique, as are inverses.

Proof. For part (i), suppose $g * h = h$. Then using the defining properties of a group, we have

$$g = g * e = g * (h * h^{-1}) = (g * h) * h^{-1} = h * h^{-1} = e.$$

Specifically, the first equality follows from the definition of the identity, the second from the definition of inverses, the third by associativity, the fourth by assumption, and the last by definition of inverses again. A similar calculation proves that if $h * g = h$, then $g = e$. Part (ii) is left as Exercise 3.1.1. □

From this, we obtain the following.

Corollary 3.1.2. *Let $(G, *)$ be a group with identity e, and let $g, h \in G$. Then $e = e^{-1}$, $(g^{-1})^{-1} = g$, and $(g * h)^{-1} = h^{-1} * g^{-1}$.*

Proof. Since $e^{-1} * e = e$, Proposition 3.1.1 part (i) shows that $e = e^{-1}$. Next, observe that $g * g^{-1} = e$, and so by Proposition 3.1.1 part (ii), g must be the inverse of g^{-1}, that is, $g = (g^{-1})^{-1}$. For the last part, we compute

$$(g * h) * (h^{-1} * g^{-1}) = (g * (h * h^{-1})) * g^{-1} = (g * e) * g^{-1} = g * g^{-1} = e.$$

By Proposition 3.1.1 part (ii), $h^{-1} * g^{-1} = (g * h)^{-1}$, as required. □

We also have "cancellation" for equations in groups.

Proposition 3.1.3. *Let $(G, *)$ be a group and $g, h, k \in G$. If $g * h = k * h$, then $g = k$. Likewise, if $h * g = h * k$, then $g = k$.*

Proof. From the assumption, we have

$$g = g * e = g * (h * h^{-1}) = (g * h) * h^{-1} = (k * h) * h^{-1} = k * (h * h^{-1}) = k * e = k.$$

The second statement is similar. □

The next fact is a generalization of Proposition 3.1.1 part (ii).

Proposition 3.1.4. *Let $(G, *)$ be a group and $g, h \in G$. Then the equations $x * g = h$ and $g * x = h$ have unique solutions $x \in G$.*

Proof. For the first equation, $x = h * g^{-1}$ is clearly a solution. If $x = k$ is any other solution, then we have $k * g = (h * g^{-1}) * g$, and so by Proposition 3.1.3, $k = h * g^{-1}$, so the solution is unique. The proof for the second case is handled similarly. □

Suppose that $(G, *)$ is a group, $g \in G$, and $n \in \mathbb{Z}$, and we wish to define g^n. We define this recursively for $n \geq 0$ by setting $g^0 = e$ and for $n \geq 1$, we set $g^n = g^{n-1} * g$. For $n \leq 0$, we define $g^n = (g^{-1})^{|n|}$.

Proposition 3.1.5. *Let $(G, *)$ be a group, $g \in G$ and $n, m \in \mathbb{Z}$. Then*

(i) $g^n * g^m = g^{n+m}$.

(ii) $(g^n)^m = g^{nm}$.

Proof. We prove part (i), and leave part (ii) as Exercise 3.1.2. Suppose $n \geq 0$ is some fixed integer. We prove that for any $m \in \mathbb{Z}$, $m \geq 0$, we have $g^n * g^m = g^{n+m}$. We do this by induction on m. For the base case $m = 0$, we have $g^n * e = g^n$, which holds by definition of the identity $e \in G$. Next, we assume that $g^n * g^k = g^{n+k}$ for $k = m - 1 \geq 0$, and prove that this also holds for $k = m$. Again, we compute

$$g^n * g^m = g^n * (g^{m-1} * g) = (g^n * g^{m-1}) * g = g^{n+m-1} * g = g^{n+m}.$$

We have used the inductive hypothesis for the third equality, while the last equality just follows from the definition of g^{n+m}.

Therefore, for all $n, m \in \mathbb{Z}$ we know $g^n * g^m = g^{n+m}$ if $n, m \geq 0$. By replacing g with g^{-1} in the proof above, we see that $g^n * g^m = g^{n+m}$ if $n, m \leq 0$. Now suppose $n \geq 0$ and we prove that $g^{-n} = (g^n)^{-1}$ by induction on n. The base case is the statement that $e = e^{-1}$. Supposing the claim is true for $n - 1 \geq 0$, we prove it for n. Applying induction, and what we've already proved, we have

$$\begin{aligned} g^{-n} * g^n &= (g^{-1})^n * g^n = ((g^{-1})^{n-1} * g^{-1}) * (g * g^{n-1}) = (g^{-1})^{n-1} * ((g^{-1} * g) * g^{n-1}) \\ &= (g^{-1})^{n-1} * (e * g^{n-1}) = (g^{-1})^{n-1} * g^{n-1} = e. \end{aligned}$$

By Proposition 3.1.1 part (ii), we see that $g^{-n} = (g^n)^{-1}$.

Finally, suppose that $m, n \in \mathbb{Z}$ are arbitrary, and we prove that $g^n * g^m = g^{n+m}$. There are four cases depending on the signs of m and n, though we have already proved the cases that the signs are the same. We complete the proof for the case $n > 0$ and $m < 0$, and leave the case $n < 0$ and $m > 0$ to the reader (which is nearly identical). Define $k = n + m$. If $k \geq 0$, then $k, -m, n \geq 0$ and we have

$$g^n * g^m = g^{k-m} * g^m = (g^k * g^{-m}) * g^m = g^k * (g^{-m} * g^m) = g^k * e = g^k = g^{n+m}.$$

If $k \leq 0$, then $k, m, -n \leq 0$ and

$$g^n * g^m = g^n * g^{k-n} = g^n * (g^{-n} * g^k) = (g^n * g^{-n}) * g^k = e * g^k = g^k = g^{n+m}.$$

□

It is important to remember that g^n is defined using the operation $*$ of the group. The next example illustrates how this is interpreted for some of the groups we are familiar with.

Example 3.1.6. In the group $(\mathbb{Z}, +)$, we note that the identity is 0, and for $g \in \mathbb{Z}$ and $n \in \mathbb{Z}$ we have $g^n = ng$. For example, in the group $\mathbb{Z}$, $4^3 = 4 + 4 + 4 = 3(4) = 12$. The same is true in $(\mathbb{Q}, +)$, $(\mathbb{R}, +)$, and $(\mathbb{C}, +)$, for example. Similarly, in **the group** $(\mathbb{Z}_n, +)$, we have $[a]^n = [na]$.

These examples illustrate an important point: When we prove statements about g^n in a group G, we must make the translation into the terminology/notation of the given group.

We end this section with a simple method for constructing new groups from old. As we learn more about groups, we will learn other constructions as well.

Suppose $(G, *)$ and $(H, \star)$ are two groups. Define an operation $\circledast$ on $G \times H$ by

$$(g, h) \circledast (g', h') = (g * g', h \star h').$$

Since $*$ and $\star$ are operations on G and H, respectively, so $\circledast$ is an operation on $G \times H$.

Proposition 3.1.7. *For any group $(G, *)$ and $(H, \star)$, $\circledast$ makes $G \times H$ into a group. Furthermore, if e_G and e_H are the identities in G and H, respectively, then (e_G, e_H) is the identity in $G \times H$. The inverse of $(g, h) \in G \times H$ is (g^{-1}, h^{-1}).*

The group $G \times H$ (with the operation $\circledast$) is called the **direct product** of G and H. Later we will see that there are sometimes other ways to define a group operation on $G \times H$ from the operations on G and H, giving rise to other kinds of "products".

Before we give the proof, we note how cumbersome the notation $*$, $\star$, and $\circledast$ is. From this point forward, we omit the operation in the notation for abstract groups, writing $gg' = g * g'$ in the group G (and likewise $hh' = h \star h'$ in H, and $(g, h)(g', h') = (g, h) \circledast (g', h')$ in $G \times H$). Because the only operation we have given is the group operation, there is no loss of information, and this greatly simplifies the expressions we will use. We will also denote the identity in G by e instead of e_G, and the identity in H by e instead of e_H. This presents some minor confusion because e denotes two different elements (one in G and one in H). However, when we write $(e, h) \in G \times H$ or $(g, e) \in G \times H$, we understand that the in the first case e is the identity in G and in the second it is the identity in H. More generally, we will always provide appropriate context so there should be no confusion.

Proof. We must show that the operation is associative. For that, we take $(g, h), (g', h'), (g'', h'') \in G \times H$, and compute:

$$\begin{aligned}(g, h)((g', h')(g'', h'')) &= (g, h)(g'g'', h'h'') = (g(g'g''), h(h'h'')) \\ &= ((gg')g'', (hh')h'') = (gg', hh')(g'', h'') \\ &= ((g, h)(g', h'))(g'', h'').\end{aligned}$$

Therefore, the operation is associative. Next we prove that (e, e) serves as the identity in $G \times H$. For that, let $(g, h) \in G \times H$ and compute:

$$(g, h)(e, e) = (ge, he) = (g, h) = (eg, eh) = (e, e)(g, h).$$

Thus, $G \times H$ has an identity element (e, e). We leave the proof that any element $(g, h) \in G \times H$ has an inverse, in fact given by (g^{-1}, h^{-1}), as Exercise 3.1.5. □

Exercises.

Exercise 3.1.1. Prove part (ii) of Proposition 3.1.1.

Exercise 3.1.2. Prove part (ii) of Proposition 3.1.5.

Exercise 3.1.3. Suppose that G is a nonempty set with an associative operation $*$ such that the following holds:

1. There exists an element $e \in G$ so that $e * g = g$ for all $g \in G$, and
2. for all $g \in G$, there exists an element $g^{-1} \in G$ so that $g^{-1} * g = e$.

Prove that $(G, *)$ is group.

The difference between this and the definition of a group is that we are only assuming that e is a "left identity", and that elements have a "left inverse". Of course, we could have replaced "left" with "right" and there is an analogous equivalent definition of group. *Hint: Start by proving that if $g \in G$ and $g * g = g$, then $g = e$. Then prove that $g * g^{-1} = e$ (that is, the left inverse is also a right inverse for the left identity). Finally, prove that the left identity is also a right identity.*

Exercise 3.1.4. Prove that a nonempty set G with an associate operation $*$ is a group if and only if the equations $g * x = h$ and $x * g = h$ have solutions in G for all $g, h \in G$. *Hint: Prove that if $e * g = g$ for some g, then $e * h = h$ for all $h \in G$. Now appeal to Exercise 3.1.3.*

Exercise 3.1.5. Complete the proof of Proposition 3.1.7 by proving that for all $(g, h) \in G \times H$, (g^{-1}, h^{-1}) serves as an inverse for (g, h).

Exercise 3.1.6. Let G and H be groups with identities, both denoted e. Prove that for all $g \in G$ and $h \in H$, (g, e) and (e, h) commute. That is, prove that

$$(g, e)(e, h) = (e, h)(g, e).$$

You should...

- Be able to do all the exercises from this section and know the proofs of statements marked by †.
- Be able to formally manipulate inside a group using only the axioms.
- Know what g^n means if G is **any** group, $g \in G$, and $n \in \mathbb{Z}$.
- Understand the direct product construction.

3.2 Subgroups and cyclic groups

Subgroups provide a useful source of examples of groups (see Proposition 2.6.8). Here we prove a few basic results about subgroups.

Remark 3.2.1. From this point forward, we will typically leave out mention of the operation in a group G, writing gh for all $g, h \in G$. We sometimes call gh the **product** of g and h, regardless of what the operation on G.

In Exercise 2.6.7 you proved that the intersection of two subgroups of a group is again a subgroup. A more general version of that statement is given by the next proposition.

Proposition 3.2.2. *Let G be a group and suppose $\mathcal{H}$ is any collection of subgroups of G. Then*

$$K = \bigcap_{H \in \mathcal{H}} H < G$$

is a subgroup.

Proof. Observe that $e \in H$ for all $H \in \mathcal{H}$, and hence $K \neq \emptyset$. If $g, h \in K$, then $g, h \in H$ for all $H \in \mathcal{H}$, and since each H is a subgroup, g^{-1} and gh are in H for each H, which implies $g^{-1}, gh \in K$. It follows that K is a subgroup, as required. □

Using this proposition, any *subset* $A \subset G$ determines a subgroup called the **subgroup generated by** A, defined by

$$\langle A \rangle = \bigcap_{H \in \mathcal{H}(A)} H$$

where $\mathcal{H}(A)$ is the set of all subgroups of G containing the set A:

$$\mathcal{H}(A) = \{H < G \mid A \subset H \text{ and } H \text{ is a subgroup of } G\}.$$

For any $g \in G$, we apply this to the set $\{g\}$ consisting of the single element g, and we write $\langle g \rangle = \langle \{g\} \rangle$. In this case, we call $\langle g \rangle$ the **cyclic subgroup generated by** g. If G is a group and there exists an element $g \in G$ so that $\langle g \rangle = G$, then we say that G is a **cyclic group** and that g is a generator. Cyclic groups have very simple descriptions.

Proposition 3.2.3. *Let G be a group and $g \in G$. Then*

$$\langle g \rangle = \{g^n \mid n \in \mathbb{Z}\}.$$

Proof. Let $H = \{g^n \mid n \in \mathbb{Z}\} \subset G$. We first prove that H is a subgroup. For this, observe that $e = g^0 \in H$, so $H \neq \emptyset$. For any two elements $g^n, g^m \in H$, Proposition 3.1.5 implies that $g^n g^m = g^{n+m} \in H$ and $(g^n)^{-1} = g^{-n} \in H$. Therefore, H is a subgroup. Since $g \in H$, it follows from the definition that $\langle g \rangle \subset H$. On the other hand, given $g^n \in H$, an inductive argument proves that any subgroup of G containing g must contain g^n. Therefore, g^n is in every subgroup containing g, and hence $g^n \in \langle g \rangle$, proving $\langle g \rangle = H$. □

Corollary 3.2.4. *If G is a cyclic group, then G is abelian.*

Recall that a group G is abelian if the operation is commutative. That is, for all $g, h \in G$, we have $gh = hg$.

Proof. Let $g \in G$ be a generator, so that $G = \langle g \rangle$. By Proposition 3.2.3, $G = \{g^n \mid n \in \mathbb{Z}\}$, so any two elements of G have the form g^n, g^m, for $m, n \in \mathbb{Z}$. By Proposition 3.1.5 we have

$$g^n g^m = g^{n+m} = g^{m+n} = g^m g^n.$$

□

Example 3.2.5. Recall that in the *group* $\mathbb{Z}$ (with addition) we have $a^n = na$. Therefore, if we consider $\langle 1 \rangle$, the subgroup generated by 1, we have

$$\langle 1 \rangle = \{1^n \mid n \in \mathbb{Z}\} = \{n1 \mid n \in \mathbb{Z}\} = \{n \mid n \in \mathbb{Z}\} = \mathbb{Z}.$$

So we see that $\mathbb{Z}$ is cyclic, and 1 is a generator.

Given a group G, the **order of** G is its cardinality $|G|$. As with sets, we will be primarily interested in the distinction between finite groups and infinite groups, and we often write $|G| < \infty$ and $|G| = \infty$, for these two situations, respectively. If $g \in G$ then we write $|g| = |\langle g \rangle| \in \mathbb{Z}_+ \cup \{\infty\}$ and call this the **order of** g. The order of an element g has the following variety of interpretations.

Proposition 3.2.6. *Let G be a group. For $g \in G$, the following are equivalent:*

(i) $|g| < \infty$.

(ii) *There exist integers $n \neq m$ so that $g^n = g^m$.*

(iii) *There exists an integer $n \neq 0$ so that $g^n = e$.*

(iv) *There exists an integer $n > 0$ so that $g^n = e$.*

If any (hence every) one of the conditions is satisfied, then $|g|$ is the smallest positive integer n so that $g^n = e$, and

$$\langle g \rangle = \{e, g, g^2, \ldots, g^{n-1}\}.$$

Proof. We first prove (i)⇒(ii), so suppose $|g| < \infty$. Since $\mathbb{Z}$ is infinite, Proposition 3.2.3 means that there must be two integers $n \neq m$ so that $g^n = g^m$ (otherwise the function $\mathbb{Z} \to \langle g \rangle$ given by $n \mapsto g^n$ would be injective, and hence $\infty = |\mathbb{Z}| \leq |\langle g \rangle| = |g| < \infty$, a contradiction). Next, we prove (ii)⇒(iii). Let $n, m \in \mathbb{Z}$, $n \neq m$, be such that $g^n = g^m$. Then $n - m \neq 0$ and $g^{n-m} = g^n g^{-m} = e$, as required. To prove (iii)⇒(iv), we suppose $g^n = e$ for some $n \neq 0$. Now observe that $g^n = e$ implies $g^{-n} = (g^n)^{-1} = e^{-1} = e$. Since either n or $-n$ is positive, this suffices.

Finally, we assume that (iv) holds, and let $n \geq 0$ be the smallest positive integer such that $g^n = e$. We prove that (i) holds, and along the way, prove that $n = |g|$, and that $\langle g \rangle = \{e, g, g^2, \ldots, g^{n-1}\}$. For this, suppose $m \in \mathbb{Z}$ and let $q, r \in \mathbb{Z}$ be as in Proposition 1.4.7 so that $m = qn + r$ and $0 \leq r < n$. Then by Proposition 3.1.5 we have

$$g^m = g^{qn+r} = (g^n)^q g^r = e^q g^r = e g^r = g^r.$$

Therefore, $\langle g \rangle = \{e, g, g^2, \ldots, g^{n-1}\}$. In particular, $|g| \leq n < \infty$, proving (i). If $|g| < n$, then for some $0 \leq i < j \leq n - 1$ we must have $g^j = g^i$. But then $0 < j - i \leq n - 1$, and $g^{j-i} = e$, contradicting the minimality of n. Therefore, $|g| = n$, completing the proof. □

Example 3.2.7. Consider the group $(\mathbb{Z}_n, +)$. This is cyclic, generated by $[1]$,

$$\langle [1] \rangle = \{[a1] \mid a \in \mathbb{Z}\} = \{[a] \mid a \in \mathbb{Z}\} = \mathbb{Z}_n,$$

and $|[1]| = n$.

Example 3.2.8. For each integer $n > 0$, let $\zeta_n = e^{2\pi i/n} \in \mathbb{C}^\times$. Viewing $\mathbb{C}^\times$ as a group with multiplication, we observe that as in Example 2.3.18

$$\langle \zeta_n \rangle = \{e^{2\pi i k/n} \mid k \in \mathbb{Z}\} = C_n,$$

is the subgroup of n^{th} roots of unity (in Exercise 2.6.4 you proved that this is a subgroup). So, C_n is cyclic with generator ζ_n.

The groups $\mathbb{Z}$ and $\mathbb{Z}_n$ (both with addition, as always) are our "prototypes" for cyclic groups. Later we will see how to transfer information about these two groups to any other cyclic group. One such piece of information is a complete description of all subgroups of these groups, starting with $\mathbb{Z}$.

Theorem 3.2.9. † *If $H < \mathbb{Z}$ is a subgroup, then either $H = \{0\}$, or else $H = \langle d \rangle$, where*

$$d = \min\{h \in H \mid h > 0\}.$$

Consequently, $a \mapsto \langle a \rangle$ defines a bijection from $\mathbb{N} = \{0, 1, 2, \ldots\}$ to the set of subgroups of $\mathbb{Z}$. Furthermore, for $a, b \in \mathbb{Z}_+$, we have $\langle a \rangle < \langle b \rangle$ if and only if $b|a$.

The following proof should look familiar (compare to the proof of Proposition 1.4.8).

Proof. Suppose $H \neq \{0\}$. So there exists $a \in H$, $a \neq 0$, and since H is a subgroup, $-a \in H$ as well. Consequently, there is a positive element in H, and we let d be the smallest such. Since H is a subgroup containing d, we see that $\langle d \rangle < H$. To prove the other containment, suppose $h \in H$, and appealing to Proposition 1.4.7, let $q, r \in \mathbb{Z}$ be such $0 \leq r < d$ and $h = qd + r$. Then since $d \in H$, we have $-qd \in H$, and so $r = h - qd \in H$. But since $0 \leq r < d$, and d was the small *positive* element of H, we must have $r = 0$. In this case, $h = qd$, and so $h \in \langle d \rangle$, as required. Therefore, $\langle d \rangle = H$.

Suppose $a, b \in \mathbb{N}$ and that $\langle a \rangle < \langle b \rangle$. Then $a \in \langle b \rangle$, and hence $a = nb$ for some $n \in \mathbb{Z}_+$. Consequently, $b|a$. Conversely, if $b|a$, then any $na \in \langle a \rangle$ is also divisible by b, and hence lies in $\langle b \rangle$. That is, $\langle a \rangle < \langle b \rangle$, completing the proof. □

Theorem 3.2.10. † *For any $n \geq 2$, if $H < \mathbb{Z}_n$ is a subgroup, then there is a positive divisor d of n so that*

$$H = \langle [d] \rangle.$$

Furthermore, this defines a bijection between divisors of H and subgroups of $\mathbb{Z}_n$. Furthermore, if $d, d' > 0$ are two divisors of n, then $\langle [d] \rangle < \langle [d'] \rangle$ if and only if $d' | d$.

Remark 3.2.11. Note that as an immediate consequence of this theorem, we see that for any subgroup $H < \mathbb{Z}_n$, we have $|H| \Big| n$, since $|H| = |\langle [d] \rangle| = n/d$. We do not state this as a corollary since *Lagrange's Theorem* (Theorem 3.5.6 below) provides a much more general theorem.

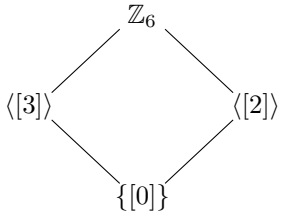
This proof mirrors the proof of Theorem 3.2.9.

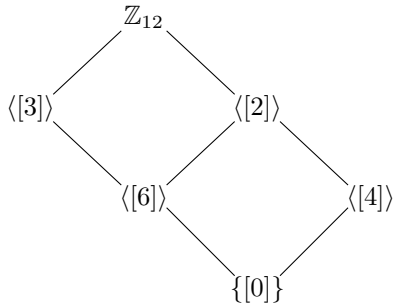
Proof. Let $d > 0$ be the smallest positive integer so that $[d] \in H$. Note that if $H = \{[0]\}$, then $d = n$. We clearly have $\langle [d] \rangle < H$. Suppose $[a] \in H$, and let $q, r \in \mathbb{Z}$ be such that $0 \leq r < d$ and $a = qd + r$. Then $[a - qd] = [r] \in H$ and by minimality of d, $r = 0$. So $[a] = [qd] \in \langle [d] \rangle$, so $\langle [d] \rangle = H$.

We leave the last statement as Exercise 3.2.1. □

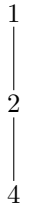
The set of all subgroups of a group of G, together with the data of which subgroups contain which others is called the **subgroup lattice**. We often picture the subgroup lattice in a diagram with the entire group at the top, the trivial subgroup $\{e\}$ at the bottom, and the intermediate subgroups in the middle, with lines drawn from subgroups *up* to larger groups.

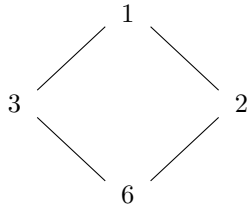
Example 3.2.12. Here we show the subgroup lattices for $\mathbb{Z}_4$, $\mathbb{Z}_6$, and $\mathbb{Z}_{12}$.

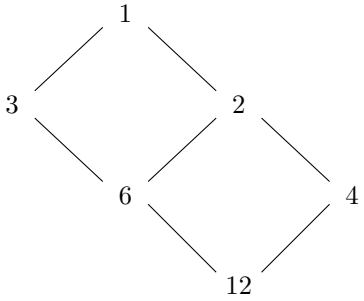




Writing down the subgroup lattice is as easy as writing down the divisibility lattice in which n is placed at the bottom, 1 at the top, and all intermediate divisors in between, connected by edges when there is divisibility. The congruence class of the divisor generates the corresponding subgroup in the subgroup lattice.







We will see that there are certain kinds of subgroups in any group. We list two of them now. For this, let G be any group. The **center of** G is the subset defined by

$$Z(G) = \{g \in G \mid gh = hg \text{ for all } h \in H\}.$$

That is, the center consists of all elements that commute with every element in the group. Note that $Z(G) = G$ if and only if G is abelian. Similarly, if $g \in G$, we define the **centralizer of** g **in** G by

$$C_G(g) = \{h \in G \mid gh = hg\}.$$

That is, $C_G(g)$ is the set of all elements that commute with the element g.

Proposition 3.2.13. *For any group G and any $g \in G$, $C_G(g)$ is a subgroup of G. Furthermore*

$$Z(G) = \bigcap_{g \in G} C_G(g).$$

In particular, $Z(G)$ is a subgroup of G.

Proof. First observe that $eg = g = ge$, so $e \in C_G(g)$, and hence $C_G(g) \neq \emptyset$. For any $h, h' \in C_G(g)$, we have

$$(hh')g = h(h'g) = h(gh') = (hg)h' = (gh)h' = g(hh'),$$

so that $hh' \in C_G(g)$. Next, note that from the equation $hg = gh$, we can "multiply" on the left and right by h^{-1} to obtain

$$h^{-1}hgh^{-1} = h^{-1}ghh^{-1} \quad \Rightarrow \quad gh^{-1} = h^{-1}g.$$

Therefore, $h^{-1} \in C_G(g)$, so $C_G(g)$ is a subgroup. Since $Z(G)$ consists of the elements of G that commute with every $g \in G$, this is precisely the intersection of all centralizers $C_G(g)$, over all $g \in G$, as stated. That $Z(G)$ is a subgroup then follows from Proposition 3.2.2. □

Exercises.

Exercise 3.2.1. Suppose $n \geq 2$ is an integer and $d, d' > 0$ are two divisors of n. Prove that $\langle [d] \rangle < \langle [d'] \rangle$ if and only if $d' | d$.

Exercise 3.2.2. Prove that the number of elements of order n in $\mathbb{Z}_n$ is exactly $\varphi(n)$, the Euler phi function of n. *Hint: You need to decide which $[a] \in \mathbb{Z}_n$ generate $\mathbb{Z}_n$.*

Exercise 3.2.3. Draw the subgroup lattice for the group $\mathbb{Z}_8$, $\mathbb{Z}_{15}$, $\mathbb{Z}_{24}$, and $\mathbb{Z}_{30}$.

Exercise 3.2.4. Draw the subgroup lattice for S_3 (a group with respect to composition $\circ$). You will need to find all the subgroups $H < S_3$ by hand (because we don't yet have any theorems that tell us what the subgroups of S_3 are). *Hint: There are exactly 6 subgroups, but you should verify this by proving that there are no other subgroups than the ones you have listed.*

Exercise 3.2.5. Prove that if G and H are groups and $K < G, J < H$ are subgroups, then $K \times J \subset G \times H$ is a subgroup. Construct an example of a subgroup of $\mathbb{Z}_2 \times \mathbb{Z}_2$ which is **not** of the form $K \times J$ for some $K < \mathbb{Z}_2$ and $J < \mathbb{Z}_2$.

Exercise 3.2.6. Consider the group S_3 (with the operation $\circ$, as usual). Find $C_{S_3}((1\ 2))$, $C_{S_3}((1\ 2\ 3))$, and $Z(S_3)$.

You should...

- Be able to do all the exercises from this section and know the proofs of statements marked by †.
- Know what the order of a group is and what the order of an element is. Understand the relationship with cyclic groups.
- Know what a generator for a cyclic group is.
- Understand what all the subgroups of $\mathbb{Z}$ and $\mathbb{Z}_n$ are. Be able to construct the subgroup lattice for $\mathbb{Z}_n$.
- Know what the centralizer of an element is and what the center of a group is.

3.3 Homomorphisms, isomorphisms, and normal subgroups

Recall that in linear algebra, the objects are vector spaces, and that the maps between vector spaces that we study are linear transformations. These are maps that *preserve the vector space structure*. This is a common theme in abstract algebra. For every type of abstract algebraic object, we are often interested in the maps between them that preserve the structure. For groups, this takes the following form.

Definition 3.3.1. *Suppose G, H are groups. A map $\phi\colon G \to H$ is a* **group homomorphism** *(or simply a* **homomorphism***) if for all $g, g' \in G$, we have*

$$\phi(gg') = \phi(g)\phi(g').$$

We emphasis that in this definition, both the operation in G and the operation in H are involved. Specifically, $g, g' \in G$, and gg' is the result of combining g and g' using the operation in G. On the other hand, $\phi(g), \phi(g') \in H$, and $\phi(g)\phi(g')$ is the result of combining $\phi(g)$ and $\phi(g')$ in H. So, we can think of this as saying that a homomorphism "turns the operation in G into the operation in H".

Remark 3.3.2. The term *homomorphism* is really the general name for maps between abstract algebraic objects preserving the structure, and so we should always say "group homomorphism" when referring to a map as in Definition 3.3.1. However, when there is no other structure involved, there is no loss of clarity referring to the map as simply a homomorphism, and we will follow this practice when convenient.

Example 3.3.3. Recall that $\pi\colon \mathbb{Z} \to \mathbb{Z}_n$ is the map defined by $\pi(a) = [a]$, for all $a \in \mathbb{Z}$. By definition of addition in $\mathbb{Z}_n$, this is a group homomorphism:

$$\pi(a+b) = [a+b] = [a] + [b] = \pi(a) + \pi(b).$$

Likewise, if $m|n$, then $\pi_{m,n}\colon \mathbb{Z}_n \to \mathbb{Z}_m$ defined in §1.5 by $\pi_{m,n}([a]_n) = [a]_m$ is also a group homomorphism since

$$\pi_{m,n}([a]_n + [b]_n) = \pi_{m,n}([a+b]_n) = [a+b]_m = [a]_m + [b]_m = \pi_{m,n}([a]_n) + \pi_{m,n}([b]_n).$$

With those basic examples in mind, we proceed to some simple observations about homomorphisms.

Proposition 3.3.4. *Suppose G and H are groups, and that e denotes the identity (in each). If $\phi\colon G \to H$ is a group homomorphism then $\phi(e) = e$. Furthermore, for all $g \in G$ and $n \in \mathbb{Z}$, we have $\phi(g)^n = \phi(g^n)$.*

We emphasize that in the equation $\phi(e) = e$, the e on the right is in H, while the one on the left is in G.

Proof. Let $g \in G$, then we have $\phi(g) = \phi(eg) = \phi(e)\phi(g)$. By uniqueness of the identity (Proposition 3.1.1 part (i)), we see that $\phi(e) = e$.

For the second statement, we first observe that $e = \phi(e) = \phi(gg^{-1}) = \phi(g)\phi(g^{-1})$, and so by uniqueness of inverses (Proposition 3.1.1 part (ii)), we have $\phi(g)^{-1} = \phi(g^{-1})$. So, if we can show that for all $g \in G$ and $n > 0$, we have $\phi(g^n) = \phi(g)^n$, then for $n < 0$, we have

$$\phi(g^n) = \phi((g^{-1})^{|n|}) = \phi(g^{-1})^{|n|} = (\phi(g)^{-1})^{|n|} = \phi(g)^n.$$

We leave the verification for $n > 0$ as Exercise 3.3.1. □

Example 3.3.5. Suppose G is a group and $g \in G$. Then we define

$$\phi\colon \mathbb{Z} \to G$$

by the formula $\phi(n) = g^n$. According to Proposition 3.2.3 we have

$$\phi(n+m) = g^{n+m} = g^n g^m = \phi(n)\phi(m).$$

and so ϕ is a homomorphism. According to Proposition 3.3.4, every homomorphism from $\mathbb{Z}$ into a group G has this form. That is, any homomorphism from $\mathbb{Z}$ into a group G is determined by $\phi(1)$.

Proposition 3.3.6. *Suppose $\phi\colon G \to H$ and $\psi\colon H \to K$ are group homomorphisms. Prove that $\psi \circ \phi\colon G \to K$ is a group homomorphism.*

Proof. This is left as Exercise 3.3.2. □

You may recall from linear algebra that the kernel of a linear transformation (in matrix terminology, this is the null space) is the set of vectors that are sent to zero by the transformation. This is a subspace, and the transformation is injective precisely when the kernel is $\{\mathbf{0}\}$. The analogous statements also hold for group homomorphisms as we now explain.

Suppose $\phi\colon G \to H$ is a group homomorphism. We define the **kernel of** ϕ to be the subset

$$\ker(\phi) = \{g \in G \mid \phi(g) = e\},$$

where, as usual, e denotes the identity (in this case, the identity in H).

Proposition 3.3.7. *If $\phi\colon G \to H$ is a group homomorphism, then $\ker(\phi) < G$ is a subgroup.*

Instead of proving this, we prove the following more general fact.

Proposition 3.3.8. † *Suppose $\phi\colon G \to H$ is a group homomorphism. Then for any subgroup $K < G$, $\phi(K)$ is a subgroup of H. If $J < H$ is a subgroup, then $\phi^{-1}(J)$ is a subgroup of G.*

Note that for the trivial subgroup $J = \{e\} < H$, we have $\phi^{-1}(\{e\}) = \ker(\phi)$.

Proof. Let $K < G$ be any subgroup. Since $K \neq \emptyset$, $\phi(K) \neq \emptyset$. For any $\phi(k), \phi(k') \in \phi(K)$, we have

$$\phi(k)\phi(k') = \phi(kk') \in \phi(K) \text{ and } \phi(k)^{-1} = \phi(k^{-1}) \in \phi(K).$$

So, $\phi(K)$ is a subgroup of H.

If $J < H$ is a subgroup, then $\phi(e) = e$, so $e \in \phi^{-1}(J) \neq \emptyset$. For any $g, g' \in \phi^{-1}(J)$, since J is a subgroup, we have

$$\phi(gg') = \phi(g)\phi(g') \in J \text{ and } \phi(g^{-1}) = \phi(g)^{-1} \in J,$$

and therefore $gg', g^{-1} \in \phi^{-1}(J)$. Consequently, $\phi^{-1}(J)$ is a subgroup of G. □

As mentioned, injectivity for homomorphisms can be determined in terms of the kernel, just as in linear algebra.

Proposition 3.3.9. † *Suppose $\phi\colon G \to H$ is a group homomorphism. Then ϕ is injective if and only if $\ker(\phi) = \{e\}$.*

Proof. Since $\phi(e) = e$ for any homomorphism, we see that if ϕ is injective, then $\phi(g) = e = \phi(e)$ implies $g = e$, and hence $\ker(\phi) = \{e\}$. Conversely, suppose $\ker(\phi) = \{e\}$, and let $g, g' \in G$ be such that $\phi(g) = \phi(g')$. Then

$$\phi(g'g^{-1}) = \phi(g')\phi(g)^{-1} = \phi(g)\phi(g)^{-1} = e,$$

and hence $g'g^{-1} \in \ker(e)$. But since $\ker(\phi) = \{e\}$, we see that $g'g^{-1} = e$, i.e. $g' = g$, and so ϕ is injective. □

3.3.1 Normal subgroups

Next we observe a special property for kernels of homomorphism. To state this, we require some additional notation. Suppose G is a group and $A, B \subset G$ are any two nonempty *subsets*. Then we define the **product** of the sets A and B as

$$AB = \{ab \mid a \in A \text{ and } b \in B\}.$$

If $A = \{a\}$ is a one element set, then we will often instead write

$$aB = \{a\}B.$$

Using the associativity, we can similarly define the product of any finite number of nonempty subsets $A_1, A_2, \ldots, A_n \subset G$

$$A_1 A_2 \cdots A_n = \{a_1 a_2 \cdots a_n \mid a_i \in A_i, \text{ for all } i \in \{1, \ldots, n\}\}.$$

Again, if any of these sets are one-element sets, say $A_i = \{a_i\}$, then in the notation we replace $\{a_i\}$ with a_i.

Example 3.3.10. Suppose $H < G$ is a subgroup of a group G. Then note that since $e \in H$ and since H is a subgroup, we have $HH = H$. Likewise, for any $h \in H$, $hH = H = Hh = hHh^{-1}$.

With this we now make an important definition.

Definition 3.3.11. *A subgroup $N < G$ of a group G is said to be a* **normal subgroup** *if for all $g \in G$, we have*

$$gNg^{-1} = N.$$

Here, as above, $gNg^{-1} = \{gng^{-1} \mid n \in N\}$.

We write $N \triangleleft G$ when N is a normal subgroup of G.

We claim that kernels of group homomorphisms are normal subgroups. The following lemma is useful in the proof.

Lemma 3.3.12. *Suppose $N < G$ is a subgroup of a group G. Then N is normal if and only if for all $g \in G$, $gNg^{-1} \subset N$.*

If $H < G$ is a subgroup and $g \in G$, we call gHg^{-1} the **conjugate of H by g**. A conjugate of H is again a subgroup of G: see Exercise 3.3.6.

Proof. If N is normal, then for all $g \in G$, we have $gNg^{-1} = N$. Since containment allows the possibility of equality, we have $gNg^{-1} \subset N$.

Conversely, suppose that for all $g \in G$, we have $gNg^{-1} \subset N$. Then we also have $g^{-1}Ng \subset N$. Then note that $gg^{-1}Ngg^{-1} = N$, and hence

$$N = gg^{-1}Ngg^{-1} \subset gNg^{-1} \subset N.$$

The only way this can hold is if the containments in the middle are in fact equalities. But then the last containment becomes $gNg^{-1} = N$. Since g was arbitrary, we see that N is a normal subgroup. □

Proposition 3.3.13. *For any group homomorphism $\phi\colon G \to H$, $\ker(\phi) \triangleleft G$ is a normal subgroup.*

Proof. Suppose $g' \in \ker(\phi)$ and $g \in G$. Then

$$\phi(gg'g^{-1}) = \phi(g)\phi(g')\phi(g)^{-1} = \phi(g)e\phi(g)^{-1} = \phi(g)\phi(g)^{-1} = e.$$

So, $gg'g^{-1} \in \ker(\phi)$, and thus

$$g\ker(\phi)g^{-1} \subset \ker(\phi).$$

According to Lemma 3.3.12, it follows that $\ker(\phi)$ is a normal subgroup. □

Example 3.3.14. Recall from §3.1 that for any two groups H, K, the direct product $H \times K$ makes the Cartesian product into a group in which the operation is given by $(h,k)(h',k') = (hh',kk')$, for all $(h,k),(h',k') \in H \times K$. The function $\pi_H \colon H \times K \to H$ given by $\pi_H(h,k) = h$ is a homomorphism:

$$\pi_H((h,k)(h',k')) = \pi_H(hh',kk') = hh' = \pi_H(h,k)\pi_H(h',k').$$

Likewise, $\pi_K(h,k) = k$ defines a homomorphism $\pi_K \colon H \times K \to K$. The kernel of π_H is a normal subgroup and is given by

$$\ker(\pi_H) = \{(h,k) \mid k \in K, h = e\} = \{(e,k) \mid k \in K\} = \{e\} \times K,$$

and similarly $\ker(\pi_K) = H \times \{e\}$ (c.f. Exercise 3.2.5).

Normal subgroups are typically less abundant than subgroups.

Example 3.3.15. Consider the subgroup $H = \langle (1\ 2) \rangle = \{(1), (1\ 2)\} < S_3$. We claim that H is not a normal subgroup. For this, we let $(2\ 3) \in S_3$ and noting that $(2\ 3)^{-1} = (2\ 3)$ we compute

$$(2\ 3)(1\ 2)(2\ 3) = (1\ 3).$$

Therefore

$$(2\ 3)H(2\ 3) \neq H$$

since $(1\ 3) \notin H$.

Later we will see that every normal subgroup of a group is the kernel of some homomorphism.

3.3.2 Isomorphisms

We have seen many groups that "look the same". For example, the additive group of congruence classes $\mathbb{Z}_n$ and the roots of unity $C_n < \mathbb{C}^\times$ have different descriptions, but they are in fact different manifestations of the "same group". We make precise the way in which two groups should be considered as "the same" as follows.

Definition 3.3.16. *Given groups G and H, a* **group isomorphism** *(or simply* **isomorphism***) from G to H is a bijective homomorphism $\phi \colon G \to H$. If there exists an isomorphism $\phi \colon G \to H$, then we say that G and H are* **isomorphic** *and we write $G \cong H$.*

Proposition 3.3.17. *Suppose $\phi \colon G \to H$ is an isomorphism. Then $\phi^{-1} \colon H \to G$ is an isomorphism. If $\psi \colon H \to K$ is an isomorphism, then $\psi \circ \phi \colon G \to K$ is also an isomorphism.*

Proof. For the first part, observe that for all $h, h' \in H$, there exists $g, g' \in G$ so that $\phi(g) = h$ and $\phi(g') = h'$ since ϕ is surjective. Then

$$\phi^{-1}(hh') = \phi^{-1}(\phi(g)\phi(g')) = \phi^{-1}(\phi(gg')) = gg' = \phi^{-1}(h)\phi^{-1}(h'),$$

so ϕ^{-1} is also a homomorphism. Since ϕ^{-1} is a bijection by Proposition 1.1.8, it is an isomorphism.

According to Proposition 3.3.6, $\psi \circ \phi$ is a homomorphism. On the other hand, Lemma 1.1.7 ensures that this is also a bijection, hence an isomorphism. □

An isomorphism $\phi \colon G \to G$ from a group G to itself is called an **automorphism of** G. The set of all automorphisms of G is denoted $\mathrm{Aut}(G)$. From Proposition 3.3.17, we immediately observe

Corollary 3.3.18. *For any group G, we have $\mathrm{Aut}(G) < \mathcal{S}ym(G)$ is a subgroup.*

Proof. Since $id \in \mathrm{Aut}(G)$, $\mathrm{Aut}(G) \neq \emptyset$. Closure under composition and inverses follows from Proposition 3.3.17. □

Suppose G is a group and $g \in G$. Then **conjugation by** g is the function $c_g \colon G \to G$ given by

$$c_g(h) = ghg^{-1}.$$

We call ghg^{-1} the **conjugate of** h **by** g. Note that for a subgroup $H < G$, the conjugate of H we defined earlier, gHg^{-1}, consists of all conjugates of elements of H.

Proposition 3.3.19. *Let G be a group and $g \in G$. Then $c_g \colon G \to G$ is an automorphism and $c_g^{-1} = c_{g^{-1}}$.*

Proof. For all $h, h' \in G$, we have

$$c_g(hh') = g(hh')g^{-1} = gheh'g^{-1} = ghg^{-1}gh'g^{-1} = c_g(h)c_g(h').$$

Next we prove that $c_g^{-1} = c_{g^{-1}}$. This will prove that c_g is a bijection, hence an isomorphism. We compute

$$c_g c_{g^{-1}}(h) = c_g(g^{-1}hg) = g(g^{-1}hg)g^{-1} = (gg^{-1})h(gg^{-1}) = h$$

and

$$c_{g^{-1}} c_g(h) = c_{g^{-1}}(ghg^{-1}) = g^{-1}(ghg^{-1})g = (g^{-1}g)h(g^{-1}g) = h.$$

So, $c_{g^{-1}} = c_g^{-1}$, as required. □

We end this section with a few more examples of isomorphisms. The first is of a general nature.

Proposition 3.3.20. *Suppose G is a cyclic group and that $g \in G$ is a generator. If $|G| = |g| = \infty$, then $\phi \colon \mathbb{Z} \to G$ given by $\phi(n) = g^n$ is an isomorphism.*

If $|G| = |g| = n < \infty$, then $\phi([a]) = g^a$ well-defines an isomorphism

$$\phi \colon \mathbb{Z}_n \to G.$$

Proof. Suppose first that $|G| = |g| = \infty$. We have already seen in Example 3.3.5 that ϕ is a homomorphism. According to Proposition 3.2.6, $\phi(n) = g^n = g^m = \phi(m)$ if and only if $n = m$, and hence ϕ is injective. Since G is cyclic, $G = \langle g \rangle = \{g^n \mid n \in \mathbb{Z}\}$ by Proposition 3.2.3, and hence ϕ is surjective. Therefore, ϕ is a bijection, hence an isomorphism.

Now suppose that $|G| = |g| = n < \infty$. We wish to define $\phi([a]) = g^a$, but we must first check that this is well-defined. For this, consider any other representative $b \in [a]$, and observe that $b = a + nk$, for $k \in \mathbb{Z}$. Thus

$$g^b = g^{a+nk} = g^a g^{nk} = g^a (g^n)^k = g^a (e^k) = g^a e = g^a,$$

and ϕ is well-defined.

To see that this is also a homomorphism, we compute:

$$\phi([a] + [b]) = \phi([a + b]) = g^{a+b} = g^a g^b = \phi([a])\phi([b]).$$

Since $G = \langle a \rangle = \{e, g, g^2, \ldots, g^{n-1}\}$ we note that any element $g^a \in G$ is in the image $\phi([a]) = g^a$, and so ϕ is surjective. Since the domain and range have order n, it follows that ϕ is bijection, and hence an isomorphism. □

Example 3.3.21. Recall from Example 3.3.14 that for the direct product $H \times K$, the two kernels $\ker(\pi_H)$ and $\ker(\pi_K)$ are normal subgroups of $H \times K$. The function $\phi_H \colon H \to \ker(\pi_K) = H \times \{e\}$ given by $\phi_H(h) = (h, e)$ is a bijection. We also have

$$\phi_H(hh') = (hh', e) = (h, e)(h', e) = \phi_H(h)\phi_H(h'),$$

so ϕ_H is a homomorphism, hence an isomorphism. There is similarly an isomorphism $\phi_K \colon K \to \ker(\pi_H) = \{e\} \times K$ given by $\phi_K(k) = (e, k)$.

Example 3.3.22. Let $\mathbb{F}$ be a field, $n > 0$ any positive integer, and consider the vector space $\mathbb{F}^n$. Recall that the set of invertible linear transformations of $\mathbb{F}^n$ is a group, denoted $GL(\mathbb{F}^n)$, with respect to composition; see Example 2.6.15. Taking matrix representatives with respect to the standard bases $\mathbf{e}_1, \ldots, \mathbf{e}_n$ of $\mathbb{F}^n$ defines a bijection $\Phi\colon \mathrm{GL}(\mathbb{F}^n) \to \mathrm{GL}(n, \mathbb{F})$, the set of $n \times n$ matrices with nonzero determinant; see §2.4 and Theorem 2.4.22.

According to Proposition 2.4.19, $\Phi(TS) = \Phi(T)\Phi(S)$. From this we see that $\mathrm{GL}(n, \mathbb{F})$ is a group (with respect to matrix multiplication), and Φ is an isomorphism. Indeed, if id is the identity on $\mathbb{F}^n$, then $\Phi(\mathrm{id}) = I$, the identity matrix, and $IA = AI = A$. Furthermore, if $A = \Phi(T)$, then the inverse matrix is $A^{-1} = \Phi(T^{-1})$ since

$$AA^{-1} = \Phi(T)\Phi(T^{-1}) = \Phi(TT^{-1}) = \Phi(\mathrm{id}) = I = \Phi(T^{-1}T) = A^{-1}A.$$

Thus $\mathrm{GL}(n, \mathbb{F})$ is a group, and Φ is an isomorphism, so $\mathrm{GL}(n, \mathbb{F}) \cong \mathrm{GL}(\mathbb{F}^n)$.

Recall that $\mathbb{F}^\times$ with multiplication is a group; see Example 2.6.5. According to Proposition 2.4.23, $\det\colon \mathrm{GL}(n, \mathbb{F}) \to \mathbb{F}^\times$ is a group homomorphism.

Exercises.

Exercise 3.3.1. Suppose $\phi\colon G \to H$ is a homomorphism, and $g \in G$. Prove that for all $n > 0$ we have $\phi(g^n) = \phi(g)^n$ by induction on n, thus completing the proof of Proposition 3.3.4.

Exercise 3.3.2. Prove Proposition 3.3.6.

Exercise 3.3.3. Prove that for any group G, the center $Z(G)$ is a normal subgroup.

Exercise 3.3.4. Prove that if G is an abelian group, then every subgroup of G is normal.

Exercise 3.3.5. In Exercise 3.2.4 you found all six subgroups of S_3. Which are normal?

Exercise 3.3.6. Prove that for any subgroup $H < G$ and element $g \in G$, gHg^{-1} is also a subgroup of G, and that $c_g(h) = ghg^{-1}$ defines an isomorphism $c_g\colon H \to gHg^{-1}$. In particular, if $H \triangleleft G$, then conjugation *in* G defines an automorphism $c_g\colon H \to H$.

Exercise 3.3.7. Suppose $\phi\colon G \to H$ is an isomorphism of groups. Prove the following statements that indicate the way in which G and H are the "same":

(i) $\phi(Z(G)) = Z(H)$. In particular, G is abelian if and only if H is abelian.

(ii) $\phi(C_G(g)) = C_H(\phi(g))$.

(iii) For all $g \in G$, $|\phi(g)| = |g|$ (that is the order of g is equal to the order of $|\phi(g)|$).

Exercise 3.3.8. Let G be a group and $H < G$ a subgroup. Prove that the set

$$N(H) = \{g \in G \mid gHg^{-1} = H\} \subset G$$

is a subgroup containing H, and that $H \triangleleft N(H)$.

The subgroup $N(H)$ from Exercise 3.3.8 is called the **normalizer of** H*, and it is (by definition) the largest subgroup of G containing H in which H is normal.*

Exercise 3.3.9. Let $n \geq 1$. Prove that for every divisor d of n, the number of elements of $\mathbb{Z}_n$ of order d is exactly $\varphi(d)$, the Euler phi function of d. Using this, prove the formula

$$n = \sum_{d|n} \varphi(d).$$

The sum is over all positive divisors d of n. *Hint: By Theorem 3.2.10, the subgroups of $\mathbb{Z}_n$ correspond precisely to the positive divisors of n. Now look at Proposition 3.3.20 and Exercises 3.2.2 and 3.3.7* (iii).

Exercise 3.3.10. Prove that if $\gcd(n, m) = 1$, then $\mathbb{Z}_{nm} \cong \mathbb{Z}_n \times \mathbb{Z}_m$ and that $\mathbb{Z}_{nm}^\times \cong \mathbb{Z}_n^\times \times \mathbb{Z}_m^\times$ (recall that in $\mathbb{Z}_k^\times$, the operation is multiplication of congruence classes). *Hint: Theorem 1.5.8 and the discussion there.*

Exercise 3.3.11. Prove that S_3 and $\mathbb{Z}_6$ are not isomorphic.

The rest of the exercises in this section extend the notion of group homomorphism to rings, and develop some of their basic properties.

Definition 3.3.23. *If R and S are rings, then a function $\phi\colon R \to S$ is a* **ring homomorphism** *if*

$$\phi(a + b) = \phi(a) + \phi(b) \text{ and } \phi(ab) = \phi(a)\phi(b),$$

for all $a, b \in R$. If R and S both have 1 and $\phi(1) = 1$, then we say that the ring homomorphism is **unital**. *Note that a ring homomorphism is a homomorphism of additive groups. The* **kernel** *of a ring homomorphism $\ker(\phi) = \phi^{-1}(0)$ is thus an additive subgroup. Exercise 3.3.13 asks you to show that the kernel is a special kind of subring.*

Exercise 3.3.12. Prove that if R, S are rings and $\phi\colon R \to S$ is a ring homomorphism, then $\phi(0) = 0$.

Exercise 3.3.13. An **ideal** (or sometimes called a **two-sided ideal**) is a subring $\mathcal{I} \subset R$ with the property that for all $r \in R$ and $a \in \mathcal{I}$, we have $ra, ar \in \mathcal{I}$. Prove that the kernel of a ring homomorphism $\phi\colon R \to S$ is an ideal.

Exercise 3.3.14. Prove that for every $n \geq 1$, $\phi\colon \mathbb{Z} \to \mathbb{Z}_n$ given by $\phi(k) = [k]$ is a ring homomorphism.

Exercise 3.3.15. If $\mathbb{F}$ is a field and $a \in \mathbb{F}$ is any element, then prove that $\phi\colon \mathbb{F}[x] \to \mathbb{F}$ defined by $\phi(p(x)) = p(a)$ is a ring homomorphism from the polynomial ring $\mathbb{F}[x]$ to the ring $\mathbb{F}$.

Exercise 3.3.16. Let $\phi\colon \mathbb{Z} \to \mathbb{Z}$ be given by $\phi(k) = 2k$. Prove that although ϕ is a homomorphism of additive groups, it is *not* a ring homomorphism.

Exercise 3.3.17. A *ring isomorphism* is a bijective ring homomorphism. Prove that if R, S are rings, and $\phi\colon R \to S$ is a ring isomorphism, then R has a 1 if and only if S does, and in this case, ϕ is unital. Prove that the inverse of a ring isomorphism is a ring isomorphism.

You should...

- Be able to do all the exercises from this section and know the proofs of statements marked by †.
- Know what a homomorphism is, and be able to check whether or not a given function is a homomorphism.
- Know what the kernel of a homomorphism is, and what it says about injectivity.
- Understand products of sets and how to manipulate them.
- Know what a normal subgroup is, and be able to check if a given subgroup is normal.
- Know what an isomorphism is. Be able to decide when two groups (for example, two cyclic groups) are isomorphic.
- Know what a ring homomorphism and isomorphism are.
- Know what an ideal of a ring is, and that the kernel of a ring homomorphism is an ideal.

3.4 Permutation groups and dihedral groups.

Because the easiest examples of groups to think about are cyclic groups, and more generally abelian groups, we may develop a warped sense of intuition if we are constantly using these to gauge our understanding. Although we have introduced other examples in §2.6, we have not discussed them as much. Here we describe a few other important classes of groups. The first class, permutation groups, are perhaps the most important to consider.

Recall that for any set X, the permutation group $\mathcal{S}ym(X)$ of all bijections of X is a group with respect to the operation of composition, and when $X = \{1, 2, \ldots, n\}$, $\mathcal{S}ym(X)$ is denoted S_n. The groups $\mathcal{S}ym(X)$ and their subgroups are incredibly robust.

To illustrate this point, let G be any group. For any $g \in G$, we define a function $L_g \colon G \to G$, given by $L_g(h) = gh$ called **left multiplication by** g.

Theorem 3.4.1 (Cayley's Theorem)**.**† *For any group G and $g \in G$, $L_g \colon G \to G$ is a bijection. Furthemore, $L_\cdot \colon G \to \mathcal{S}ym(G)$, given by $g \mapsto L_g$ is an injective homomorphism.*

Proof. For all $g, g', h \in G$, we have

$$L_{gg'}(h) = (gg')h = g(g'h) = L_g(g'h) = L_g(L_{g'}(h)) = L_g \circ L_{g'}(h).$$

That is, $L_{gg'} = L_g \circ L_{g'}$. Furthermore, if $e \in G$ is the identity, then $L_e = \mathrm{id}$, the identity function $\mathrm{id} \colon G \to G$.

From this we see that for all $g \in G$, $L_{g^{-1}} \circ L_g = L_{g^{-1}g} = L_e = \mathrm{id}$ and $L_g \circ L_{g^{-1}} = L_{gg^{-1}} = L_e = \mathrm{id}$, and hence L_g is a bijection with $L_g^{-1} = L_{g^{-1}}$. Since $L_{gg'} = L_g \circ L_{g'}$, $L_\cdot \colon G \to \mathcal{S}ym(G)$ is a homomorphism. Finally, $g \in \ker(L_\cdot)$ if and only if $L_g(h) = h$ for all $h \in G$. In this case, $g = ge = L_g(e) = e$, so $\ker(L_\cdot) = \{e\}$, and by Proposition 3.3.9, $L_\cdot$ is injective. □

Remark 3.4.2. For any $g \in G$, there is also **right multiplication by** g, $R_g \colon G \to G$, given by $R_g(h) = hg$. This is also a bijection, but because we compose right-to-left, $R_\cdot$ is not a homomorphism.

The following corollary distills Cayley's Theorem into a more transparent form.

Corollary 3.4.3. *Let G be any group. Then G is isomorphic to a subgroup of $\mathcal{S}ym(G)$.*

Proof. An injective homomorphism is an isomorphism onto its image. □

A more striking corollary for finite groups is the following.

Corollary 3.4.4. *For any integer $n > 0$, S_n contains an isomorphic copy of every group of order n. That is, for every group G of order n, there exists a subgroup $H < S_n$ such that $G \cong H$.*

For the proof, we require a lemma.

Lemma 3.4.5. *If $\Phi \colon X \to Y$ is a bijection, then there is an isomorphism $c_\Phi \colon \mathcal{S}ym(X) \to \mathcal{S}ym(Y)$ defined by $c_\Phi(\sigma) = \Phi\sigma\Phi^{-1}$.*

Proof. To see that c_Φ is a homomorphism, we compute

$$c_\Phi(\sigma\tau) = \Phi\sigma\tau\Phi^{-1} = \Phi\sigma\,\mathrm{id}\,\tau\Phi^{-1} = \Phi\sigma\Phi^{-1}\Phi\tau\Phi^{-1} = c_\Phi(\sigma)c_\Phi(\tau).$$

Since Φ^{-1} is also a bijection, we define $c_{\Phi^{-1}}$ similarly, and observe that $c_\Phi^{-1} = c_{\Phi^{-1}}$. Thus, c_Φ is a bijection, hence an isomorphism. □

Proof of Corollary 3.4.4. Let G be any group of order n, let $\Phi \colon G \to \{1, \ldots, n\}$ be any bijection, and let $c_\Phi \colon \mathcal{S}ym(G) \to S_n$ be the isomorphism from Lemma 3.4.5. Then the composition $c_\Phi \circ L_\cdot \colon G \to S_n$ is an injective homomorphism (see Proposition 3.3.6), and so G is isomorphic to its image, a subgroup of S_n. □

As n increases, the groups S_n become more and more complicated. For example, note that any permutation in S_n can also be viewed as a permutation in S_{n+k}, for any $k > 0$, by requiring it to fix every element not in $\{1, \ldots, n\}$ (in fact, the disjoint cycle notation for the element in S_{n+k} is exactly the same as in S_n). This defines an injective homomorphism $S_n \to S_{n+k}$, and so one has the further corollary of Corollary 3.4.4.

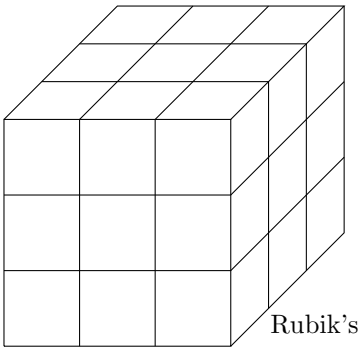
Corollary 3.4.6. *For any integer $n > 0$, S_n contains an isomorphic copy of every group of order $\leq n$.*

Proof. This is immediate from the preceding paragraph and Corollary 3.4.4. □

Some groups can be more naturally described using the symmetric group.

Example 3.4.7. The **Rubik's cube** is a 3–dimensional mechanical puzzle. It is a cube in which each of the 6 faces have been subdivided into 9 squares called **facets**. The facets are colored one of 6 colors, and in the *solved* state, each of the 9 facets on each face have the same color. The goal of the puzzle is to return it to the solved state after it has been "jumbled".

There are 6 generating *moves* obtained by rotating each of the 6 faces 90° clockwise. The first thing we notice is that the *center facets* of each face do not change position upon application of any move. Therefore, each move permutes the remaining 48 facets. If we enumerate those 48 facets, then we can define $\mathcal{R}u < S_{48}$ to denote the subgroup generated by the permutations induced by the 6 moves. The figure shows one such labeling.



Rubik's cube

			1	2	3						
			4		5						
			6	7	8						
9	10	11	12	13	14	15	16	17	18	19	20
21		22	23		24	25		26	27		28
29	30	31	32	33	34	35	36	37	38	39	40
			41	42	43						
			44		45						
			46	47	48						

"unfolded" cube with labeled facets

It is useful to label the faces based on their location relative to the observer. To this end, we label the faces as follows: (F) *front*, (B) *back*, (R) *right*, (L) *left*, (U) *up*, and (D) *down*. (A better choice might replace "up" and "down" with "top" and "bottom", respectively, but then two faces are denoted by "B"). We also let $F, B, R, L, U, D \in \mathcal{R}u$ denote the permutations obtained by clockwise rotation in the face of the same name. For example,

$$F = (6\ 15\ 43\ 31)(7\ 25\ 42\ 22)(8\ 35\ 41\ 11)(12\ 14\ 34\ 32)(13\ 24\ 33\ 23).$$

We will return to this group later and see how group theory can help us to understand the structure of this group and along the way, how one can solve the Rubik's cube.

3.4.1 Sign of a permutation.

There is an important homomorphism $\epsilon\colon S_n \to \{\pm 1\} = C_2$ called the **sign homomorphism** (or sometimes the **parity homomorphism**. If $\epsilon(\sigma) = 1$, we say that σ is **even**, and if $\epsilon(\sigma) = -1$, we say that σ is **odd**. The definition of ϵ is somewhat involved, but the next proposition explains the even/odd terminology. First, recall from Proposition 1.3.9 that every permutation in S_n can be expressed as a composition of 2–cycles.

Proposition 3.4.8. *If $\sigma \in S_n$ is a composition of k 2–cycles, then $\epsilon(\sigma) = (-1)^k$.*

Before we can prove this proposition, we must first *define* the homomorphism ϵ. To do this, consider any real valued function of n variables $q(x_1, \ldots, x_n)$, and for $\sigma \in S_n$, define

$$\sigma \cdot q(x_1, \ldots, x_n) = q(x_{\sigma(1)}, \cdots, x_{\sigma(n)}).$$

That is, $\sigma \cdot q$ is the function obtained from q by permuting the variables according to σ. Observe that if $\sigma, \tau \in S_n$, then

$$(\sigma\tau) \cdot q(x_1, \ldots, x_n) = q(x_{\sigma\tau(1)}, \ldots, x_{\sigma\tau(n)}) = \sigma \cdot q(x_{\tau(1)}, \ldots, x_{\tau(n)}) = \sigma \cdot (\tau \cdot q).$$

Next, consider the function

$$p(x_1, \ldots, x_n) = \prod_{1 \leq i < j \leq n} (x_j - x_i).$$

Observe that $p(1, 2, 3, \ldots, n) > 0$, so $p \neq 0$. Also, we have $\sigma \cdot p = \pm p$, for any $\sigma \in S_n$, and we define $\epsilon(\sigma) \in \{\pm 1\}$ to be such that $\sigma \cdot p = \epsilon(\sigma)p$. Since $\tau \cdot (-p) = -\tau \cdot p$, it follows that

$$\epsilon(\sigma\tau)p = (\sigma\tau) \cdot p = \sigma \cdot (\tau \cdot p) = \sigma \cdot (\epsilon(\tau)p) = \epsilon(\tau)\sigma \cdot p = \epsilon(\tau)\epsilon(\sigma)p = \epsilon(\sigma)\epsilon(\tau)p.$$

That is, $\epsilon(\sigma\tau) = \epsilon(\sigma)\epsilon(\tau)$, and so ϵ is a homomorphism.

Proof of Proposition 3.4.8. Suppose τ is a 2–cycle, say $\tau = (i\ j)$, for some $1 \leq i < j \leq n$. For each $k \neq l$, either $k < l$ and $(x_l - x_k)$ is one of the factors in the product, or else $k > l$ and $(x_k - x_l)$ is one of the factors. Whichever of these exists in the product is called the $\{k, l\}$ factor. When we consider $\tau \cdot p$, we can think of the effect of τ on the factors, which it permutes, possibly changing the sign. If $\{k, l\} \cap \{i, j\} = \emptyset$, then the $\{k, l\}$ factor is unaffected. If $k \notin \{i, j\}$, then the $\{k, i\}$ and $\{k, j\}$ factors are swapped, and either both change signs, or neither does. The only remaining factor is the $\{i, j\}$ factor, which *necessarily* changes sign. Consequently, there are an odd number of sign changes, and hence $\epsilon(\tau) = -1$. If σ is a product of k 2–cycles, then since ϵ is a homomorphism, $\epsilon(\sigma) = (-1)^k$. □

Corollary 3.4.9. *If $\sigma \in S_n$ is a k–cycle, then σ is even if k is odd, and σ is odd if k is even.*

Proof. According to Exercise 1.3.3, a k–cycle can be expressed as composition of $(k-1)$ 2–cycles. According to Proposition 3.4.8 $\epsilon(\sigma) = (-1)^{k-1}$. □

Since S_n contains a 2–cycle as soon as $n \geq 2$, we see that for all $n \geq 2$, ϵ is surjective. The kernel of ϵ consists precisely of the even permutations and is called the **alternating group on n elements**, and is denoted

$$A_n = \ker(\epsilon) \lhd S_n.$$

Example 3.4.10. For $n = 3$, we can just list the elements of S_3, and observe that there are three 2–cycles, two 3–cycles (which are each products of two 2–cycles), and the identity, $(1) = (1\ 2)(1\ 2)$. So,

$$A_3 = \{(1), (1\ 2\ 3), (1\ 3\ 2)\}.$$

For A_4 we can similarly list all the elements

$$A_4 = \{(1), (1\ 2\ 3), (1\ 3\ 2), (1\ 2\ 4), (1\ 4\ 2), (1\ 3\ 4), (1\ 4\ 3), (2\ 3\ 4), (2\ 4\ 3), (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4), (2\ 3)\}.$$

We note that S_3 has 6 elements and A_3 has 3, while S_4 has 24 elements and A_4 has 12. In general, S_n has $n!$ elements (since there n choice for where 1 goes, $n-1$ for where 2 goes, etc.) and as we will see later (by Theorem 3.5.6), A_n has $\frac{n!}{2}$ elements.

3.4.2 Dihedral groups

Another interesting class of non-abelian groups to keep in mind are the dihedral groups, first introduced in Example 2.6.18. Recall that $P_n \subset \mathbb{R}^2$ is a regular n–gon, and D_n is the group of isometries of $\mathbb{R}^2$ that preserve P_n. There are many regular n–gons in the plane, and we haven't specified which of them we are talking about. If $T \in \mathrm{Isom}(\mathbb{R}^2)$ is an isometry, then $T(P_n)$ is another regular n–gon, and Exercise 3.4.7 asks you to show that conjugation by T, the isomorphism $c_T\colon \mathrm{Isom}(\mathbb{R}^2) \to \mathrm{Isom}(\mathbb{R}^2)$ given by $c_T(\Phi) = T\Phi T^{-1}$, restricts to an isomorphism from the subgroup preserving P_n to the subgroup preserving $T(P_n)$. Consequently, we can view either the group of isometries preserving P_n or those preserving $T(P_n)$, and since these groups are isomorphic, we view them as basically the same group (and consequently, will not distinguish between them).

Since any isometry of P_n must preserve the centroid, by applying some isometry to P_n, we may assume that this centroid is at the origin, $\mathbf{0}$, and consequently, the dihedral group D_n as a subgroup of

$$G_{\mathbf{0}} = \{T \in \mathrm{Isom}(\mathbb{R}^2) \mid T(\mathbf{0}) = \mathbf{0}\} = \{T_A \in \mathrm{Isom}(\mathbb{R}^2) \mid A \in \mathrm{O}(2)\}.$$

Here, $T_A(\mathbf{x}) = A\mathbf{x}$. On the other hand, $A \mapsto T_A$ defines an isomorphism $\mathrm{O}(2) \to G_{\mathbf{0}}$, so we can instead simply identify the dihedral group as a subgroup of $\mathrm{O}(2)$, and denote the elements by matrices.

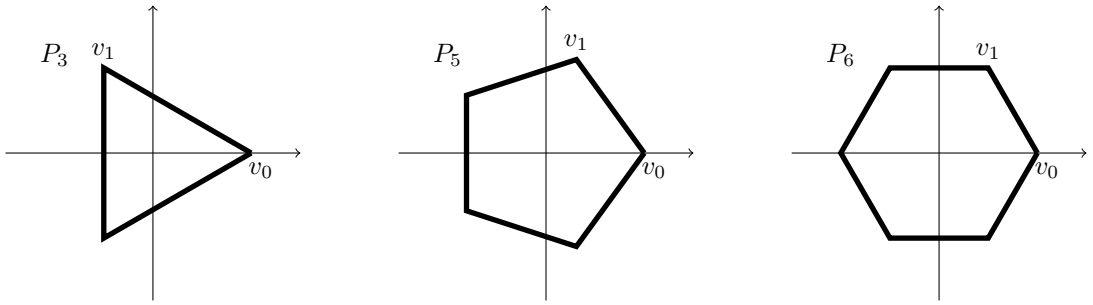
To do this, let's look more closely at $\mathrm{O}(2)$. A 2×2 matrix with entries in $\mathbb{R}$ is in $\mathrm{O}(2)$ if and only if $A^tA = I$. That is, the columns must form an orthonormal basis for $\mathbb{R}^2$. From this, we can easily express any element $A \in \mathrm{O}(2)$ in one of following two useful forms:

$$r_\theta = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \qquad \text{and} \qquad j_\psi = \begin{pmatrix} \cos(2\psi) & \sin(2\psi) \\ \sin(2\psi) & -\cos(2\psi) \end{pmatrix}.$$

The matrix r_θ represents a rotation through an angle θ (counterclockwise), while j_ψ represents a reflection through the line that makes an angle ψ with the positive x–axis (counterclockwise). Simple computations (or geometric considerations) prove the following formulas for all θ, ψ:

$$j_\psi^2 = I, \qquad r_\theta^{-1} = r_{-\theta}, \qquad r_\theta r_\psi = r_{\theta+\psi}, \qquad j_\theta j_\psi = r_{\theta-\psi}, \qquad j_\psi r_{-\theta} = r_\theta j_\psi = j_{\psi+\theta} r_\theta = j_{\frac{\psi+\theta}{2}}.$$

Now, to pin down $D_n < \mathrm{O}(2)$ exactly, we assume the centroid is at $\mathbf{0}$ and that one of the vertices of P_n, call it v_0 is along the positive x–axis. Let v_1 be the next vertex counterclockwise from v_0. Viewing v_0 and v_1 as vectors (based at $\mathbf{0}$), we see that the angle from v_0 to v_1 is $2\pi/n$ (counterclockwise). See the figure below.



Every element of D_n must take v_0 to one of the n vertices of P_n, and must send v_1 to a vertex adjacent to the image of v_1. Furthermore, any element of D_n is determined by what it does to v_0 and v_1. This means that there are at most $2n$ elements of D_n. On the other hand, we can list $2n$ elements (and hence all elements of D_n) as follows.

Fix $n \geq 3$, and set $r = r_{2\pi/n}$ and $j = j_0$. Then we have

$$D_n = \{I, r, r^2, r^3, \ldots, r^{n-1}, j, rj, r^2j, r^3j, \ldots, r^{n-1}j\}.$$

To see that these are all distinct, observe that the first n elements are all rotations, where the last are all reflections (by the above calculations), and $r^k = r_{2\pi k/n}$ and $r^k j = j_{\pi k/n}$. We can also use the calculations above to prove the following *relations* among the elements of D_n:

$$r^k r^\ell = r^{k+\ell}, \qquad r^k j = j r^{-k}, \qquad r^k j r^\ell j = r^{k-\ell}.$$

The exercises provide more opportunities to explore the structure of D_n. We note that this description is independent of the *size* of P_n, so up to isomorphism, D_n is independent of the choice of regular n–gon in the plan.

Exercises.

Exercise 3.4.1. Let $\tau \in S_n$ and suppose that $\sigma = (k_1\ k_2\ \dots\ k_j)$ is a j–cycle. Prove that the conjugate of σ by τ is also a j–cycle, and is given by

$$\tau\sigma\tau^{-1} = (\tau(k_1)\ \tau(k_2)\ \dots\ \tau(k_j)).$$

Further prove that if $\sigma' \in S_n$ is any other j–cycle, then σ and σ' are conjugate. *Hint: For the second part, you should explicitly find a conjugating element $\tau \in S_n$.*

The **cycle structure** of an element $\sigma \in S_n$ denotes the number of cycles of each length in the disjoint cycle representation of σ. We can encode the cycle structure with a **partition of** n:

$$n = j_1 + j_2 + \cdots + j_r,$$

where $\{j_i\}$ are positive integers giving the length of the distinct cycles (where we include "1's" for every number that is fixed, which we view as a "1–cycle" i.e. the identity). For example, the cycle structure of $(1\ 2\ 3)(5\ 6) \in S_6$ is $1 + 2 + 3 = 6$, since there is a 1–cycle, a 2–cycle, and a 3–cycle in the disjoint cycle representation.

Exercise 3.4.2. Suppose $\sigma_1, \sigma_2 \in S_n$. Using the previous exercise, prove that σ_1 and σ_2 have the same cycle structure if and only if they are conjugate.

Exercise 3.4.3. Proposition 1.3.9 shows that every permutation is a composition of 2–cycles, and thus the set of all 2–cycles generates S_n (i.e. the subgroup $G < S_n$ generated by the set of 2–cycles is all of S_n). Prove that $(1\ 2)$ and $(1\ 2\ 3\ \dots\ n)$ generates S_n; that is, prove

$$H = \langle (1\ 2), (1\ 2\ 3\ \dots\ n) \rangle = S_n.$$

Hint: Consider $\sigma = (1\ 2)(1\ 2\ 3\ \cdots\ n) \in H$ and then $\sigma^k(1\ 2)\sigma^{-k} \in H$ for $k \geq 1$. See also Exercise 3.4.1.

Exercise 3.4.4. If $\tau \in S_n$ is a k–cycle, prove $|\tau| = k$. If $\sigma \in S_n$ has cycle structure given by the partition $j_1 + j_2 + \cdots + j_r = n$ then prove that $|\sigma|$ is the least common multiple of $\{j_1, j_2, \dots, j_r\}$ (that is, $|\sigma|$ is the least common multiple of the orders of the cycles in the disjoint cycle representation of σ). *Hint: Use the fact that disjoint cycles commute, and look at Proposition 3.2.6 and its proof.*

Exercise 3.4.5. Let G be a group and for all $g \in G$, let $R_g \colon G \to G$ be right-multiplication $R_g(h) = hg$. Prove that the function $F \colon G \to \mathcal{S}ym(G)$ given by $F(g) = R_{g^{-1}}$ defines an injective homomorphism.

Exercise 3.4.6. Prove $D_3 \cong S_3$.

Exercise 3.4.7. Suppose $P \subset \mathbb{R}^n$ is any nonempty subset and $G_P < \mathrm{Isom}(\mathbb{R}^n)$ is the subgroup consisting of isometries preserving P:

$$G_P = \{\Phi \in \mathrm{Isom}(\mathbb{R}^n) \mid \Phi(P) = P\}$$

(see Proposition 2.6.17). Prove that conjugation by T, $c_T \colon \mathrm{Isom}(\mathbb{R}^n) \to \mathrm{Isom}(\mathbb{R}^n)$ restricts to an isomorphism from G_P to $G_{T(P)}$.

Exercise 3.4.8. Let $n \geq 3$. Prove that $R_n = \{I, r, r^2, r^3, \ldots, r^{n-1}\} \subset D_n$, the cyclic subgroup generated by r, is a normal subgroup. This is called the **subgroup of rotations**.

Exercise 3.4.9. Prove that if n is odd, then any two reflections $r^k j$ and $r^{k'} j$ in D_n are conjugate. Is the same true for n even? Explain.

You should...

- Be able to do all the exercises from this section and know the proofs of statements marked by †.
- Understand the complexity of subgroups of S_n (e.g. Corollary 3.4.6 and Example 3.4.7).
- Know what the sign homomorphism $\epsilon\colon S_n \to \{\pm 1\}$ is and what it tells you about a permutation (i.e. Proposition 3.4.8 and Corollary corollary:odd if even, even if odd). Know what an even and odd permutation is.
- Be able to compute the order of a permutation form its disjoint cycle representation (see Exercise 3.4.4).
- Be comfortable computing in the groups S_n and D_n, and understand what the elements of these groups look like.

3.5 Cosets and Lagrange's Theorem.

In this section we prove *Lagrange's Theorem* which relates the order of a finite group to the orders of its subgroups and explains some of the facts we seen so far (c.f. Remark 3.2.11 and Example 3.4.10). To state this, we require a few important definitions.

Definition 3.5.1. *Let G be a group, $H < G$ a subgroup, and $g \in G$. The* **left coset of H (by g)** *is the set*

$$gH = \{gh \mid h \in H\}.$$

The **right coset of H (by g)** *is similarly defined as $Hg = \{hg \mid h \in H\}$.*

The set of left cosets of H in G is denoted

$$G/H = \{gH \mid g \in G\},$$

and the number (cardinality) of left cosets is called the **index** *and is denoted $[G : H] = |G/H|$.*

Remark 3.5.2. We will often work with left cosets, stating and proving facts about them. There are analogous results for right cosets, but we will primarily restrict our discussion of right cosets to those results that are directly relevant to our interests.

Note that for every element $g \in G$, $g = ge \in gH$, so every element in G is in some coset. The next lemma provides different ways of thinking about when two elements belong to the same coset.

Lemma 3.5.3. *Let G be a group, $H < G$ a subgroup. For all $g_1, g_2 \in G$, the following are equivalent:*

(i) *g_1, g_2 belong to the same left coset of H,*

(ii) *$g_1 H = g_2 H$,*

(iii) *$g_1^{-1} g_2 \in H$.*

Proof. (i)⇒(ii): Suppose g_1, g_2 are in the coset gH. Then there exists $h_1, h_2 \in H$ so that $g_1 = gh_1$, $g_2 = gh_2$. Since $hH = L_h(H) = H$ for all $h \in H$, we have

$$g_1H = gh_1H = gH = gh_2H = g_2H.$$

(ii)⇒(iii): Next, suppose $g_1H = g_2H$. Then multiplying on the left by g_1^{-1} implies

$$H = eH = g_1^{-1}g_1H = g_1^{-1}g_2H.$$

Thus $e = g_1^{-1}g_2h$ for some $h \in H$. This implies $g_1^{-1}g_2 = h^{-1} \in H$.
(iii)⇒(i): Finally, suppose $g_1^{-1}g_2 \in H$. Then

$$g_2 = (g_1g_1^{-1})g_2 = g_1(g_1^{-1}g_2) \in g_1H.$$

Since $g_1 \in g_1H$, we see that g_1, g_2 are in the same left coset. □

Corollary 3.5.4. *If $H < G$ is a subgroup of a group G, then G/H is a partition of G.*

Proof. We've already pointed out that for any $g \in G$, we have $g \in gH$. If $g \in g_1H \cap g_2H$, then by Lemma 3.5.3, $g_1H = gH = g_2H$. It follows that G/H is a partition. □

If $H < G$ is a subgroup, a collection of elements $\mathcal{H} \subset G$ is called a set of **left coset representatives for H in G** if for each left coset $g'H \subset G$, there exists *exactly one* $g \in \mathcal{H}$ so that $gH = g'H$.

Example 3.5.5. For every integer $n > 0$, consider the subgroup $n\mathbb{Z} = \langle n \rangle < \mathbb{Z}$. For any $a \in \mathbb{Z}$, the left coset of $n\mathbb{Z}$ by a is precisely

$$a + n\mathbb{Z} = \{a + nk \mid k \in \mathbb{Z}\} = [a],$$

the congruence class of a. Consequently $\mathbb{Z}_n = \mathbb{Z}/n\mathbb{Z}$.

The integers $\{0, 1, 2, \ldots, n-1\}$ are coset representatives since for any $[a] \in \mathbb{Z}_n = \mathbb{Z}/n\mathbb{Z}$, there is exactly one integer $r \in \{0, 1, \ldots, n-1\}$ so that $[a] = [r]$. Here r is the remainder of a upon division by n; see Proposition 1.5.2.

Theorem 3.5.6 (Lagrange's Theorem)**.** † *If $H < G$ is any subgroup of a finite group G, then*

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

In particular, the order of H divides the order of G, $|H| \Big| |G|$.

Proof. For all $g \in G$, the bijection $L_g \colon G \to G$ (see Theorem 3.4.1) restricts to a bijection $L_g \colon H \to gH$ from H to gH, and therefore $|gH| = |H|$. Now suppose $[G : H] = n$, and let $\{g_1, \ldots, g_n\} \subset G$ be a set of coset representatives. Then we have

$$|G| = \sum_{i=1}^{n} |g_iH| = \sum_{i=1}^{n} |H| = n|H| = [G : H]|H|.$$

□

This has some rather surprising consequences.

Corollary 3.5.7. *Suppose G is a finite group and $g \in G$. Then $|g| \Big| |G|$.*

Proof. The order of g is the order of the cyclic subgroup generated by g, and hence divides the order of G by Theorem 3.5.6 □

From this we can describe all subgroups of prime order.

Corollary 3.5.8. *Suppose $p > 1$ is a prime integer. Then any group G of order p is cyclic, hence isomorphic to $\mathbb{Z}_p$. Furthermore, any non-identity element is a generator.*

Proof. Let $g \in G$, $g \neq e$. Then $|g| > 1$, and by Corollary 3.5.7, $|g| \Big| |G| = p$. Since p is prime, $|g| = p$, and so $|\langle g \rangle| = p$, and because $\langle g \rangle < G$, we have $\langle g \rangle = G$. Therefore, G is cyclic (generated by g), and by Proposition 3.3.20, $G \cong \mathbb{Z}_p$. □

Corollary 3.5.9. *Let G be a finite group and $K < H < G$ subgroups. Then*

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

Proof. According to Theorem 3.5.6 we have

$$[G : H][H : K] = \frac{|G|}{|H|}\frac{|H|}{|K|} = \frac{|G|}{|K|} = [G : K].$$

□

In Exercise 3.5.5 you are asked to prove that this corollary holds without the assumption that G is finite.

As we will see, the cosets of normal subgroups are particularly important. The following proposition provides some key special properties of normal subgroups, as well as characterizations of normality in terms of cosets.

Proposition 3.5.10. *Let G be a group and $N < G$ a subgroup. Then the following are equivalent.*

(i) *N is normal in G,*

(ii) *for all $g \in G$, $gN = Ng$,*

(iii) *for all $g \in G$, there exists $h \in G$ so that $gN = Nh$.*

Proof. (i)⇔(ii) For any $g \in G$, the equality $gN = Ng$ is equivalent to $gNg^{-1} = N$ by multiplying on the right by g (or g^{-1}).

(ii)⇒(iii) This is obvious: let $h = g$.

(iii)⇒(ii) Suppose $gN = Nh$. Then $g \in Nh$ and as in Lemma 3.5.3 we have $g \in Nh$, so $Ng = Nh$. Therefore $gN = Ng$ and hence gNg^{-1} □

As we have already seen, kernels of homomorphisms are normal subgroups. Here we see that the cosets of the kernel are also related to the homomorphism.

Proposition 3.5.11. *Suppose G, H are groups, $\phi\colon G \to H$ is a homomorphism, $N = \ker(\phi)$. Then for all $g \in G$, we have*

$$gN = Ng = \phi^{-1}(\phi(g)).$$

We recall that the function ϕ determines an equivalence relation on G whose equivalence classes are the fibers. This proposition says that this equivalence relation is the same as the one defined by the partition into cosets G/H (c.f. Corollary 3.5.4).

Proof. By Proposition 3.5.10, we have $gN = Ng$. Now, for all $n \in N$, we have

$$\phi(gn) = \phi(g)\phi(n) = \phi(g),$$

and so $gN \subset \phi^{-1}(\phi(g))$. If $g' \in \phi^{-1}(\phi(g))$, then $\phi(g') = \phi(g)$, so $\phi(g^{-1}g') = \phi(g)^{-1}\phi(g') = e$, and $g^{-1}g' \in \ker(\phi) = N$. By Lemma 3.5.3, $g' \in gN$, proving $\phi^{-1}(\phi(g)) \subset gN$, and hence $\phi^{-1}(\phi(g)) = gN$. □

Exercises.

Exercise 3.5.1. Prove *Fermat's Little Theorem*: For every prime $p \geq 2$ and $a \in \mathbb{Z}$, we have $a^p \equiv a \bmod p$. *Hint: Consider the two cases $p|a$ and $p \not| a$, in the latter case thinking about the group $\mathbb{Z}_p^\times$.*

Exercise 3.5.2. Prove *Euler's Theorem*: For every $a, n \in \mathbb{Z}$, $n \geq 1$, $\gcd(a, n) = 1$, we have $a^{\varphi(n)} \equiv 1 \bmod n$, where $\varphi(n) = |\mathbb{Z}_n^\times|$ is Euler's phi-function.

Exercise 3.5.3. Prove that for every group G, subgroup H and element $g \in G$, that $gH \mapsto Hg^{-1}$ defines a bijection between the set of left cosets and the set of right cosets. Conclude that the number of left cosets is equal to the number of right cosets (consequently the index could have been defined as either of these).

Exercise 3.5.4. Suppose G is a group and $N < G$ is a subgroup with $[G : N] = 2$. Prove that $N \triangleleft G$ is a normal subgroup.

Exercise 3.5.5. Prove Corollary 3.5.9 without the assumption that G is finite. Specifically, suppose $\mathcal{H} \subset G$ and $\mathcal{K} \subset H$ are coset representatives for $H < G$ and $K < H$, respectively, and prove that

$$\mathcal{J} = \{gh \mid g \in \mathcal{H}, h \in \mathcal{K}\}$$

is a set of coset representatives for K in G. Consequently,

$$[G : K] = |\mathcal{J}| = |\mathcal{H}||\mathcal{K}| = [G : H][H : K].$$

Note that the left hand side is infinite if and only if (either or both of) the factors on the right hand side is infinite.

Exercise 3.5.6. Suppose $K, H < G$ are subgroups of a group G. Prove that for all $g \in G$, $H \cap gK$ is either empty, or is equal to a coset of $K \cap H$ in H. Using this, prove that

$$[H : K \cap H] \leq [G : K].$$

Exercise 3.5.7. Suppose G_1, G_2 are groups and $H_1 < G_1$, $H_2 < G_2$ subgroups, and $\phi\colon G_1 \to G_2$ is an isomorphism with $\phi(H_1) = H_2$. Prove that $[G_1 : H_1] = [G_2 : H_2]$.

Exercise 3.5.8. Suppose G is a group and $H < G$ is a subgroup. According to Exercise 3.3.6, $gHg^{-1} < G$ is a subgroup for all $g \in G$. Combined with Proposition 3.2.2, this shows that the intersection of all conjugates is also a subgroup:

$$N = \bigcap_{g \in G} gHg^{-1} < G.$$

Prove that this is normal in G. The subgroup N is sometimes called the **normal core** of H in G.

Exercise 3.5.9. Prove that if $H < G$ is a finite index subgroup, then there are only finitely many conjugates $\{gHg^{-1} \mid g \in G\}$. Using this and Exercise 3.5.6, prove that the subgroup N from Exercise 3.5.8 also has finite index. *Hint: Exercise 3.3.6 and Exercise 3.5.7 might also be helpful.*

You should...

- Be able to do all the exercises from this section and know the proofs of statements marked by †.
- Know the definition of cosets and that cosets of a subgroup form a partition. Know what the index of a subgroup is.
- Know Lagrange's Theorem, its proof, and its consequences.
- Be able to decide when a subgroup is normal (e.g. Proposition 3.5.10).

3.6 Quotient groups and the isomorphism theorems.

The function $\pi\colon \mathbb{Z} \to \mathbb{Z}_n$, given by $\pi(a) = [a]$ is a surjective homomorphism with kernel $n\mathbb{Z} = \langle n \rangle$. In Example 3.5.5, we saw that $\mathbb{Z}_n$ is precisely the set of left cosets of $n\mathbb{Z}$, i.e. $\mathbb{Z}_n = \mathbb{Z}/n\mathbb{Z}$. We can thus think of $\mathbb{Z}_n$ as obtained from $\mathbb{Z}$ and the normal subgroup $n\mathbb{Z}$ (normality follows since $\mathbb{Z}$ is abelian), and the group structure ensures that $\mathbb{Z} \to \mathbb{Z}_n$ is a (surjective) homomorphism.

In this section, we describe a vast generalization of this construction, proving that normal subgroups are precisely the kernels of homomorphisms. After that, we prove the **three isomorphism theorems**: Theorems 3.6.5, 3.6.11, and 3.6.13. These theorems relate this construction to homomorphisms and subgroups in important and useful ways.

The construction hinges on the following.

Lemma 3.6.1. *If G is a group, $N \triangleleft G$ a normal subgroup, and $g, h \in G$, then the product of two left cosets is again a left coset. More precisely,*

$$(gN)(hN) = (gh)N.$$

Proof. From Proposition 3.5.10 we have $hN = Nh$. Therefore we have

$$(gN)(hN) = g(Nh)N = g(hN)N = (gh)NN = (gh)N,$$

where we have used the fact that since N is a subgroup, $NN = N$. □

Theorem 3.6.2. *Suppose G is a group and $N \triangleleft G$ is a normal subgroup. Then the product of cosets defines an operation on G/N making it into a group. Furthermore, the canonical map $\pi\colon G \to G/N$ given by $\pi(g) = gN$ is a surjective homomorphism with kernel N.*

Proof. The map $\pi\colon G \to G/N$ is the usual map sending an element of G to the element of the partition containing it (see Proposition 1.2.12 and Corollary 3.5.4), and so is surjective. Furthermore, by Lemma 3.6.1, we have

$$\pi(gh) = (gh)N = (gN)(hN) = \pi(g)\pi(h).$$

Thus, if we prove that G/N with the operation given by "product of cosets" is a group, then π will be a surjective homomorphism.

We must verify the three axioms for a group—these basically follow from the fact that G is a group and Lemma 3.6.1, but we elaborate. First, we have

$$(gN)((hN)(kN)) = (gN)((hk)N) = (g(hk))N = ((gh)k)N = ((gh)N)(kN) = ((gN)(hN))(kN),$$

so the operation is associative. Next we claim that $N = eN$ is the identity (where e is the identity in G). Indeed, for any $g \in G$, we have

$$N(gN) = (eN)(gN) = (eg)N = gN = (ge)N = (gN)(eN) = (gN)N.$$

Finally, we claim that the inverse of gN is $g^{-1}N$. For this, we again compute

$$(gN)(g^{-1}N) = (gg^{-1})N = eN = N = (g^{-1}g)N = (g^{-1}N)(gN).$$

This completes the proof. □

Definition 3.6.3. *If G is a group and $N \triangleleft G$ is a normal subgroup, the group G/N from Theorem 3.6.2 is called the* **quotient group** *of G by N. The homomorphism π is called the* **quotient homomorphism**.

Example 3.6.4. We have already pointed out that, as sets, $\mathbb{Z}_n = \mathbb{Z}/n\mathbb{Z}$. The operation on the quotient group is $\mathbb{Z}/n\mathbb{Z}$ is induced by the operation $+$ on $\mathbb{Z}$ from Lemma 3.6.1, and so is given by

$$(a + n\mathbb{Z}) + (b + n\mathbb{Z}) = (a + b) + n\mathbb{Z}.$$

On the other hand, $a + n\mathbb{Z} = [a]$, and so we have

$$[a] + [b] = (a + n\mathbb{Z}) + (b + n\mathbb{Z}) = (a + b) + n\mathbb{Z} = [a + b],$$

which is our original operation. Therefore, $\mathbb{Z}_n$ **is** the quotient group of $\mathbb{Z}$ by $n\mathbb{Z}$.

The first isomorphism theorem gives us a method of "identifying" a quotient group. Specifically, if we are given a normal subgroup $N \triangleleft G$, then we may want to prove that G/N is isomorphic to some other group H (maybe one we are more familiar with, for example). The next theorem tells us that to do this it suffices to find a surjective homomorphism from G onto H whose kernel is exactly N.

Theorem 3.6.5 (First Isomorphism Theorem)**.** † *Suppose $\phi\colon G \to H$ is a homomorphism of groups and $N = \ker(\phi)$. Then there exists a unique isomorphism $\widetilde{\phi}\colon G/N \to \phi(G) < H$ such that $\widetilde{\phi}\pi = \phi$. Thus, $G/N \cong \phi(G)$.*

The fact that $\phi(G) < H$ is a subgroup follows from Proposition 3.3.8. Of course, $\phi(G) = H$ if and only if ϕ is surjective.

Proof. Proposition 1.2.13 proves that there is a unique map $\widetilde{\phi}\colon G/N \to H$ satisfying $\widetilde{\phi}\pi = \phi$, and thus given by

$$\widetilde{\phi}(gN) = \phi(g),$$

and furthermore, $\widetilde{\phi}$ is a bijection from G/N to $\phi(G)$.

All that remains is to prove that $\widetilde{\phi}$ is a homomorphism. This follows easily from the fact that ϕ and π are homomorphisms:

$$\widetilde{\phi}((gN)(hN)) = \widetilde{\phi}(\pi(g)\pi(h)) = \widetilde{\phi}(\pi(gh)) = \phi(gh) = \phi(g)\phi(h) = \widetilde{\phi}(\pi(g))\widetilde{\phi}(\pi(h)) = \widetilde{\phi}(gN)\widetilde{\phi}(hN).$$

□

Remark 3.6.6. This is a good time to go back and review Proposition 1.2.13.

Example 3.6.7. Suppose $n = dk$ for positive integers n, d, k. Then we can form the quotient group $\mathbb{Z}_n/\langle [d]_n \rangle$ since $\mathbb{Z}_n$ is abelian, and hence all subgroups are normal. On the other hand, we have the homomorphism $\pi_{d,n}\colon \mathbb{Z}_n \to \mathbb{Z}_d$, and the kernel is the set of all congruence classes $[a]_n \in \mathbb{Z}_n$ such that $d|a$. This is precisely the subgroup generated by $[d]_n$, and so $\ker(\pi_{d,n}) = \langle [d]_n \rangle$. Therefore by Theorem 3.6.5, we have

$$\mathbb{Z}_n/\langle [d]_n \rangle \cong \mathbb{Z}_d.$$

Example 3.6.8. Since $\mathbb{R}$, the group of real numbers with addition, is abelian, the subgroup $\mathbb{Z} \triangleleft \mathbb{R}$ is normal, and we wish to "identify" the quotient $\mathbb{R}/\mathbb{Z}$. For this, observe that the function $\phi\colon \mathbb{R} \to S^1$, given by

$$\phi(t) = e^{2\pi i t}$$

is a homomorphism. This follows from the geometric interpretation of complex multiplication:

$$\phi(t+s) = e^{2\pi i(t+s)} = e^{2\pi i t + 2\pi i s} = e^{2\pi i t} e^{2\pi i s} = \phi(t)\phi(s).$$

The kernel is $\ker(\phi) = \{t \in \mathbb{R} \mid e^{2\pi i t} = 1\} = \mathbb{Z}$. By Theorem 3.6.5, we have $\mathbb{R}/\mathbb{Z} \cong S^1$.

The second isomorphism theorem involves the product of subgroups of a group. More precisely, suppose $H, K < G$ are subgroups. The product of sets HK is a *subset* of G, but need not be a subgroup in general. The next proposition gives the necessary and sufficient conditions for HK to be a subgroup.

Proposition 3.6.9. *Let $H, K < G$ be subgroups of a group G. Then HK is a subgroup of G if and only if $HK = KH$.*

Proof. First, suppose that HK is a subgroup. Then since H is a subgroup, the set of all inverses of elements in H, denoted H^{-1} is again equal to H, and likewise for K and HK:

$$H^{-1} = H, \quad K^{-1} = K, \text{ and } \quad (HK)^{-1} = HK.$$

But since the inverse of a product hk is given by $(hk)^{-1} = k^{-1}h^{-1}$, we have

$$HK = (HK)^{-1} = K^{-1}H^{-1} = KH.$$

Next, suppose that $HK = KH$. Since $e = ee \in HK$, $HK \neq \emptyset$. We must show that $HKHK = HK$ and that $(HK)^{-1} = HK$. For the first, we observe that

$$HKHK = H(KH)K = H(HK)K = (HH)(KK) = HK.$$

For the second, we argue similar to the above:

$$(HK)^{-1} = K^{-1}H^{-1} = KH = HK.$$

□

One way to ensure that $HK = KH$ is if one of the subgroups is normal.

Corollary 3.6.10. *Suppose $H, K < G$ are subgroups of a group G with K normal. Then $HK = KH$, and hence $HK < G$ is a subgroup.*

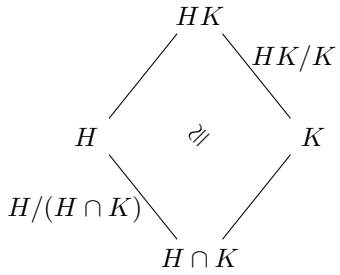
Proof. Since $K \triangleleft G$, we know that $hK = Kh$ for all $h \in H$ (see Proposition 3.5.10). Therefore

$$HK = \bigcup_{h \in H} hK = \bigcup_{h \in H} Kh = KH.$$

□

Theorem 3.6.11 (Second Isomorphism Theorem)**.** *Suppose $H, K < G$ are subgroups of a finite group G with K normal in G. Then $H \cap K \triangleleft H$ and*

$$HK/K \cong H/(H \cap K).$$



This is sometimes called the **Diamond Isomorphism Theorem** since the groups involved are arranged in a diamond in the subgroup lattice so that "parallel side quotients" are isomorphic.

Proof. The proof is an application of Theorem 3.6.5. For this, let $\phi\colon H \to HK$ denote the inclusion, which is a homomorphism, and let $\pi\colon HK \to HK/K$ be the quotient homomorphism. We claim that the kernel of the homomorphism $\pi\phi$ is exactly $H \cap K$. Indeed, for $h \in H$, we have $\pi\phi(h) = K$ (the identity in HK/K) if and only if $\phi(h) \in K$. That is, $\pi\phi(h) = K$ if and only if $h \in H \cap K$, and so $\ker(\pi\phi) = H \cap K$. By Theorem 3.6.5, $H/(H \cap K) \cong HK/K$. □

The primary application of Theorem 3.6.11 will be to the Sylow Theorems, which we will not encounter until the next chapter.

Before we get to the third isomorphism theorem, we must expand on the observations from Proposition 3.3.8. Given a group G, let $\mathcal{S}(G)$ denote the set of all subgroups of G. If $K < G$ is a subgroup, we let $\mathcal{S}(G, K)$ denote the set of all subgroups of G *that also contain* K:

$$\mathcal{S}(G, K) = \{L < G \mid K < L\}.$$

Theorem 3.6.12 (Correspondence Theorem)**.** *Suppose $\phi\colon G \to H$ is a surjective homomorphism with kernel N. Then there are inverse bijections*

$$\phi_*\colon \mathcal{S}(G,N) \to \mathcal{S}(H) \text{ and } \phi^*\colon \mathcal{S}(H) \to \mathcal{S}(G,N)$$

given by

$$\phi_*(K) = \phi(K) \text{ and } \phi^*(L) = \phi^{-1}(L).$$

Furthermore these bijections preserve normality. That is, $K \in \mathcal{S}(G,N)$ is a normal subgroup of G if and only if $\phi_*(K)$ is a normal subgroup of H.

Proof. By Proposition 3.3.8, ϕ_* defines a map, but to see that ϕ^* does as well, we need to know that $\phi^{-1}(L)$ contains N for any $L < H$. This follows from the fact that $\{e\} < L$ and hence

$$N = \ker(\phi) = \phi^{-1}(e) < \phi^{-1}(L).$$

Now we prove that ϕ_* and ϕ^* are inverses, and hence bijections.

For this, let $K \in \mathcal{S}(G,N)$ and observe that just because ϕ is a function, we have $K \subset \phi^{-1}(\phi(K))$. On the other hand, if $g \in \phi^{-1}(\phi(K))$, then $\phi(g) \in \phi(K)$, so there exists $k \in K$ with $\phi(g) = \phi(k)$. This means that $k^{-1}g \in \ker(\phi) = N < K$. Since $k \in K$, $g = kk^{-1}g \in K$, and so $\phi^{-1}(\phi(K)) \subset K$, proving equality $\phi^{-1}(\phi(K)) = K$.

Next, we note that just because ϕ is a surjective function, we have $\phi(\phi^{-1}(L)) = L$ for any subset L of H. In particular, this holds for $L \in \mathcal{S}(H)$. Consequently, ϕ_* and ϕ^* are inverses, as required.

Finally, to prove that normality is preserved by these bijections, suppose $K \in \mathcal{S}(G,N)$ is a normal subgroup of G and we must show that $\phi(K)$ is a normal subgroup of H. For this, let $h \in H$ be any element. Since ϕ is surjective, there exists $g \in G$ so that $\phi(g) = h$, then

$$h\phi(K)h^{-1} = \phi(g)\phi(K)\phi(g^{-1}) = \phi(gKg^{-1}) = \phi(K).$$

Conversely, suppose that $\phi(K)$ is normal and let $g \in G$. Then observe that $gKg^{-1} \in \mathcal{S}(G,N)$ (since $gNg^{-1} = N$) and since $\phi(g)\phi(K)\phi(g)^{-1} = \phi(K)$ we have

$$K = \phi^{-1}(\phi(K)) = \phi^{-1}(\phi(g)\phi(K)\phi(g^{-1})) = \phi^{-1}(\phi(gKg^{-1})) = \phi^*(\phi_*(gKg^{-1})).$$

But since ϕ^* and ϕ_* are inverse bijections, we conclude

$$K = gKg^{-1}.$$

It follows that K is normal. □

The third isomorphism theorem addresses the following situation that arises from Theorem 3.6.12: Suppose $\phi\colon G \to H$ is a surjective homomorphism and $K \triangleleft G$ is a normal subgroup that contains $N = \ker(\phi)$. Therefore $\phi(K)$ is a normal subgroup of H. So, what is the relationship between the quotient G/K and $H/\phi(K)$?

Theorem 3.6.13 (Third Isomorphism Theorem)**.** *Suppose $\phi\colon G \to H$ is a surjective homomorphism and $K \triangleleft G$ is a normal subgroup that contains $N = \ker(\phi)$. Then*

$$G/K \cong H/\phi(K).$$

Proof. This is yet another application of Theorem 3.6.5. Specifically, let $\pi\colon H \to H/\phi(K)$ denote the quotient homomorphism. Then $\pi\phi\colon G \to H/\phi(K)$ is a homomorphism, surjective since both π and ϕ are surjective. The kernel consists of the elements $g \in G$ so that $\pi\phi(g) = \phi(K)$, the identity in $H/\phi(K)$. Since π is just the quotient homomorphism, $\pi\phi(g) = \phi(K)$ if and only if $\phi(g) \in \phi(K)$. This happens if and only if $g \in K$, since $\phi^{-1}(\phi(K)) = K$. Thus, $\ker(\pi\phi) = K$, and so by Theorem 3.6.5, we have $G/K \cong H/\phi(K)$. □

Example 3.6.14. Recall from Example 3.6.7 that if d and n are positive integers with $d|n$, then the quotient of $\mathbb{Z}_n$ by $\langle [d]_n \rangle$ is isomorphic to $\mathbb{Z}_d$. This can be viewed as an instance of Theorem 3.6.13. Indeed, under the surjective homomorphism $\pi\colon \mathbb{Z} \to \mathbb{Z}_n$, we have $\langle d \rangle = \pi^{-1}(\langle [d]_n \rangle)$ (c.f. Theorem 3.6.12). Therefore, $\mathbb{Z}_n/\langle [d]_n \rangle \cong \mathbb{Z}/\langle d \rangle = \mathbb{Z}_d$.

Another familiar form of the third isomorphism theorem involves the quotient group construction directly.

Theorem 3.6.15 (Third Isomorphism Theorem (second version))**.** *Suppose G is a group and $K, N \triangleleft G$ are normal subgroups with $K < N < G$. Then N/K is a normal subgroup of G/K, and*

$$(G/K)/(N/K) \cong G/N.$$

Proof. We leave the derivation of this theorem from Theorems 3.6.12 and 3.6.13 as an exercise. □

The advantage of this version of the third isomorphism theorem is that it can be easier to remember: G/N is obtained from $(G/K)/(N/K)$ by "canceling the K's".

Exercises.

Exercise 3.6.1. Suppose G is a group, $N, K < G$ with N normal in G, $N \cap K = \{e\}$, and $G = NK$. Prove that $G/N \cong K$.

Exercise 3.6.2. Let $\mathbb{F}$ be a field and let $\hat{\mathbb{F}} = \mathbb{F} \cup \{\infty\}$ (where ∞ is just a symbol). An $\mathbb{F}$**–linear fractional transformation** is a function

$$T\colon \hat{\mathbb{F}} \to \hat{\mathbb{F}}$$

given by

$$T(x) = \frac{ax + b}{cx + d}$$

where $ad - bc \neq 0$ and $T(\infty) = a/c$, while $T(-d/c) = \infty$ (recall that in field, a/c means ac^{-1}). Prove that the set of all linear fractional transformations $\mathcal{M}(\hat{\mathbb{F}})$ is a subgroup of $\mathcal{S}ym(\hat{\mathbb{F}})$. Further prove that if we let $\Delta < \mathrm{GL}(2, \mathbb{F})$ denote the subgroup

$$\Delta = \left\{ \left(\begin{array}{cc} a & 0 \\ 0 & a \end{array} \right) \middle| \, a \in \mathbb{F}^\times \right\},$$

then Δ is a normal subgroup and

$$\mathrm{GL}(2, \mathbb{F})/\Delta \cong \mathcal{M}(\hat{F}).$$

In fact, Δ is the center of $\mathrm{GL}(2, \mathbb{F})$.

Exercise 3.6.3. Prove the following generalization of Theorem 3.6.5. Let G be a group, $\phi\colon G \to H$ a homomorphism, and $N \triangleleft G$ a normal subgroup *contained in* $\ker(\phi)$ and $\pi\colon G \to G/N$ the quotient homomorphism. Prove that there exists a unique homomorphism $\widetilde{\phi}\colon G/N \to H$ such that $\widetilde{\phi}\pi = \phi$. *Hint: Look at the proofs of Theorem 3.6.5 and Proposition 1.2.14.*

Exercise 3.6.4. Suppose that G_1, G_2 are groups, $N_1 \triangleleft G_1$, $N_2 \triangleleft G_2$ are normal subgroups. Prove that $N_1 \times N_2 \triangleleft G_1 \times G_2$, and $(G_1 \times G_2)/(N_1 \times N_2) \cong G_1/N_1 \times G_2/N_2$. *Hint: Theorem 3.6.5.*

Exercise 3.6.5. Suppose G is a group and $N, K \triangleleft G$. Prove $NK \triangleleft G$.

Exercise 3.6.6. Suppose G is a finite group and $N, H < G$ with N normal in G. Prove

$$|HN| = \frac{|H||N|}{|H \cap N|}.$$

Hint: Second Isomorphism Theorem.

Exercise 3.6.7. Prove Theorem 3.6.15.

Exercise 3.6.8. Let G be any group and $g, h \in G$. The **commutator** of g and h, is the element $[g,h] = ghg^{-1}h^{-1}$. The **commutator subgroup of** G, denoted G', is the subgroup generated by all commutators:

$$G' = \langle \{[g,h] \mid g,h \in G\} \rangle.$$

(See Section 3.1 to review the "subgroup generated by a subset".) Prove that $G' \triangleleft G$ and that G/G' is abelian. Further prove that if $\phi\colon G \to H$ is any homomorphism from a group G to an abelian group H, then $G' < \ker(\phi)$.

Exercise 3.6.9. Recall that for any group G, $\mathrm{Aut}(G)$ denotes the group of all automorphisms of G (see Corollary 3.3.18) and that conjugation by an element $g \in G$ defines an automorphism $c_g\colon G \to G$, given by $c_g(h) = ghg^{-1}$. An automorphism of the form c_g for some $g \in G$ is called an **inner automorphism**. Prove that the set of all inner automorphisms, $\mathrm{Inn}(G) \subset \mathrm{Aut}(G)$ is a group isomorphic to $G/Z(G)$, where $Z(G)$ is the center of G (see §3.2 and Exercise 3.3.3.).

Exercise 3.6.10. Let G be any group. With the notation introduced in Exercise 3.6.9, prove that $\mathrm{Inn}(G)$ is a normal subgroup of $\mathrm{Aut}(G)$. The quotient group $\mathrm{Aut}(G)/\mathrm{Inn}(G)$ is called the group of **outer automorphisms**. If G is abelian, prove that $\mathrm{Out}(G) \cong \mathrm{Aut}(G)$.

The remaining exercises in this section describe extensions of the results of this section to rings and ring homomorphisms.

Exercise 3.6.11. Suppose R is a ring and $\mathcal{I} \subset R$ is a (two-sided) ideal (see Exercise 3.3.13 for the definition of ideal). The quotient additive group $R/\mathcal{I}$ is the set of cosets, which in the additive notation have the form $r + \mathcal{I}$, for $r \in R$. Prove that

$$(r+\mathcal{I})(s+\mathcal{I}) = rs + \mathcal{I}$$

well-defines an operation on $R/\mathcal{I}$ (i.e. if $r+\mathcal{I} = r'+\mathcal{I}$ and $s+\mathcal{I} = s'+\mathcal{I}$, then $rs+\mathcal{I} = r's'+\mathcal{I}$). Furthermore, prove that this makes $R/\mathcal{I}$ into a ring, so that the quotient *group* homomorphism $\pi\colon R \to R/\mathcal{I}$ is also a ring homomorphism.

Exercise 3.6.12. Prove the *First Isomorphism Theorem for Rings*: If $\phi\colon R \to S$ is a surjective ring homomorphism, then there exists a unique ring isomorphism

$$\widetilde{\phi}\colon R/\ker(\phi) \to S$$

such that $\widetilde{\phi}\pi = \phi$. Hint: You already have a unique additive group isomorphism $\widetilde{\phi}$ so you just need to check that this is a ring homomorphism.

Exercise 3.6.13. Formulate and prove a *Second Isomorphism Theorem for Rings*, by analogy with Theorem 3.6.11.

Exercise 3.6.14. Prove the ring version of the correspondence theorem: If $\phi\colon R \to S$ is a surjective ring homomorphism, then the bijection between additive subgroups of S and those of R containing $\ker(\phi)$ from Theorem 3.6.12 restricts to a bijection on the set of subrings and ideals.

Exercise 3.6.15. Formulate and prove a *Third Isomorphism Theorem for Rings*, by analogy with Theorem 3.6.13.

You should...

- Be able to do all the exercises from this section and know the proofs of statements marked by †.

- Know how to construct the quotient group by a normal subgroup (i.e. how is operation defined), and know what the quotient homomorphism is (i.e. how is this homomorphism defined).
- Be able to construct isomorphisms (and more generally, homomorphisms) from a quotient group G/N to some other group H by constructing homomorphisms $G \to H$ (i.e. Theorem 3.6.5 and Exercise 3.6.3).
- Know what the product of two subgroups is and conditions required for it to be a subgroup (e.g. Proposition 3.6.9).
- Understand the second isomorphism theorem, the correspondence theorem and the third isomorphism theorem.
- Understand the extension of the group theoretic constructions and theorems here to rings (i.e. Exercises 3.6.11 – 3.6.15).

3.7 Products of groups.

We defined the direct product of groups in Section 3.1. In Section 3.6, we also saw that given two subgroups $H, K < G$, their product HK can also sometimes be a subgroup of G. In this section we discuss these and other "product constructions" and relations between them.

Proposition 3.7.1. † *Suppose that $H, K \triangleleft G$ are normal subgroups, $H \cap K = \{e\}$. Then the function $\phi\colon H \times K \to HK$ given by $\phi(h,k) = hk$ is an isomorphism. In particular, if $G = HK$, then $G \cong H \times K$.*

Proof. Let $h \in H$ and $k \in K$, and consider the **commutator** $[h,k] = hkh^{-1}k^{-1}$. This can be viewed in two ways:

$$(hkh^{-1})k^{-1} \quad \text{and} \quad h(kh^{-1}k^{-1}).$$

Since K is normal and $k \in K$, we have $hkh^{-1} \in K$. From the expression for $[h,k]$ on the left, we see that $[h,k] \in K$. Similarly, $kh^{-1}k^{-1} \in H$ and normality of H implies $[h,k] \in H$. But then $[h,k] \in H \cap K = \{e\}$. Consequently $[h,k] = e$, or equivalently, $hk = kh$. Because of this, the function ϕ is a homomorphism:

$$\phi((h,k)(h',k')) = \phi(hh',kk') = hh'kk' = h(h'k)k' = h(kh')k' = (hk)(h'k') = \phi(h,k)\phi(h',k').$$

By definition of HK, ϕ is surjective. Furthermore, if $\phi(h,k) = e$, then $hk = e$, and so $h = k^{-1}$. But since $H \cap K = \{e\}$ and $h \in H$ and $k^{-1} \in K$, we must have $h = k^{-1} = e$. Consequently, $\ker(\phi) = \{(e,e)\}$, and ϕ is injective, so bijective. Since ϕ is a homomorphism it is an isomorphism, as required. □

If H, K are any two groups, we have already seen in Example 3.3.14 that we have normal subgroups

$$\ker(\pi_H) = \{e\} \times K, \ker(\pi_K) = H \times \{e\} \triangleleft H \times K$$

It is also clear that $\ker(\pi_H) \cap \ker(\pi_K) = \{(e,e)\}$ and $H \times K = \ker(\pi_K)\ker(\pi_H)$. The proposition therefore gives us an isomorphism

$$\ker(\pi_K) \times \ker(\pi_H) \cong H \times K.$$

On the other hand, $\ker(\pi_K) \cong H$ and $\ker(\pi_H) \cong K$. The isomorphism from the proposition is merely the extension of these isomorphisms to the direct products of the domain and range.

Given a set of groups $H_1, \ldots, H_n$, we construct the direct product $H_1 \times H_2 \times \cdots \times H_n$ with the operation given by $(h_1, h_2, \ldots, h_n)(h_1', h_2', \ldots, h_n') = (h_1h_1', h_2h_2', \ldots, h_nh_n')$. We leave the verification that this makes the n–fold product into a group as an easy exercise. In addition, Proposition 3.7.1 has a natural generalization.

Proposition 3.7.2. *Suppose G is a group and $H_1, H_2, \ldots, H_n \triangleleft G$ are normal subgroups and that for all $1 \leq i \leq n$ we have*

$$H_i \cap (H_1H_2 \cdots H_{i-1}H_{i+1} \cdots H_n) = \{e\}.$$

Then $\phi\colon H_1 \times H_2 \times \cdots \times H_n \to H_1H_2 \cdots H_n$, given by $\phi(h_1, h_2, \ldots, h_n) = h_1h_2 \cdots h_n$ is an isomorphism.

Proof. This is left as Exercise 3.7.3. □

The intersection criteria of this proposition may seem like overkill. Why not simply assume $H_i \cap H_j = \{e\}$ whenever $i \neq j$? The next example shows that this is generally insufficient.

Example 3.7.3. Consider the subgroups of $\mathbb{Z} \times \mathbb{Z}$:

$$H_1 = \{(n, 0) \mid n \in \mathbb{Z}\}, H_2 = \{(0, n) \mid n \in \mathbb{Z}\}, \text{ and } H_3 = \{(n, n) \mid n \in \mathbb{Z}\}.$$

The first two are subgroups of the product, since they are the kernels of the homomorphisms onto the two factors, and the third is easily seen to be a subgroup. Since $\mathbb{Z} \times \mathbb{Z}$ is abelian (see Exercise 3.7.1), we see that all H_i are normal. Furthermore, for $i \neq j$, we have $(n, m) \in H_i \cap H_j$ if and only if $n = 0 = m$, and so $H_i \cap H_j = \{(0, 0)\}$. However, $\phi\colon H_1 \times H_2 \times H_3 \to H_1H_2H_3 = \mathbb{Z} \times \mathbb{Z}$ is not injective since, for example

$$\phi((0, n), (n, 0), (-n, -n)) = (n - n, n - n) = (0, 0)$$

for all $n \in \mathbb{Z}$.

3.7.1 Semidirect products

Proposition 3.7.2 provides one generalization of Proposition 3.7.1 obtained by allowing more than two normal subgroups. We can also relax the assumption in another way by requiring only one of the two subgroups to be normal (c.f. Corollary 3.6.10).

To explain this, let $H, K < G$ be subgroups with H normal in G and $H \cap K = \{e\}$. We note that the function $\phi\colon H \times K \to HK$ is still a bijection since it is clearly surjective, and because $H \cap K = \{e\}$ implies

$$\phi(h, k) = \phi(h', k') \Rightarrow hk = h'k' \Rightarrow (h')^{-1}h = k'k^{-1} \Rightarrow (h')^{-1}h = e = k'k^{-1} \Rightarrow h = h' \text{ and } k = k'.$$

Therefore, we can use the bijection to define a *new* group structure on $H \times K$ making ϕ into an isomorphism. To see what this group structure looks like, we should look at the operation on HK.

Recall from §3.3.2 and Exercise 3.3.6 that since $H \triangleleft G$, every element $g \in G$ defines an automorhpism $c_g\colon H \to H$ given by conjugating by g. Then, for any $h, h' \in H$ and $k, k' \in K$ we have

$$(hk)(h'k') = h(kh')k' = h(kh'(k^{-1}k))k' = (h(kh'k^{-1}))(kk') = (hc_k(h'))(kk').$$

Now we define the new operation on $H \times K$ by

$$(h, k)(h', k') = (hc_k(h'), kk'),$$

for all $(h, k), (h', k') \in H \times K$. Then $H \times K$ becomes a group with this operation and ϕ is an isomorphism. We stress that this is **not** in general the direct product group structure on $H \times K$. Specifically, if $c_k\colon H \to H$ is not the identity automorphism for some $k \in K$, then there exists $h' \in H$ so that $c_k(h') \neq h'$ and hence for any $h \in H$ and $k' \in K$, we have

$$(h, k)(h', k') = (hc_k(h'), kk') \neq (hh', kk').$$

On the other hand if each $c_k\colon H \to H$ is the identity for all $k \in K$, then it is the direct product.

With this in mind, we generalize this construction as follows.

Definition 3.7.4. *Suppose H and K are groups and $\alpha\colon K \to \mathrm{Aut}(H)$ is a homomorphism (which we denote $k \mapsto \alpha_k \in \mathrm{Aut}(H)$; see Remark 1.1.11). Let $H \rtimes_\alpha K$ denote the set $H \times K$ with the operation*

$$(h, k)(h', k') = (h\alpha_k(h'), kk')$$

and call this the **semidirect product of H and K by α.** *We sometimes write $H \rtimes K$ with the homomorphism α understood.*

Remark 3.7.5. We emphasize that *as a set* $H \rtimes_\alpha K$ is just $H \times K$. The symbol $\rtimes_\alpha$ instead of $\times$ is used to indicate that we have given the Cartesian produce a **different** operation. This notation is also suggestive: as the next proposition shows, there is a subgroup naturally isomorphic to H in $H \rtimes_\alpha K$ which is a normal subgroup (thus the "$\triangleleft$ part" of the symbol $\rtimes$).

Proposition 3.7.6. *For any two groups H and K and homomorhphism $\alpha\colon K \to \mathrm{Aut}(H)$, the semidirect product $H \rtimes_\alpha K$ is a group. Furthermore, the subsets $H_0 = H \times \{e\}$ and $K_0 = \{e\} \times K$ are subgroups isomorphic to H and K, respectively (by the obvious bijections). Furthermore, H_0 is normal and $H_0K_0 = H \rtimes K$.*

Proof. We first show that the operation is associative. To this end, let $(h,k)(h',k'),(h'',k'') \in H \rtimes K$. Then we have

$$((h,k)(h',k'))(h'',k'') = (h\alpha_k(h'),kk')(h'',k'') = (h\alpha_k(h')\alpha_{kk'}(h''),kk'k'')$$

and

$$(h,k)((h',k')(h'',k'')) = (h,k)(h'\alpha_{k'}(h''),k'k'') = (h\alpha_k(h'\alpha_{k'}(h'')),kk'k'').$$

To see that these are equal, we first observe that their second coordinates clearly are. For the first coordinates, we use the fact that α is a homomorphism to an automorphism group. This implies $\alpha_k \circ \alpha_{k'} = \alpha_{kk'}$ *and* for any h''' we have $\alpha_k(h')\alpha_k(h''') = \alpha_k(h'h''')$. In our situation, this gives

$$\alpha_k(h')\alpha_{kk'}(h'') = \alpha_k(h')\alpha_k(\alpha_{k'}(h'')) = \alpha_k(h'\alpha_{k'}(h'')),$$

so the first coordinates in our two calculations above are also equal, and the operation is associative.

The identity element is (e,e) since

$$(h,k)(e,e) = (h\alpha_k(e),ke) = (he,k) = (h,k)$$

and

$$(e,e)(h,k) = (e\alpha_e(h),ek) = (eh,k) = (h,k).$$

We claim that the inverse of (h,k) is given by $(h,k)^{-1} = (\alpha_{k^{-1}}(h^{-1}),k^{-1})$. We leave the verification of this claim, together with the rest of the proof of the proposition, as Exercise 3.7.4 □

Given that the definition of the semidirect product was motivated by a generalization of Proposition 3.7.1, the next fact should not come as a surprise.

Proposition 3.7.7. *Suppose G is a group with subgroups $H,K < G$ such that H is normal and $H \cap K = \{e\}$. Then the map*

$$\phi\colon H \rtimes_c K \to HK$$

defined by $\phi(h,k) = hk$ is an isomorphism, where $c\colon K \to \mathrm{Aut}(H)$ is conjugation (in G).

Proof. The calculations we did above precisely show that the bijection ϕ satisfies

$$\phi((h,k)(h',k')) = \phi(hc_k(h'),kk') = hc_k(h')kk' = (hk)(h'k') = \phi(h,k)\phi(h',k'),$$

and hence is an isomorphism. □

Example 3.7.8. For all $n \geq 3$, recall from §3.4.2 that the dihedral group D_n has $2n$ elements

$$D_n = \{I, r, r^2, r^3, \ldots, r^{n-1}, j, rj, r^2j, r^3j, \ldots, r^{n-1}j\}.$$

From Exercise 3.4.8, there is a normal, cyclic subgroup $R_n = \langle r \rangle$ of rotations. On the other hand, $\langle j \rangle < D_n$ is a cyclic subgroup of order 2, hence $\langle j \rangle \cong \mathbb{Z}_2$. From the description of the elements of D_n, we clearly have $R_n\langle j \rangle = D_n$, and since $j \notin R_n$, we see that $R_n \cap \langle j \rangle = \{I\}$. By Proposition 3.7.7, we have

$$R_n \rtimes \mathbb{Z}_2 \cong R_n \rtimes \langle j \rangle \cong D_n.$$

Example 3.7.9. Recall from §2.5 the two types of isometries $T_A, \tau_{\mathbf{v}} \in \mathrm{Isom}(\mathbb{R}^n)$, where $A \in \mathrm{O}(n)$, $\mathbf{v} \in \mathbb{R}^n$, and

$$T_A(\mathbf{x}) = A\mathbf{x} \text{ and } \tau_{\mathbf{v}}(\mathbf{x}) = \mathbf{x} + \mathbf{v},$$

for all $\mathbf{x} \in \mathbb{R}^n$. This gives us two subsets

$$H = \{\tau_{\mathbf{v}} \mid \mathbf{v} \in \mathbb{R}^n\} \quad \text{and} \quad K = \{T_A \mid A \in \mathrm{O}(n)\}.$$

These are in fact subgroups, and the bijections $\phi_H \colon \mathbb{R}^n \to H$ and $\phi_K \colon \mathrm{O}(n) \to K$ given by

$$\phi_H(\mathbf{v}) = \tau_{\mathbf{v}} \text{ and } \phi_K(A) = T_A$$

are isomorphisms. For H, we have

$$\phi_H(\mathbf{v}+\mathbf{u})(\mathbf{x}) = \tau_{\mathbf{v}+\mathbf{u}}(\mathbf{x}) = \mathbf{x}+\mathbf{v}+\mathbf{u} = \tau_{\mathbf{v}}(\mathbf{x}+\mathbf{u}) = \tau_{\mathbf{v}}(\tau_{\mathbf{u}}(\mathbf{x})) = \tau_{\mathbf{v}} \circ \tau_{\mathbf{u}}(\mathbf{x}) = \phi_H(\mathbf{v}) \circ \phi_H(\mathbf{u})(\mathbf{x}),$$

that is, $\phi_H(\mathbf{v}+\mathbf{u}) = \phi_H(\mathbf{v}) \circ \phi_H(\mathbf{u})$. Since ϕ_H is a bijection, this proves that $H = \phi_H(\mathbb{R}^n)$ is a subgroup and ϕ_H is an isomorphism. For K, this basically follows from the computations in §2.4 or Example 3.3.22.

In Exercise 3.7.5 you are asked to use these facts to prove

$$\mathrm{Isom}(\mathbb{R}^n) \cong \mathbb{R}^n \rtimes \mathrm{O}(n).$$

Exercises.

Exercise 3.7.1. If H, K are abelian groups, prove that $H \times K$ is abelian.

Exercise 3.7.2. Suppose G is a finite group, $H, K \triangleleft G$ are normal subgroups, $\gcd(|H|, |K|) = 1$, and $|G| = |H||K|$. Prove that $G \cong H \times K$.

Exercise 3.7.3. Prove that if $H_1, \ldots, H_n \triangleleft G$ are normal subgroups, then $H_1 H_2 \cdots H_n < G$ is a subgroup, then prove Proposition 3.7.2.

Exercise 3.7.4. Complete the proof of Proposition 3.7.6. Specifically, prove

1. $(h,k)^{-1} = (\alpha_{k^{-1}}(h^{-1}), k^{-1})$.
2. H_0, K_0 are subgroups and $h \mapsto (h,e)$ and $k \mapsto (e,k)$ define isomorphisms $H \cong H_0$ and $K \cong K_0$, respectively.
3. $H_0 \triangleleft H \rtimes_\alpha K$ and $H_0 K_0 = H \rtimes_\alpha K$.

Exercise 3.7.5. Prove that for the subgroups $H, K < \mathrm{Isom}(\mathbb{R}^n)$ described in Example 3.7.9 we have $H \triangleleft \mathrm{Isom}(\mathbb{R}^n)$, $HK = \mathrm{Isom}(\mathbb{R}^n)$, and $H \cap K = \{\mathrm{id}\}$. Deduces that

$$\mathrm{Isom}(\mathbb{R}^n) = HK \cong H \rtimes K \cong \mathbb{R}^n \rtimes \mathrm{O}(n).$$

Exercise 3.7.6. Prove that for all $n \geq 2$, we have $S_n \cong A_n \rtimes \mathbb{Z}_2$, where S_n is the symmetric group and A_n the alternating subgroup (consisting of the even permutations). *Hint: use Proposition 3.7.7. You have to* **find** *a subgroup $K < S_n$ with $K \cong \mathbb{Z}_2$ (c.f. Example 3.7.8).*

Exercise 3.7.7. In this exercise you will identify $\mathrm{Aut}(\mathbb{Z}_n)$.

1. Prove that for any $[a] \in \mathbb{Z}_n^\times$, multiplication by $[a]$ defines an automorphism $\alpha_{[a]} \colon \mathbb{Z}_n \to \mathbb{Z}_n$, given by $\alpha_{[a]}([k]) = [ak]$.
2. Prove that $\alpha \colon \mathbb{Z}_n^\times \to \mathrm{Aut}(\mathbb{Z}_n)$ defines an injective homomorphism.

3. Prove that any automorphism $\phi \in \mathrm{Aut}(\mathbb{Z}_n)$ is determined by $\phi([1])$, and that $\phi([1])$ must be a generator of $\mathbb{Z}_n$.

4. Prove that $\alpha\colon \mathbb{Z}_n^\times \to \mathrm{Aut}(\mathbb{Z}_n)$ is an isomorphism.

Exercise 3.7.8. For every $n \geq 3$, construct a nonabelian group of order $n\varphi(n)$, where $\varphi(n) = |\mathbb{Z}_n^\times|$ is the Euler phi function of n. *Hint: Use Exercise 3.7.7.*

You should...

- Be able to do all the exercises from this section and know the proofs of statements marked by †.
- Be able to recognize when a group is isomorphic to a direct product and semidirect product in terms of its subgroups.
- Know how the semidirect product is defined: know what all the data necessary is and how it is assembled into a group structure. Be able to construct a semidirect product when given sufficient data.
- Be familiar with examples of semidirect products.

Chapter 4

Group actions

The notion of a group is motivated in part by the notion of symmetries of geometric or algebraic objects. Theorem 3.4.1 shows that all groups can be viewed as symmetries of a set. In this chapter we describe an important extension of this idea in which the elements of a group *act* as symmetries through the notion of a *group action*.

4.1 Basics and the orbit-stabilizer theorem

There are two equivalent ways of defining a group action. We give one as the definition, and prove that the second is equivalent in a proposition.

Definition 4.1.1. *Given a group G and a nonempty set X, an* **action** *of G on X is a function*

$$G \times X \to X,$$

denoted $(g, x) \mapsto g \cdot x$, satisfying the following two properties:

1. *$e \cdot x = x$ for all $x \in X$*
2. *$g \cdot (h \cdot x) = gh \cdot x$ for all $x \in X$ and $g, h \in G$.*

Here $e \in G$ is the identity.

Proposition 4.1.2. *Suppose G is a group and X is a nonempty set. If $G \times X \to X$ is an action, then for all $g \in G$, the map $\alpha_g \colon X \to X$ defined by $\alpha_g(x) = g \cdot x$ for all $x \in X$ is a bijection, and $g \mapsto \alpha_g$ defines a homomorphism $\alpha \colon G \to \mathcal{S}ym(X)$.*

Conversely, if $\alpha \colon G \to \mathcal{S}ym(X)$ is a homomorphism, then $g \cdot x = \alpha_g(x)$ defines an action $G \times X \to X$.

Proof. First suppose $G \times X \to X$ is an action. By the second property of an action, we see that for all $g, h \in G$ and $x \in X$ we have

$$\alpha_g \circ \alpha_h(x) = \alpha_g(\alpha_h(x)) = \alpha_g(h \cdot x) = g \cdot (h \cdot x) = gh \cdot x = \alpha_{gh}(x),$$

and hence $\alpha_g \circ \alpha_h = \alpha_{gh}$. Combining this with the first property, we have

$$\alpha_g \circ \alpha_{g^{-1}} = \alpha_{gg^{-1}} = \alpha_e = \mathrm{id}_X = \alpha_e = \alpha_{g^{-1}g} = \alpha_{g^{-1}} \circ \alpha_g.$$

Therefore, $\alpha_{g^{-1}} = \alpha_g^{-1}$, and hence α_g is a bijection. By the first computation, $g \mapsto \alpha_g$ is a homomorphism from G to $\mathcal{S}ym(X)$, as required.

Now suppose $\alpha \colon G \to \mathcal{S}ym(X)$ is a homomorphism (denoted $g \mapsto \alpha_g$). Setting $g \cdot x = \alpha_g(x)$ for all $x \in X$ and $g \in G$, we have

$$e \cdot x = \alpha_e(x) = \mathrm{id}_X(x) = x.$$

If $g, h \in G$ and $x \in X$, then

$$gh \cdot x = \alpha_{gh}(x) = \alpha_g \circ \alpha_h(x) = \alpha_g(\alpha_h(g)) = \alpha_g(h \cdot x) = g \cdot (h \cdot x).$$

Therefore, $g \cdot x$ defines an action. □

So, we can either think of an action as a function $G \times X \to X$ satisfying the properties in its definition, or else as a homomorphism $G \mapsto \mathcal{S}ym(X)$, and we will use these interchangeably (via Proposition 4.1.2). In fact, we will often confuse an element $g \in G$ with the function $\alpha_g \colon X \to X$ which it defines.

There are a few important definitions associated to any action.

Definition 4.1.3. *Suppose $G \times X \to X$ is an action of G on a set X. For any $x \in X$, the* **stabilizer of x in G** *is defined to be*

$$\mathrm{stab}_G(x) = \{g \in G \mid g \cdot x = x\}.$$

That is, the stabilizer of x is the set of all elements $g \in G$ which fix x. The **kernel of the action** *is the set of all elements fixing every element of X:*

$$K = \{g \in G \mid g \cdot x = x \text{ for all } x \in X\}.$$

Alternatively, this is the kernel of the associated homomorphism $\alpha \colon G \to \mathcal{S}ym(X)$.

The **orbit of x** *is the subset*

$$G \cdot x = \{g \cdot x \mid g \in G\} \subset X,$$

consisting all images of x under the elements of G. The action is called **transitive** *if for all $x, y \in X$, there exists $g \in G$ so that $g \cdot x = y$. The* set *of orbits of the action is denoted*

$$\mathcal{O}_G(X) = \{G \cdot x \mid x \in X\}.$$

Remark 4.1.4. The notation for the stabilizer, kernel, and orbit of an action must always be understood in the context of a given action. If there are multiple actions, then we must be more careful in our use of this notation.

Some easy facts are listed in the following proposition.

Proposition 4.1.5. *Suppose $G \times X \to X$ is an action. Then for all $x \in X$, $\mathrm{stab}_G(x)$ is a subgroup of G and the kernel K of the action is the intersection of all stabilizers*

$$K = \bigcap_{x \in X} \mathrm{stab}_G(x).$$

The orbits $\mathcal{O}_G(X)$ form a partition of X, and the action of G on X is transitive if and only if $G \cdot x = X$ for some $x \in X$.

Proof. We leave this as Exercise 4.1.1 □

The last sentence tells us that an action of G on a set X determines an equivalence relation on X in which two elements are equivalent if and only if they are in the same orbit.

Example 4.1.6. If $G < \mathcal{S}ym(X)$ for some X, then G acts on X with associated homomorphism given by the inclusion $G \to \mathcal{S}ym(X)$. Equivalently, G acts on X by the formula $g \cdot x = g(x)$ (which makes sense since g is a function from X to itself). As a special case, we see that if $G < S_n = \mathcal{S}ym(\{1, 2, \ldots, n\})$, then G acts on $\{1, \ldots, n\}$.

Example 4.1.7. Suppose G acts on a set X. For any two sets X and Y, recall that Y^X is the set of functions from X to Y. If G acts on X, we claim that

$$g \cdot f(x) = f(g^{-1} \cdot x)$$

defines an action of G on Y^X. To see this, we first observe that if $e \in G$ is the identity, then for all $f \in Y^X$ and $x \in X$, we have

$$e \cdot f(x) = f(e^{-1} \cdot x) = f(e \cdot x) = f(x),$$

and thus $e \cdot f = f$. If $g, h \in G$, $f \in Y^X$ and $x \in X$, then

$$g \cdot (h \cdot f)(x) = h \cdot f(g^{-1} \cdot x) = f(h^{-1} \cdot (g^{-1} \cdot x)) = f(h^{-1}g^{-1} \cdot x) = f((gh)^{-1} \cdot x) = gh \cdot f(x).$$

Therefore, $g \cdot (h \cdot f) = gh \cdot f$, and so this is indeed an action of G on Y^X.

In the case that Y is a field, call it $\mathbb{F}$, the set $\mathbb{F}^X$ is a vector space (see 2.4.4). In Exercise 4.1.4 you are asked to show that in this case, the action of G on $\mathbb{F}^X$ is by *linear transformations*. That is, the action homomorphism $\alpha\colon G \to \mathcal{S}ym(\mathbb{F}^X)$ has image in $\mathrm{GL}(\mathbb{F}^X) < \mathcal{S}ym(\mathbb{F}^X)$, the invertible linear transformations from $\mathbb{F}^X$ to itself. Any action of a group G on a vector space V such that the action homomorphism has image in $\mathrm{GL}(V) < \mathcal{S}ym(V)$ is called a **representation** of G, or sometimes a **linear representation** of G (thus Exercise 4.1.4 asks you to show that α is a representation of G).

Example 4.1.8 (Left regular action)**.** If G is a group, then $G \times G \to G$ given by $g \cdot h = gh$ defies an action of G on itself—this is the content of Theorem 3.4.1. This is sometimes called the **left regular action** of G on itself. Since $gh = h$ if and only if $g = e$, we see that $\mathrm{stab}_G(h) = \{e\}$ for all $h \in G$. Consequently, the kernel is also trivial, $K = \{e\}$. For all $h, h' \in G$, setting $g = h'h^{-1}$ we have $g \cdot h = h'h^{-1}h = h'$, and hence the action is transitive: $G \cdot h = G$.

Example 4.1.9 (Conjugation action)**.** Let G be any group and define $c\colon G \to \mathrm{Aut}(G) < \mathcal{S}ym(G)$ by $c_g(h) = ghg^{-1}$. Since $c_g \circ c_{g'}(h) = g(g'hg'^{-1})g^{-1} = c_{gg'}(h)$, it follows that c is a homomorphism (compare Section 3.7), and hence c defines an action of G on itself. This action is sometimes called the **conjugation action** of G on itself. Note that with respect to this action,

$$\mathrm{stab}_G(h) = \{g \in G \mid c_g(h) = h\} = \{g \in G \mid ghg^{-1} = h\} = C_G(h),$$

that is, the stabilizer of h is the centralizer of h. The kernel of the action K is the intersection of the centralizers, and is thus the center of G, $K = Z(G)$.

The orbit of h is the set of all elements conjugate to h, this is often called the **conjugacy class of** h and is denoted

$$[h] = G \cdot h = \{ghg^{-1} \mid g \in G\}.$$

If $|G| > 1$, then there are at least two conjugacy classes, since the identity e is only conjugate to itself: $geg^{-1} = e$ for all $g \in G$. More generally, for all $h \in Z(G)$ and $g \in G$, $ghg^{-1} = h$, and so $[h] = \{h\}$. In particular, the action of G by conjugation is only transitive when $G = \{e\}$.

Example 4.1.10 (Coset action)**.** Suppose now that G is a group and $H < G$ is a subgroup. Recall that G/H is the set of left cosets of H in G. The group G acts on the set G/H by left multiplication:

$$G \times G/H \to G/H$$

given by $g \cdot aH = gaH$, for all $g, a \in G$. To prove that this an action, we verify the two axioms. Let $e, g, g', a \in G$ with e the identity. Then we have

$$e \cdot aH = eaH = aH \quad \text{and} \quad g \cdot (g' \cdot aH) = g \cdot g'aH = gg'aH = gg' \cdot aH,$$

proving that this is an action.

In Exercise 4.1.2, you are asked to prove that for all $aH \in G/H$, the stabilizer is given by $\text{stab}_G(aH) = aHa^{-1}$. Consequently, the kernel of the action, K, is the intersection of all conjugates of H That is, the kernel is the *core* of H in G (see Exercise 3.5.8). Given $aH, bH \in G/H$, setting $g = ba^{-1} \in G$, we have

$$g \cdot aH = gaH = ba^{-1}aH = bH,$$

and so the orbit of aH is all of G/H, proving that the action is transitive.

Example 4.1.10 is quite important. We will see below that any action can be broken into pieces that are essentially the same as the one just described (see Theorem 4.1.12).

Remark 4.1.11. Example 4.1.10 is really obtained from Example 4.1.8, by subsets that are permuted by the action. The example given in Exercise 4.1.3 is similarly obtained from Example 4.1.9.

In Section 4.2 we provide many more examples of group actions.

4.1.1 The orbit-stabilizer theorem

The next theorem, though quite simple, provides an invaluable tool for studying groups through their actions. Conversely, it can be used to understand actions through properties of the group.

Theorem 4.1.12 (Orbit-stabilizer Theorem)**.** † *Suppose that G acts on a set X, $x \in X$ and $H = \text{stab}_G(x)$. Then there is a bijection*

$$\theta \colon G/H \to G \cdot x$$

given by $\theta(aH) = a \cdot x$.

If $G \times G/H \to G/H$ is the action of G on the cosets of H from Example 4.1.10, then

$$\theta(g \cdot aH) = g \cdot \theta(aH).$$

Proof. We first prove that $\theta(aH) = a \cdot x$ is a well-defined map from G/H to $G \cdot x$. That is, we suppose $aH = bH$, and prove that $a \cdot x = b \cdot x$. The point is that $aH = bH$ if and only if $a^{-1}b \in H = \text{stab}_G(x)$. Then $a^{-1}b \cdot x = x$ and hence

$$a \cdot x = a \cdot (a^{-1}b \cdot x) = aa^{-1}b \cdot x = b \cdot x,$$

so θ is well-defined. Given any element $a \cdot x \in G \cdot x$, by definition $\theta(aH) = a \cdot x$, and so θ is surjective. If $\theta(aH) = \theta(bH)$, then $a \cdot x = b \cdot x$ and hence

$$a^{-1}b \cdot x = a^{-1} \cdot (b \cdot x) = a^{-1} \cdot (a \cdot x) = a^{-1}a \cdot x = e \cdot x = x,$$

so $a^{-1}b \in \text{stab}_G(x) = H$, and $aH = bH$. Consequently, θ is injective, hence a bijection.

Finally, by definition of the coset action of G on G/H, we have

$$\theta(g \cdot aH) = \theta(gaH) = ga \cdot x = g \cdot (a \cdot x) = g \cdot \theta(aH),$$

as required. This completes the proof. □

We analyze the theorem a bit below, but first we describe some applications.

Corollary 4.1.13. *Suppose that G is a finite group acting on a finite set X and $x \in X$. Then*

$$|G| = |G \cdot x||\text{stab}_G(x)|.$$

Proof. Let $H = \text{stab}_G(x)$. Then by Theorem 3.5.6, we have $|G/H| = |G|/|H|$ and by Theorem 4.1.12 $|G \cdot x| = |G/H| = |G|/|H|$. □

As an application, we obtain the so-called **class equation** for a finite group:

Theorem 4.1.14 (Class Equation)**.** *Let G be a finite group and $\mathcal{C}$ be the set of conjugacy classes of elements in G. Then*

$$|G| = \sum_{[g]\in\mathcal{C}} \frac{|G|}{|C_G(g)|} = \sum_{[g]\in\mathcal{C}} [G : C_G(g)].$$

Proof. Consider the action of G on itself by conjugation (see Example 4.1.9). Since the stabilizer of g is $C_G(g)$ and it's orbit is $[g]$, Corollary 4.1.13 implies $|G| = |[g]||C_G(g)|$. Since $\mathcal{C}$ is a partition of G, summing over each $[g] \in \mathcal{C}$ we obtain the order of G, and hence

$$|G| = \sum_{[g]\in\mathcal{C}} |[g]| = \sum_{[g]\in\mathcal{C}} \frac{|G|}{C_G(g)} = \sum_{[g]\in\mathcal{C}} [G : C_G(g)].$$

□

Note that $g \in Z(G)$ if and only if $G = C_G(g)$, which happens if and only if $[g] = \{g\}$. So, letting $\mathcal{C}_0 \subset \mathcal{C}$ denote the conjugacy classes of elements **not** in $Z(G)$, we have the following alternate formulation of the class equation.

Corollary 4.1.15 (Class Equation (second form))**.** *If G is a finite group and $\mathcal{C}_0$ is the set of conjugacy classes of elements in $G - Z(G)$, then*

$$|G| = |Z(G)| + \sum_{[g]\in\mathcal{C}_0} [G : C_G(g)].$$

Note that each term $[G : C_G(g)]$ in the sum is greater than one, and hence is a divisor of $|G|$ by Lagrange's Theorem.

If G acts transitively on X and $H = \text{stab}_G(x)$, then for any $x \in X$, $G \cdot x = X$ and θ defines a bijection $\theta\colon G/H \to X$. The last part of Theorem 4.1.12 then basically says that the action of G on G/H is "the same" as the action on X. One way to interpret this is provided by Lemma 3.4.5, which states that a bijection between sets gives rise to an isomorphism between their associated permutation groups. In the context of a transitive action, Theorem 4.1.12 provides a bijection $\theta\colon G/H \to X$, and hence an isomorphism $c_\theta\colon \mathcal{S}ym(G/H) \to \mathcal{S}ym(X)$. Let $\alpha\colon G \to \mathcal{S}ym(G/H)$ be the homomorphism defining the coset action of G on G/H and $\beta\colon G \to \mathcal{S}ym(X)$ be the homomorphism defining the original action of G on X. Since $\theta(g \cdot aH) = g \cdot \theta(aH)$ for all $aH \in G/H$ and $g \in G$, we also have $\theta^{-1}(g \cdot x) = g \cdot \theta^{-1}(x)$ for all $x \in X$ and $g \in G$. Consequently, for all $g \in G$ and $x \in X$, we have

$$c_\theta(\alpha_g)(x) = \theta\alpha_g\theta^{-1}(x) = \theta(g \cdot \theta^{-1}(x)) = \theta(\theta^{-1}(g \cdot x)) = \theta\theta^{-1}\beta_g(x) = \beta_g(x).$$

Therefore, $c_\theta \circ \alpha = \beta$.

More generally, if G acts on a set X (not necessarily transitively), then we can *restrict* the action to any orbit. That is, if $x \in X$, setting $Y = G \cdot x$, we note that G also acts on Y, since for all $g \in G$, and $a \cdot x \in Y$, we have $g \cdot (a \cdot x) = ga \cdot x \in Y$, and so there is a well-defined map

$$G \times Y \to Y.$$

Because $G \times X \to X$ satisfies the properties of an action, so does $G \times Y \to Y$. By construction, this action is transitive, and so the action of G on Y is really "the same" as the action of G on G/H, where $H = \text{stab}_G(x)$. This discussion naturally gives rise to the notion of *equivalent actions* (when two actions are "the same"), and a way in which any action $G \times X \to X$ can be *decomposed* into actions equivalent to those coming from coset actions.

Exercises.

Exercise 4.1.1. Prove Proposition 4.1.5.

Exercise 4.1.2. Let G be a group and $H < G$ a subgroup and consider the coset action $G \times G/H \to G/H$ from Example 4.1.10. Prove that the stabilizer of the coset $aH \in G/H$ is the conjugate of H by a,

$$\mathrm{stab}_G(aH) = aHa^{-1}.$$

Exercise 4.1.3. Let G be any group and recall that $\mathcal{S}(G)$ denotes the set of all subgroups of G. Let

$$G \times \mathcal{S}(G) \to \mathcal{S}(G)$$

be given by $g \cdot H = gHg^{-1}$. Prove that this is an action and that $\mathrm{stab}_G(H) = N(H)$, the normalizer of H (c.f. Exercise 3.3.8). For which groups G is this action transitive?

Exercise 4.1.4. Let G act on a set X and let $\mathbb{F}$ be any field. According to Example 4.1.7, G acts on the set of functions $\mathbb{F}^X$ from X to $\mathbb{F}$, and denote the action homomorphism by $\alpha\colon G \to \mathcal{S}ym(\mathbb{F}^X)$. Prove that for all $g \in G$, $\alpha_g\colon \mathbb{F}^X \to \mathbb{F}^X$ is a linear transformation. That is, prove that $\alpha\colon G \to \mathrm{GL}(\mathbb{F}^X)$ and so α is a representation.

Exercise 4.1.5. Consider the action of S_n on $\{1, \ldots, n\}$ by $\sigma \cdot x = \sigma(x)$ (see Example 4.1.6). According to the previous problem, the associated action of S_n on $\mathbb{R}^{\{1,\ldots,n\}} \cong \mathbb{R}^n$ is a representation, and we let $\alpha\colon S_n \to \mathrm{GL}(\mathbb{R}^n) \cong \mathrm{GL}(n, \mathbb{R})$ be the associated action homomorphism. For each $\sigma \in S_n$, we view α_σ as an $n \times n$ matrix (defining the associated linear transformation from $\mathbb{R}^n$ to itself).

a. Prove that the if $e_1, \ldots, e_n$ are the standard basis vectors of $\mathbb{R}^n$, then $\sigma \cdot e_i = \alpha_\sigma(e_i) = e_{\sigma(i)}$. That is, the i^{th} column of the matrix σ_α is $e_{\sigma(i)}$. *Hint: To unravel the definitions, note that e_i is the function $\{1, \ldots, n\} \to \mathbb{R}$ given by $e_i(j) = 1$ if $i = j$ and 0 otherwise.*

b. Using part a (and the fact that swapping two columns of a matrix exactly changes the sign of the determinant), prove that the composition $\det \circ \alpha\colon S_n \to \{\pm 1\}$ is precisely the sign homomorphism: $\epsilon = \det \circ \alpha$.

Exercise 4.1.6. Suppose $G \times X \to X$ is an action. If $H = \mathrm{stab}_G(x)$ and $g \in G$, prove that $\mathrm{stab}_G(g \cdot x) = gHg^{-1}$.

Exercise 4.1.7. Suppose $\sigma \in S_n$ is any element. Since $\langle \sigma \rangle < S_n$, it follows that $\langle \sigma \rangle$ acts on $\{1, \ldots, n\}$. Prove that the partition of $\{1, \ldots, n\}$ into the orbits of $\langle \sigma \rangle$ is precisely the partition into subsets cyclically permuted by σ determining its disjoint cycle representation.

Suppose that G acts on nonempty sets X and Y. A map $f\colon X \to Y$ is said to be **equivariant** (with respect to these actions) if $f(g \cdot x) = g \cdot f(x)$. Note that the orbit-stabilizer theorem provides an equivariant bijection $\theta\colon G/\mathrm{stab}_G(x) \to G \cdot x$.

Exercise 4.1.8. Suppose G acts transitively on X and Y, and that $f\colon X \to Y$ is equivariant with respect to these actions. Prove that f is surjective. Further prove that if X and Y are finite, then $|Y|$ divides $|X|$.

Exercise 4.1.9. Consider the subgroup $G < S_6$ given by $G = \{(1), (1\ 3), (2\ 4\ 6), (1\ 3)(2\ 4\ 6), (2\ 6\ 4), (1\ 3)(2\ 6\ 4)\}$ (you do not need to **prove** that this is a subgroup, but you should convince yourself of this fact). For the action of G on $X = \{1, 2, 3, 4, 5, 6\}$ (compare Example 4.1.6) explicitly list the orbits (without repetition). For each orbit $G \cdot x$, pick one element from the orbit (call it x), and explicitly write down the bijection from the set of cosets $G/\mathrm{stab}_G(x) \to G \cdot x$. These are each finite sets so you can just say explicitly where each coset is sent by the bijection.

You should...

- Be able to do all the exercises from this section and know the proofs of statements marked by †.
- Be familiar with the basic definitions associated with group actions (Definitions 4.1.1 and 4.1.3).
- Know the equivalent formulation of a group action (Proposition 4.1.2) and basic properties of actions (Proposition 4.1.5).
- Know the examples of actions from this section.
- Know the orbit-stabilizer theorem.

4.2 Geometric actions

In this section we describe a variety of geometrically and physically motivated group actions and use them to better understand the structure of certain groups.

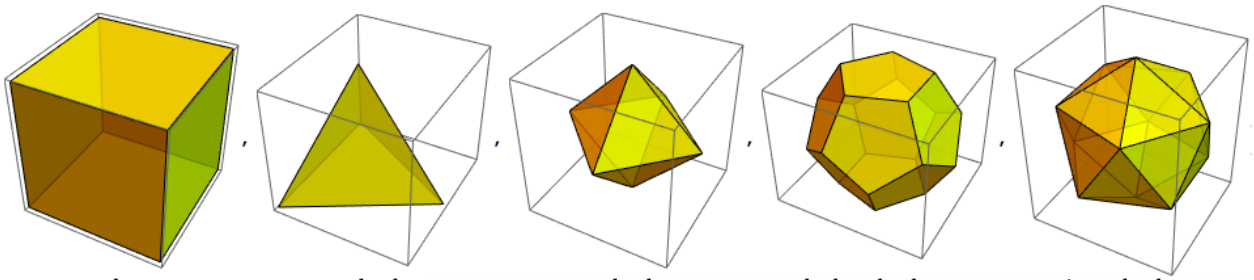
4.2.1 Platonic solids

The dihedral groups are obtained as the groups of symmetries of regular polygons. Here we consider the analogues in dimension 3.

A half-space $H \subset \mathbb{R}^3$ is a subspace defined by a linear inequality

$$H = \{(x, y, z) \in \mathbb{R}^3 \mid ax + by + cz \leq d\},$$

for some $a, b, c, d \in \mathbb{R}$. These are all points on one side of the plane $ax+by+cz = d$. We define a 3–dimensional polyhedron $P \subset \mathbb{R}^3$ to be an intersection of finitely many half-spaces which is bounded and is not contained in any plane (to rule out degenerate polyhedra). A polyhedron has finitely many vertices (corners), finitely many edges, and finitely many faces. The figure shows five important examples of polyhedra.



cube tetrahedron octahedron dodecahedron icosahedron

The five polyhedra shown are the (regular) **platonic solids**, which are the 3–dimensional analogues of regular polygons. Let $\mathrm{Isom}(P)$ denote the subgroup of isometries of $\mathbb{R}^3$ that preserve P. This group acts on the set of vertices of P, as well as the set of edges, and the set of faces. The faces of each platonic solid is a regular polygon, and these polyhedra are unique in that they have the "maximal amount of symmetry". Specifically, $\mathrm{Isom}(P)$ acts transitively on the faces, and the stabilizer of each face, which is a regular n–gon for some $n \geq 3$, is isomorphic to D_n—thus every symmetry of the face extends to a symmetry of P.

It is not difficult to see that platonic solids shown are the only polyhedra with maximal symmetry in this sense. To sketch the idea of the proof, suppose P is a polyhedron with maximal symmetry. Note that (for any polyhedron) the sum of the angles between consecutive edges around any vertex must add up to less than 2π (why?). Since there must be at least three edges around every vertex, the faces of P are n–gons with $n \leq 5$ since the angle at a vertex of regular n–gon is $\frac{\pi(n-2)}{n}$. Furthermore, for pentagon or square faces, the are exactly 3 edges around a vertex, while for triangles, there can be between 3 and 5 edges. Once the

number n of edges around a face and the number of edges meeting a vertex are determined, the polyhedron is determined by the maximal symmetry property. Thus, the five possibilities for these two parameters are precisely those for the five platonic solids.

People have been fascinated by the beauty of platonic solids for thousands of years because of their highly symmetric form. Here we determine precisely the symmetry groups of the platonic solids. For this, we first make a simplifying observation. The platonic solids occur in *dual pairs*: the cube and the octahedron are dual as are the dodecahedron and the icosahedron, while the tetrahedron is *self-dual*. What this means is that (after scaling if necessary), the each polyhedron can be inscribed inside its dual so that each vertex is situated precisely at the centroid of a face of the dual. The table below shows the number of vertices, faces, and edges of each polyhedron, which clearly reflects this duality.

	tetrahedron	cube	octahedron	dodecahedron	icosahedron
vertices	4	8	6	20	12
faces	4	6	8	12	20
edges	6	12	12	30	30

Since any element of $\mathrm{Isom}(P)$ must send the centroid of a face to the centroid of the image of the face, it follows that if we inscribe a polyhedron P' in its dual P, then the $\mathrm{Isom}(P) = \mathrm{Isom}(P')$. So up to isomorphism, there are three symmetry groups of platonic solids, and for the sake of concreteness, we pick the tetrahedron, the cube, and the dodecahedron, and let G_3, G_4, G_5 denote their respective symmetry groups (with the subscript denoting the number of edges around a face in each—thus $D_n < G_n$, for each $n = 3, 4, 5$). We will analyze these groups using group actions. The first is the simplest.

Proposition 4.2.1. $G_3 \cong S_4$.

Proof. Let P be the tetrahedron, and consider the action of G_3 on the set X consisting of the 4 faces of P. Since G_3 acts transitively on X and the stabilizer of a face is isomorphic to D_3, Corollary 4.1.13 implies

$$|G_3| = |X||D_3| = 4 \cdot 6 = 24.$$

On the other hand, G_3 acts on the set of vertices, with trivial kernel (any symmetry that fixes all the vertices is the identity). Thus we have an injective homomorphism $G_3 \to S_4$. Since $|S_4| = 24$, it follows that this is also surjective, hence an isomorphism. □

Recall that $\mathrm{Isom}(\mathbb{R}^k) = \{\Phi_{A,v} \mid A \in \mathrm{O}(k), v \in \mathbb{R}^k\}$. We let

$$\mathrm{Isom}^+(\mathbb{R}^k) = \{\Phi_{A,v} \mid A \in \mathrm{SO}(k), v \in \mathbb{R}^k\}.$$

We call this the subgroup of *orientation preserving isometries*. This is an index two subgroup: indeed, the function $\Phi_{A,v} \mapsto \det(A)$ defines a surjective homomorphism $\mathrm{Isom}(\mathbb{R}^k) \to \{\pm 1\}$, and $\mathrm{Isom}^+(\mathbb{R}^k)$ is the kernel. We let $G_4^+ < G_4$ and $G_5^+ < G_5$ denote the index two (normal) subgroups defined by

$$G_n^+ = G_n \cap \mathrm{Isom}^+(\mathbb{R}^3).$$

In fact, if we assume that the polyhedra have their centroids at the origin (which we can arrange by translation), then we can view G_n as a subgroup of $\mathrm{O}(3)$ and $G_n^+ = G_n \cap \mathrm{SO}(3)$. It turns out G_4 and G_5 are easy to describe once we know G_4^+ and G_5^+.

Proposition 4.2.2. $G_4^+ \cong S_4$ *and* $G_4 \cong S_4 \times \{\pm 1\}$.

Proof. Let P be a cube with centroid at the origin. As for the tetrahedron, we can compute $|G_4|$ and hence $|G_4^+| = |G_4|/2$ using the action of G_4 on the set X of 6 faces of P. From Corollary 4.1.13 we have

$$|G_4| = |X||D_4| = 6 \cdot 8 = 48,$$

and so $|G_4^+| = 24$. To prove $G_4^+ \cong S_4$, we need a set with four elements that G_4^+ can act on with trivial kernel. This is provided by the set Y of diagonals of the cube as shown in the figure below.

In G_4, there are exactly two elements that fix all the diagonals: the identity I and $-I$ (that is, $(x, y, z) \mapsto (-x, -y, -z)$). Of these two, only I is in G_4^+, and hence the action of G_4 on Y has trivial kernel. Since $|Y| = 4$, this action yields an isomorphism $G_4^+ \to S_4$, proving the first part.

For the second part, note that $-I \in G_4 \setminus G_4^+$ and $\langle -I \rangle = \{I, -I\} \cong \{\pm 1\}$. Furthermore, $\langle -I \rangle \cap G_4^+ = \{I\}$ and since $-I$ is central in $\mathrm{O}(3)$ (hence in G_4), we see that both $\langle -I \rangle$ and G_4^+ are normal subgroups. Therefore, by Proposition 3.7.1, we have

$$G_4 \cong G_4^+ \times \langle -I \rangle = G_4^+ \times \{\pm 1\}.$$

□

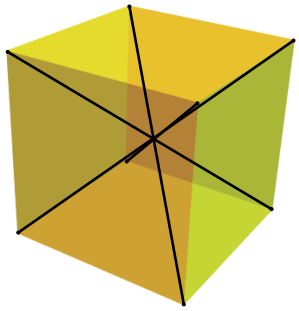
Proposition 4.2.3. *$G_5^+ \cong A_5$ and $G_5 \cong A_5 \times \{\pm 1\}$.*

Proof. The idea for the proof is similar to the previous two cases. Since G_5 acts transitively on the set of 12 faces, and the stabilizer of a face is isomorphic to D_5, Corollary 4.1.13 implies

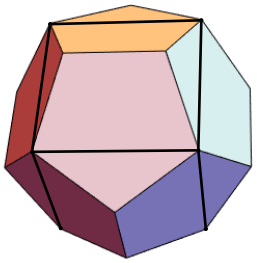
$$|G_5| = 12|D_5| = 120,$$

and consequently $|G_5^+| = 60$.

To prove that $G_5^+ \cong A_5$ (and $G_5 \cong A_5 \times \{\pm 1\}$), we note that there are 5 cubes inscribed in the dodecahedron which are permuted by G_5. The visible edges of one such cube are shown in the figure below. (Alternatively, there is a partition of the 30 edges into 5 subsets which are also preserved). The action of G_5^+ is easily seen to have trivial kernel, and hence this provides an injective homomorphism $G_5^+ \to S_5$. Since $|G_5^+| = 60 = |S_5|/2$, Theorem 3.5.6 implies that the image has index 2 inside S_5. According to Exercise 4.2.2, the image must be A_5, proving $G_5^+ \cong A_5$. The fact that $G_5 \cong A_5 \times \{\pm 1\}$ follows as in the proof of Proposition 4.2.2. □



The 4 diagonals of a cube



The visible edges of one of the five cubes inscribed inside a dodecahedron.

4.2.2 Rubik's cube group

The Rubik's cube group was defined in Example 3.4.7 as a subgroup $\mathcal{R}u < S_{48}$, consisting of the permutations of the 48 facets generated by the 6 legal moves F, B, R, L, U, D. The Rubik's cube group has the following structure.

Theorem 4.2.4.

$$\mathcal{R}u \cong (\mathbb{Z}_3^7 \times \mathbb{Z}_2^{11}) \rtimes ((A_8 \times A_{12}) \rtimes \mathbb{Z}_2).$$

We will sketch a proof of this below. For details see, for example

http://www.math.harvard.edu/~jjchen/docs/Group%20Theory%20and%20the%20Rubik's%20Cube.pdf

First, let us make an interesting observation.

Observation 4.2.5. There exist Rubik's cube configurations that have never been realized on any Rubik's cube, anywhere in the world, ever.

"Proof". This is based on a few *reasonable* assumptions. The Rubik's cube was first sold in 1980, and let's suppose that every person on Earth, from 1980 until 2016 has been performing an average of no more than 3 Rubik's cube moves per second, every second of their life. The current estimate is that the population of the Earth is about 7.4 billion people (in fact it was just over half this in 1980). So, we can bound the number of Rubik's cube moves performed each second on Earth over the past 36 years as $3 \times 7.4 \times 10^9 = 22.2 \times 10^9$. Thus, the number of moves that have been performed on Rubik's cubes is less than

$$22.2 \times 10^9 \times 365 \times 24 \times 60 \times 60 \approx 7.0 \times 10^{17}.$$

On the other hand, $|\mathcal{R}u| = |3^7 2^{12} \frac{8!}{2} \frac{12!}{2}| \approx 4.3 \times 10^{19}$. Therefore, the estimates above imply that not every element of $\mathcal{R}u$ has actually been physically performed on a Rubik's cube. □

The proof of Theorem 4.2.4 will appeal to group actions, which requires some additional notation and set-up. Note that the Rubik's cube is made up of 27 smaller **cubies**: there are 20 moving cubies consisting of the 8 **corner cubies** and the 12 **edge cubies** as well as the 6 **center cubies** and the "hidden" **middle cubie**. Let C denote the set of corner cubies and E the set of edge cubies. Each element of $\mathcal{R}u$ permutes the cubies sending corner cubies to corner cubies and edge cubies to edge cubies. Thus, $\mathcal{R}u$ acts on each of the sets C and E, so we have homomorphisms $\phi_c \colon \mathcal{R}u \to \mathcal{S}ym(C) \cong S_8$ and $\phi_e \colon \mathcal{R}u \to \mathcal{S}ym(E) \cong S_{12}$. We can take these two homomorphisms together into the product,

$$\phi = \phi_c \times \phi_e \colon \mathcal{R}u \to \mathcal{S}ym(C) \times \mathcal{S}ym(E) \cong S_8 \times S_{12},$$

(or alternatively, think of this as an action on the set of moveable cubies $C \cup E$ preserving the partition $\{E, C\}$; see Exercise 4.2.4). Define $\epsilon^2 = \epsilon \cdot \epsilon \colon S_8 \times S_{12} \to \{\pm 1\}$ by taking the product of the values of the sign homomorphisms on each of the factors (see Exercise 4.2.5).

Lemma 4.2.6. $\ker(\epsilon^2) \cong (A_8 \times A_{12}) \rtimes \mathbb{Z}_2$.

Proof. The product of the sign homomorphisms into the product is a surjective homomorphism

$$\epsilon \times \epsilon \colon S_8 \times S_{12} \to \{\pm 1\} \times \{\pm 1\}$$

and the surjective homomorphism ϵ^2 factors through $\epsilon \times \epsilon$

$$S_8 \times S_{12} \to \{\pm 1\} \times \{\pm 1\} \to \{\pm 1\}.$$

So, $A_8 \times A_{12} = \ker(\epsilon \times \epsilon) < \ker(\epsilon^2)$. In fact, $\ker(\epsilon^2) = (\epsilon \times \epsilon)^{-1}(\{(1,1),(-1,-1)\})$.

Since $[S_8 \times S_{12} : \ker(\epsilon^2)] = 2$ and $[S_8 \times S_{12} : A_8 \times A_{12}] = 4$ (why?), $[\ker(\epsilon^2) : A_8 \times A_{12}] = 2$ by Corollary 3.5.9. On the other hand, $\sigma = ((1\ 2), (1\ 2)) \in \ker(\epsilon^2) - \ker(\epsilon \times \epsilon)$ and $|\sigma| = 2$, so by Proposition 3.7.7

$$\ker(\epsilon^2) \cong (A_8 \times A_{12}) \rtimes \langle \sigma \rangle \cong (A_8 \times A_{12}) \rtimes \{\pm 1\} \cong (A_8 \times A_{12}) \rtimes \mathbb{Z}_2.$$

□

Lemma 4.2.7. *The image of $\phi_c \times \phi_e$ is contained in* $\ker(\epsilon^2)$.

Proof. See Exercise 4.2.6. □

Proposition 4.2.8. *There is a subgroup $\mathcal{R}u_p < \mathcal{R}u$ such that the restriction to $\mathcal{R}u_p$ of $\phi_c \times \phi_e$,*

$$\phi_c \times \phi_e|_{\mathcal{R}u_p} \colon \mathcal{R}u_p \to \ker(\epsilon^2),$$

is an isomorphism. In particular, $\mathcal{R}u_p \cong (A_8 \times A_{12}) \rtimes \mathbb{Z}_2$.

Sketch of proof. To describe this subgroup, we first choose a "preferred facet" for each of the cubies in C and E. There are many choices, but we use one described as follows. First, each corner cubie has either a facet on the top face or a facet on the bottom face, and we choose these to be the preferred facets on the corner cubies. For the edge cubies, we similarly chose the facet on the top face or the facet on the bottom face if one exists. This takes care of 8 of the 12 edge cubies. The other four have a facet on either the front or the back face, and in this situation, we choose these as the preferred facet.

Now define

$$\mathcal{R}u_p = \{\sigma \in \mathcal{R}u \mid \sigma \text{ sends each preferred facet of a cubie to a preferred facet }\}.$$

It is easy to see that this is a subgroup, and moreover,

$$\ker(\phi_c \times \phi_e|_{\mathcal{R}u_p}) = \ker(\phi_c \times \phi_e) \cap \mathcal{R}u_p = \{id\}.$$

This is because any $\sigma \in \ker(\phi_c \times \phi_e)$ sends each cubie to itself, and if σ also sends the preferred facet of that cubie to itself, it must be the identity. Therefore, we are left to prove that $\phi_c \times \phi_e|_{\mathcal{R}u_p}$ is onto $\ker(\epsilon^2)$.

We can explicitly list the preferred facets. On the cubies in C these are $C' = \{1, 3, 6, 7, 41, 43, 46, 48\}$, and on those in E they are $E' = \{2, 4, 5, 7, 23, 24, 27, 28, 42, 44, 45, 47\}$; see the figure from Example 3.4.7. By definition, $\mathcal{R}u_p$ acts on C' and E', written $\phi_{c'} \colon \mathcal{R}u_p \to \mathcal{S}ym(C')$ and $\phi_{e'} \colon \mathcal{R}u_p \to \mathcal{S}ym(E')$. These actions are obtained from ϕ_c and ϕ_e, respectively, by composing with the isomorphisms $\mathcal{S}ym(C) \to \mathcal{S}ym(C')$ and $\mathcal{S}ym(E) \to \mathcal{S}ym(E')$ determined by the bijections $C \to C'$ and $E \to E'$ sending a cubie to its preferred facet (see Lemma 3.4.5). Thus, we may consider $\phi_{c'} \times \phi_{e'}$ instead of $\phi_c \times \phi_e$ in the rest of the proof.

Now consider the following element

$$\tau_0 = F^2URL^{-1}F^2R^{-1}LUF^2,$$

(recall that we compose elements right-to-left, and so this is obtained by first performing F^2, then U, then L, then R^{-1}...) This permutes three of the edge cubies in the top, sending the preferred facets to preferred facets, and hence $\tau_0 \in \mathcal{R}u_p$. With respect to the actions $\phi_{c'}$ and $\phi_{e'}$ we have

$$\phi_{e'}(\tau_0) = (7\ 4\ 5) \text{ and } \phi_{c'}(\tau_0) = id.$$

Conjugating by appropriate elements, the image by $\phi_{e'}$ gives all 3–cycles of the form $(7\ 4\ a)$ for every $a \in E' - \{4, 7\}$. For example, $U\tau_0U^{-1}, L\tau_0L^{-1} \in \mathcal{R}u_p$ and

$$\phi_{e'}(U\tau_0U^{-1}) = (7\ 4\ 2), \phi_{c'}(U\tau_0U^{-1}) = id \text{ and } \phi_{e'}(L\tau_0L^{-1}) = (7\ 4\ 23), \phi_{c'}(L\tau_0L^{-1}) = id.$$

According to Exercise 4.2.7, these 3–cycles generate A_{12}, and hence $\phi_{c'} \times \phi_{e'}(\mathcal{R}u_p)$ contains $\{id\} \times A_{12}$.

Next, consider the element

$$\tau_1 = U^{-1}R^2B^2R^{-1}F^{-1}RB^2R^{-1}FR^{-1}.$$

One can check that this also permutes only cubies in the top, and lies in $\mathcal{R}u_p$, acting as

$$\phi_{e'}(\tau_1) = (2\ 5\ 7\ \ 4) \text{ and } \phi_{c'}(\tau_1) = (6\ 8).$$

Note that $\epsilon(\phi_{e'}(\tau_1)) = \epsilon(\phi_{c'}(\tau_1)) = -1$, so $\phi_{c'} \times \phi_{e'}(\tau_1) \notin A_8 \times A_{12}$. Another computation shows

$$\phi_{e'}((U\tau_1U^{-1}\tau_1^{-1})) = id \text{ and } \phi_{c'}(U\tau_1U^{-1}\tau_1^{-1}) = (1\ 6\ 8).$$

Arguing as above, the $\phi_{c'}$–images of appropriate conjugates of $U\tau_1U^{-1}\tau_1^{-1}$ gives all 3–cycles of the form $(1\ 6\ a)$ for $a \in C' - \{1, 6\}$. Therefore, $\phi_{c'} \times \phi_{e'}(\mathcal{R}u_p)$ contains $A_8 \times \{id\}$. Combining this with the above, we see that $\phi_{c'} \times \phi_{e'}(\mathcal{R}u_p)$ properly contains $A_8 \times A_{12}$. Since the image is contained in $\ker(\epsilon^2)$ which contains $A_8 \times A_{12}$ with index 2, it follows that $\phi_{c'} \times \phi_{e'}(\mathcal{R}u_p) = \ker(\epsilon^2)$, as required. □

This gets us part of the way to Theorem 4.2.4:

Corollary 4.2.9.

$$\mathcal{R}u \cong \ker(\phi_c \times \phi_e) \rtimes \mathcal{R}u_p \cong \ker(\phi_c \times \phi_e) \rtimes ((A_8 \times A_{12}) \rtimes \mathbb{Z}_2).$$

Proof. The first isomorphism follows form Propositions 3.7.7 and 4.2.8. The second is Lemma 4.2.6. □

The next proposition will complete the proof of Theorem 4.2.4.

Proposition 4.2.10. $\ker(\phi_c \times \phi_e) \cong \mathbb{Z}_3^7 \times \mathbb{Z}_2^{11}$.

Sketch of proof. Let C'' be the set of all facets of corner cubies and E'' the set of all facets of edge cubies (so $C' \subset C''$ and $E' \subset E''$), and observe that $\mathcal{R}u$ acts on both C'' and E''. Define two functions $x\colon C'' \to \mathbb{Z}_3$ and $y\colon E'' \to \mathbb{Z}_2$ as follows. For each corner cubie, declare x to take value $0 \in \mathbb{Z}_3$ on the preferred facet, value $1 \in \mathbb{Z}_3$ for the next facet clockwise from the preferred facet, and value $2 \in \mathbb{Z}_3$ for the remaining facet in that cubie. Likewise, for each edge cubie, we require y to take value $0 \in \mathbb{Z}_2$ on the preferred facet and value $1 \in \mathbb{Z}_2$ on the other facet.

Recall from Example 4.1.7 that the action of G on C'' and E'' determine actions of $\mathcal{R}u$ on $\mathbb{Z}_3^{C''}$ and $\mathbb{Z}_2^{E''}$, respectively, by $\sigma \cdot \psi(z) = \psi(\sigma^{-1} \cdot z)$. Now we claim that for all $\sigma \in \mathcal{R}u$, we have the following two identities

$$\sum_{z \in C'} \sigma \cdot x(z) = 0 \in \mathbb{Z}_3 \quad \text{and} \quad \sum_{z \in E'} \sigma \cdot y(z) = 0 \in \mathbb{Z}_2.$$

We explain the proof for the sum on the left, and leave the one on the right to the reader.

First, observe that by definition

$$\sum_{z \in C'} \sigma \cdot x(z) = \sum_{z \in C'} x(\sigma^{-1} \cdot z).$$

If we think of the function x as assigning a label to each facet in C'', then this sum can be thought of as the sum over each preferred facets $z \in C'$ of the label on the facet *sent by* σ to z. That is, we apply σ, look at the labels appearing on each preferred facet, then sum these up. Now write σ as a composition of the generators R, L, U, D, F, B and their inverses, say with $n \geq 1$ elements in the product. We prove that the sum is 0 by induction on n. For $n = 0$, $\sigma = id$, and the sum is zero by definition of x. Suppose the sum is zero for any composition of $n-1$ of the generators and their inverses and suppose that $\sigma = g\sigma'$, where $g \in \{R^{\pm 1}, L^{\pm 1}, U^{\pm 1}, D^{\pm 1}, F^{\pm 1}, B^{\pm 1}\}$. By assumption, after applying σ', the sum of the labels appearing in the preferred facets is 0. Now we consider what happens when we further apply g. Observe that $U^{\pm 1}$ and $D^{\pm 1}$ preserve the preferred facets, so the sum is unchanged if $g \in \{U^{\pm 1}, D^{\pm 1}\}$. Next, consider $g = F$ (the other cases are similar, so we content ourselves with this case). Whatever labels we see in the preferred facets $1, 3, 46, 48$ after applying σ' are unchanged by F, while the sum of the labels appearing $6, 8, 41, 43$ will be replaced by the sum of the labels appearing in $11, 15, 31, 35$. Note that 15 and 31 are clockwise from 8 and 41, respectively, and 11 and 35 are counterclockwise from 6 and 43, respectively, and thus two will increase by 1 and two will decrease by 1 (mod 3). Therefore, there is no change in the sum, and consequently it is still zero after applying F. The other generators are similar, and thus the sum is zero for any g.

Now we turn to the subgroup $\ker(\phi_c \times \phi_e)$. The elements of $\ker(\phi_c \times \phi_e)$ preserve each cubie, possibly rotating it. If we let $G < S_{48}$ be the group obtained by allowing all rotations of the edge and corner cubies (but again leaving each cubie invariant), we see that $\ker(\phi_c \times \phi_e) < G$ and

$$G \cong \mathbb{Z}_3^8 \times \mathbb{Z}_2^{12},$$

(why?). Now, G also acts on C'' and E'', hence on the sets of functions as above. For each $\sigma \in G$, the two sums on $\mathcal{R}u$ we described above actually become surjective homomorphisms

$$\delta_c(\sigma) = \sum_{z \in C'} \sigma \cdot x(z) \text{ and } \delta_e(\sigma) = \sum_{z \in E'} \sigma \cdot y(z).$$

In fact, an isomorphism $G \cong \mathbb{Z}_3^8 \times \mathbb{Z}_2^{12}$ as above can be given by

$$\sigma \mapsto ((\sigma \cdot x(z))_{z \in C'}, (\sigma \cdot y(z))_{z \in E'})$$

where $\sigma \in G$ and the expression on the right is an 8+12–tuple in $\mathbb{Z}_3^8 \times \mathbb{Z}_2^{12}$. With respect to this isomorphism, δ_c and δ_e are obtained by adding up the first 8 and the last 12 entries, respectively.

Since both sums for $\sigma \in \mathcal{R}u$ are 0, it follows that $\ker(\phi_c \times \phi_e) < \ker(\delta_c) \cap \ker(\delta_e)$, and so $\ker(\phi_c \times \phi_e)$ has index at least 6 inside G. On the other hand, if we consider the elements

$$\tau_3 = U^2B^{-1}RD^2R^{-1}BU^2B^{-1}RD^2R^{-1}B \quad \text{and} \quad \tau_4 = U^{-1}R^{-1}D^{-1}B^2UBUB^{-1}U^2B^2DUR,$$

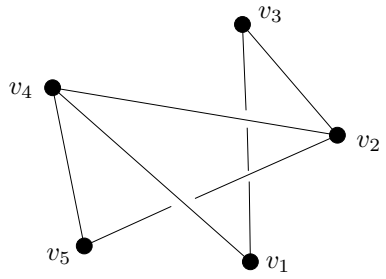
then one can show that τ_3 twists two corner cubies in opposite directions (fixing all other cubies) while the second twists a pair of edge cubies (fixing all other cubies). In particular, these are elements of $\ker(\phi_c \times \phi_e)$, and considering appropriate conjugates of these, it is easy to see that $\ker(\phi_c \times \phi_e)$ has index exactly 6 inside G and is isomorphic to the product $\mathbb{Z}_3^7 \times \mathbb{Z}_2^{11}$. □

The proof above provides the tools necessary for the most difficult steps in solving the Rubik's cube. Indeed, it is relatively simple to solve one layer of the Rubik's cube (that is, to arrange that all cubies meeting any particular face are in the solved position). The challenge then becomes moving the other cubies into the solved position without mixing up the ones already solved. In the proofs we have listed several mysterious elements of $\mathcal{R}u$ which we can check permute a controlled few cubies (and leave the others unchanged). These are precisely the kinds of elements needed to move the unsolved cubies to the solved positions leaving alone those cubies that are already solved. Can you find other elements of $\mathcal{R}u$ that do permute only a few cubies, or that do other interesting things?

4.2.3 Cayley graph

A **graph** Γ is a combinatorial object consisting of a set $V = V(\Gamma)$ called the **vertex set** of Γ together with a set $E = E(\Gamma) \subset V \times V$ called the **edge set** of Γ with the property that if $(v_1, v_2) \in E$, then $(v_2, v_1) \in E$. A **geometric realization** of a graph Γ is obtained by identifying the vertex set V with a subset of $\mathbb{R}^n$ and connecting v and v' by an arc when $(v, v') \in E$ (with any two such distinct arcs meeting at most in their endpoints). For example, the figure below shows a geometric realization of the graph Γ with $V(\Gamma) = \{v_1, v_2, v_3, v_4, v_5\}$ and

$$E(\Gamma) = \{(v_1, v_3), (v_3, v_1), (v_1, v_4), (v_4, v_1), (v_2, v_3), (v_3, v_2), (v_2, v_4), (v_4, v_2), (v_2, v_5), (v_5, v_2), (v_4, v_5), (v_5, v_4)\}.$$



We will think of Γ as either the combinatorial data of $V(\Gamma)$ and $E(\Gamma)$, or the geometric realization. With this view, we define an action of group on a graph in two different (but equivalent) ways. We can either think of an action of a group G on a graph Γ as an action on the geometric realization so that each $g \in G$ defines a *continuous* function $g\colon \Gamma \to \Gamma$ sending the vertex set to itself, *or*, as an action of G on the set V with the property that for all $v, v' \in V$ and $g \in G$,

$$(v, v') \in E \text{ if and only if } (g \cdot v, g \cdot v') \in E.$$

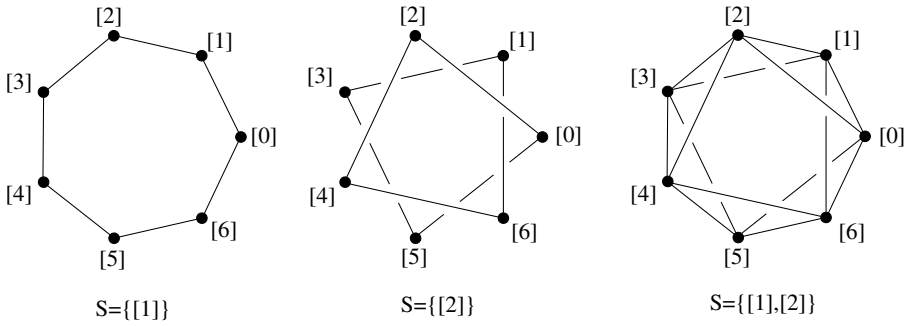
As we now explain, every group acts on a graph, and so we can view the group geometrically. Suppose G is a group and S is a **generating set** for G: that is, $S \subset G$ is a subset so that $G = \langle S \rangle$ (see Section 3.2).

There is always a generating set $S = G$, but that is usually not particularly interesting. Note that if S is a generating set, then every non-identity element $g \in G$ can be expressed in the form $g = s_1^{\epsilon_1} s_2^{\epsilon_2} \cdots s_n^{\epsilon_n}$, for some $n \geq 1$, $s_1, \ldots, s_n \in S$, and $\epsilon_1, \ldots, \epsilon_n \in \{\pm 1\}$.

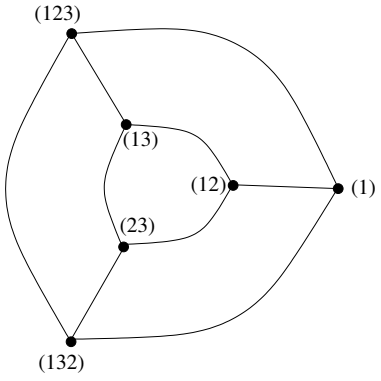
From a group G and a generating set S, define the **Cayley graph of G with respect to S**, denoted $\Gamma(G, S)$ as follows. The vertex set *is* the group, $V = G$, and the edge set is defined by

$$E = \{(g, gs^\epsilon) \mid g \in G\,,\ s \in S\,,\ \text{and } \epsilon \in \{\pm 1\}\}.$$

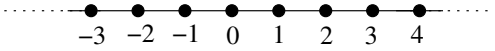
Note that if $(g, gs^\epsilon) \in E$, then $(gs^\epsilon, g) = (gs^\epsilon, (gs^\epsilon)s^{-\epsilon}) \in E$, so V, E defines a graph. Furthermore, supposing $(g, gs^\epsilon) \in E$ and $g' \in G$ note that $(g'g, g'gs^\epsilon) \in E$, and hence G acts on $\Gamma(G, S)$, acting on $V = G$ simply by $g' \cdot g = g'g$ (i.e. by the left regular action). The Cayley graphs can look very different depending on the generating set. For example, the figure below shows three Cayley graphs for the group $G = \mathbb{Z}_7$ with respect to the generating sets $\{[1]\}, \{[2]\}, \{[1], [2]\}$. In each case, the action of $[a] \in \mathbb{Z}_7$ can be visualized geometrically as rotation through an angle $2\pi a/7$ on the vertices (sending edges continuously to edges).

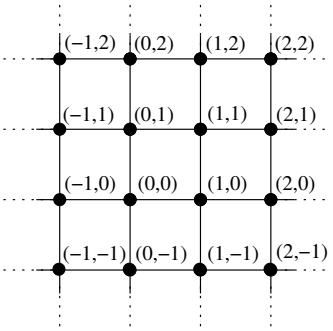


Another example is S_3 with generating set $S = \{(1\ 2), (1\ 2\ 3)\}$ (see Exercise 3.4.3). The Cayley graph with respect to this generating set is drawn in the next figure.



Even when G is infinite, it may have a finite generating set. The next two examples show (parts of) the Cayley graphs for $\mathbb{Z}$ with respect to the generating set $\{1\}$ and $\mathbb{Z} \times \mathbb{Z}$ with respect to the generating set $\{(1, 0), (0, 1)\}$.





Exercises.

Exercise 4.2.1. For each $m \geq 3$, let $n = 3m$, and consider the group $D_n = D_{3m}$. Prove that there is a normal subgroup isomorphic to $\mathbb{Z}_m$. *Hint: Connecting every third vertex in the n–gon P_n produces a regular m-gon. Prove that there are three such, and that D_n acts on this set, then analyze the kernel.*

Exercise 4.2.2. Prove that $A_n < S_n$ is the unique subgroup of index two. *Hint: For this, first suppose that $N < S_n$ is a subgroup of index 2, which by Exercise 3.5.4 is a normal subgroup. Consider the quotient homomorphism $\pi\colon S_n \to S_n/N$ and prove that any 2–cycle must have nontrivial image in S_n/N (Exercise 3.4.1 is helpful here). Now appeal to Propositions 1.3.9 and 3.4.8 to deduce that $N = \ker(\pi) = A_n$.*

Exercise 4.2.3. Edges e, e' of a tetrahedron T are said to be *opposite* if they are disjoint (that is, they do not share a vertex). The 6 edges can be partitioned into a set X of three pairs of opposite edges. Prove that G_3, the group of symmetries of T, acts on X and the kernel $K < G_3$ is a normal subgroup of order 4.

The previous exercise is interesting in that it provides a geometric description of a normal subgroup of $G_3 \cong S_4$ different than A_4. It turns out that for any $n \neq 4$, the only normal subgroup of S_n is A_n (and $\{(1)\}$ and S_n, of course.

Exercise 4.2.4. Suppose X is a set and $\{Y, Z\}$ is a partition of X with two elements: that is, $Y, Z \subset X$ are nonempty subsets, $X = Y \cup Z$, and $Y \cap Z = \emptyset$. Prove that $H = \{\sigma \in \mathcal{S}ym(X) \mid \sigma(Y) = Y \text{ and } \sigma(Z) = Z\}$ is a subgroup of $\mathcal{S}ym(X)$ and $H \cong \mathcal{S}ym(Y) \times \mathcal{S}ym(Z)$.

Exercise 4.2.5. Suppose G_1, G_2 are groups and $\phi_i\colon G_i \to H$ is a homomorphism to an abelian group, for each $i = 1, 2$. Prove that $\phi_1 \cdot \phi_2\colon G_1 \times G_2 \to H$ given by $\phi_1 \cdot \phi_2(g_1, g_2) = \phi_1(g_1)\phi_2(g_2)$ defines a homomorphism.

Exercise 4.2.6. Prove that the image of $\phi_c \times \phi_e$ is contained in $\ker(\epsilon^2)$. *Hint: Prove that for each generator $s \in \{R, L, U, D, F, B\}$, $\phi_c \times \phi_e(s) \in \ker(\epsilon^2)$, then prove that this implies $\phi_c \times \phi_e(\mathcal{R}_u) < \ker(\epsilon^2)$.*

Exercise 4.2.7. Suppose $n \geq 3$ and let $A_n \triangleleft S_n$ be the alternating group (the kernel of the sign homomorphism, i.e. the subgroup of even permutation).

a. Prove that A_n is generated by the set of 3–cycles. That is, prove that every element of A_n is a composition of 3–cycles. *Hints: Given an element $\sigma \in A_n$, first write it as a composition of an* even number *of 2–cycles (why can you do that?). Observe that a composition of two 2–cycles is either a 3–cycle or a composition of two disjoint 2–cycles (explain why). In the latter case, figure out a way to write a composition of two disjoint 2–cycles as a composition of 3–cycles. Combine all these facts to prove A_n is generated by 3–cycles.*

b. Prove that A_n is generated by 3–cycles $\{(1\ 2\ 3), (1\ 2\ 4), (1\ 2\ 5), \ldots, (1\ 2\ n)\}$. (This means you don't need *all* the 3–cycles to generate A_n.)

Exercise 4.2.8. Draw the Cayley graph of D_4 with respect to the generating set j, r. See Section 3.4.

Exercise 4.2.9. Prove that $S = \{2, 3\}$ is a generating set for $\mathbb{Z}$ and draw (a part of) the Cayley graph for $\mathbb{Z}$ with respect to S.

You should...

- Be able to do all the exercises from this section.
- Know the examples of actions from this section.
- Be able to spot group actions, and use them together with the orbit-stabilizer theorem to compute orders of group, and find structure/identify groups, up to isomorphism.

4.3 Sylow Theorems

This section provides some deep and important applications of group action to the *general structure of finite groups*. The three main theorems are called the **Sylow theorems**, and like Lagrange's Theorem (Theorem 3.5.6), these theorems inform us about the structure of subgroups, specifically, subgroups of prime-power order. In fact, one of the theorems (Theorem 4.3.6), provides a partial converse of Lagrange's Theorem by asserting the existence of subgroups of all prime-power orders which divide the order of the group.

Before we set out, we make a few definitions.

Definition 4.3.1. *If $p \geq 2$ is a prime, then a p–***group** *is a group such that every element has order which is a power of p.*

We begin with *Cauchy's Theorem*, which tells us that every prime divisor of the order of a group is the order of some element. In particular, this implies that a finite p–group has order a power of p (see Corollary 4.3.4 below).

Theorem 4.3.2 (Cauchy's Theorem)**.** *If G is a finite group and $p \geq 2$ is a prime dividing the order of G, then there exists an element $g \in G$ such that $|g| = p$.*

The proof relies on the following lemma which we will use multiple times in what follows. To state it, we suppose G is a group acting on a set X, and denote the set of elements fixed by every element of G as

$$\mathrm{Fix}_X(G) = \{x \in X \mid g \cdot x = x \text{ for all } g \in G\}.$$

Lemma 4.3.3. *Suppose $p \geq 2$ is a prime and G is a group of order p^k for some $k \geq 1$. For any action of G on a finite set X we have*

$$|\mathrm{Fix}_X(G)| \equiv |X| \pmod{p}.$$

Proof. Note that $x \in \mathrm{Fix}_X(G)$ if and only if $\mathrm{stab}_G(x) = G$, or equivalently, if and only if $G \cdot x = \{x\}$. We also recall that $\mathcal{O}_G(X)$, the set of G–orbits of X, is a partition of X. So, by Theorem 4.1.12

$$|X| = \sum_{G\cdot x \in \mathcal{O}_G(X)} |G \cdot x| = \sum_{G\cdot x \in \mathcal{O}_G(X)} \frac{|G|}{|\mathrm{stab}_G(x)|}.$$

The sum on the right splits into a sum over the orbits with exactly one element (i.e. the orbits of elements of $\mathrm{Fix}_X(G)$), and all the other orbits, which we denote $\mathcal{O}_G^\circ(X) = \{G \cdot x \mid x \notin \mathrm{Fix}_X(G)\}$:

$$|X| = \sum_{G\cdot x \in \mathcal{O}_G^\circ(X)} \frac{|G|}{|\mathrm{stab}_G(x)|} + |\mathrm{Fix}_X(G)|.$$

Now, observe that for all $x \notin \operatorname{Fix}_X(G)$, we have $\operatorname{stab}_G(x)$ is a proper subgroup of G, and hence $\frac{|G|}{|\operatorname{stab}_G(x)|}$ is a positive power of p (hence divisible by p). Therefore, p divides their sum and so

$$p \Big| \left(\sum_{G \cdot x \in \mathcal{O}_G^\circ(X)} \frac{|G|}{|\operatorname{stab}_G(x)|} \right) = |X| - |\operatorname{Fix}_X(G)|,$$

and hence $|X| \equiv |\operatorname{Fix}_X(G)| \bmod p$, proving the lemma. □

Proof of Theorem 4.3.2. Write $n = |G|$ and assume $p \geq 2$ is a prime with $p|n$. Let

$$G^p = \{(g_1, \ldots, g_p) \mid g_i \in G \text{ for all } i\}.$$

We will consider the indices i of the entries g_i as an element of $\mathbb{Z}_p$ (that is, we are actually viewing i as $[i]_p \in \mathbb{Z}_p$). This is helpful in that it allows us to define an action of $\mathbb{Z}_p$ on G^p by

$$a \cdot (g_1, \ldots, g_p) = (g_{a+1}, \ldots, g_{a+p})$$

for all $a \in \mathbb{Z}_p$ (again, we are choosing a representative a of the congruence class, but it is easy to see that if $a \equiv b \bmod p$, then $a \cdot (g_1, \ldots, g_p) = b \cdot (g_1, \ldots, g_p)$). In Exercise 4.3.1 you are asked to show that this is an action, and moreover, it preserves the subset

$$X = \{(g_1, \ldots, g_p) \in G^p \mid g_1 g_2 \cdots g_p = e\} \subset G^p,$$

so that G acts on X (by the same formula as above).

Now observe that

$$\operatorname{Fix}_X(\mathbb{Z}_p) = \{(g, g, \ldots, g) \in X\} = \{(g, \ldots, g) \mid g^p = e\}.$$

Clearly $(e, e, \ldots, e) \in \operatorname{Fix}_X(\mathbb{Z}_p)$, and we want to see that there is at least one other element. By Lemma 4.3.3, $|\operatorname{Fix}_X(\mathbb{Z}_p)| \equiv |X| \bmod p$, while on the other hand, we have a bijection $G^{p-1} \to X$ given by

$$(g_1, \ldots, g_{p-1}) \mapsto (g_1, \ldots, g_{p-1}, g_{p-1}^{-1} g_{p-2}^{-1} \cdots g_2^{-1} g_1^{-1}),$$

and hence $|X| = |G|^{p-1} = n^{p-1}$. Since $p|n$, it follows that $p| |X|$ and so $p| |\operatorname{Fix}_X(\mathbb{Z}_p)| > 0$. Therefore, there are at least p elements $g \in G$ with $g^p = e$ and hence at least $p - 1$ nonidentity such elements. □

Combining Theorem 4.3.2 and Theorem 3.5.6 we immediately obtain the following.

Corollary 4.3.4. *A finite group G is a p–group if and only if $|G|$ is a power of p.*

Combining this with the class equation (Corollary 4.1.15) we also have

Corollary 4.3.5. † *If $p \geq 2$ is a prime and G is a p–group, then the center $Z_p(G) \neq \{e\}$.*

Proof. According to Corollary 4.1.15 we have

$$|G| = |Z(G)| + \sum_{[g] \in \mathcal{C}_0} [G : C_G(g)],$$

where recall $\mathcal{C}_0$ is the set of conjugacy classes $[g]$ with *more than one element*, and hence for which $C_G(g) < G$ is a **proper** subgroup. Consequently $[G : C_G(g)] > 1$. By Theorem 3.5.6, $|G| = [G : C_G(g)]|C_G(g)|$, and so by Corollary 4.3.5, p divides $|G|$ and $[G : C_G(g)]$. Therefore, p divides

$$|G| - \sum_{[g] \in \mathcal{C}_0} [G : C_G(g)] = |Z(G)|,$$

as required. □

Cauchy's Theorem tells us that there's an element of order p in any group G in which p divides $|G|$. If we assume p^2 divides $|G|$, there may not be an element of order p^2 in G; for example $\mathbb{Z}_p \times \mathbb{Z}_p$ has no such element although the order of the group is p^2. However, if we ask for a *subgroup* of order p^2, then the First Sylow Theorem ensures that there will be such a subgroup.

Theorem 4.3.6 (First Sylow Theorem)**.** *Suppose G is a finite group of order $|G| = p^k m$, where $p \nmid m$. Then for each $0 \leq i < k$, if $H < G$ is a subgroup of order p^i, then there exists a subgroup $K < G$ with $|K| = p^{i+1}$ such that $H \triangleleft K$. In particular, for each $1 \leq i \leq k$, there exists a subgroup $K < G$ with $|K| = p^i$.*

Proof. The case $i = 0$ is a consequence of Cauchy's Theorem (Theorem 4.3.2). Indeed, $|H| = p^0 = 1$ means $H = \{e\}$, and according to Theorem 4.3.2, there exists $g \in G$ such that $p = |g| = |\langle g \rangle|$, so setting $K = \langle g \rangle$, we have $|K| = p = p^{0+1}$ and trivially $H \triangleleft K$.

Now suppose $H < G$ is any subgroup of order p^i with $1 \leq i < k$. Recall from Example 4.1.10 that G acts on G/H by $g \cdot aH = gaH$. We *restrict* this action to an action of H on G/H. Observe that $h \cdot H = H$ for all $h \in H$, so $H \in \mathrm{Fix}_{G/H}(H)$. On the other hand, by Lemma 4.3.3 we have

$$|\mathrm{Fix}_{G/H}(H)| \equiv |G/H| \bmod p.$$

Since $|G/H| = |G|/|H| = p^k m/p^i = p^{k-i}m$, it follows that p divides $|\mathrm{Fix}_{G/H}(H)|$ and so there exists a coset $aH \neq H$ such that for all $h \in H$ we have $haH = h \cdot aH = aH$. Appealing to Lemma 3.5.3 we have

$$aH \in \mathrm{Fix}_{G/H}(H) \Leftrightarrow haH = aH,\ \forall\ h \in H \Leftrightarrow a^{-1}ha \in H,\ \forall\ h \in H \Leftrightarrow a^{-1}Ha = H \Leftrightarrow a \in N_G(H),$$

where $N_G(H)$ is the normalizer of H in G—the largest subgroup of G in which H is normal (see Exercise 3.3.8). So, $N_G(H)$ is the union of all cosets $aH \in \mathrm{Fix}_{G/H}(H)$. Since each coset of H has the same number of elements (namely $|H|$), we have

$$|N_G(H)| = |\mathrm{Fix}_{G/H}(H)||H|.$$

Since p divides $|\mathrm{Fix}_{G/H}(H)|$ and $|H| = p^i$, it follows that p^{i+1} divides $|N_G(H)|$. Therefore, the quotient group $N_G(H)/H$ and has order divisible by p. Now let $gH \in N_G(H)/H$ be an element of the coset group of order p (which exists by Theorem 4.3.2). Letting $K_0 = \langle gH \rangle < N_G(H)/H$ so that $|K_0| = p$, we set $K = \pi^{-1}(K_0)$ where $\pi\colon N_G(H) \to N_G(H)/H$ is the quotient homomorphism. Then $H < K$ by Theorem 3.6.12, and $K/H \cong K_0$, so $|K|/|H| = p$ and hence $|K| = p^{i+1}$, as required. Since $K < N_G(H)$, we also have $H \triangleleft K$, completing the proof. □

Definition 4.3.7. *For a finite group G and prime $p \geq 2$, a subgroup $P < G$ is called a* **Sylow p–subgroup** *if P is a* maximal *p–subgroup. That is $P < G$ is a p–subgroup and $|G|/|P|$ is not divisible by p ($|P|$ is the largest power of p that divides $|G|$).*

Theorem 4.3.8 (Second Sylow Theorem)**.** *Let G be a finite group, $p \geq 2$ a prime, and $P < G$ a Sylow p–subgroup. Then for any p–subgroup $H < G$, there exists $g \in G$ so that $gHg^{-1} < P$. In particular, any two Sylow p–subgroups are conjugate.*

Proof. Let X be the set of conjugates of P (which are all Sylow p–subgroups), so that G acts transitively, by conjugation, on X. By Corollary 4.1.13 we have $|X| = |G|/|N_G(P)|$ (see Exercise 4.1.3). Since $P < N_G(P)$, $|P|$ divides $|N_G(P)|$ by Theorem 3.5.6, and consequently, $|G|/|N_G(P)|$ divides $|G|/|P|$, which does *not* have p as a factor. Consequently, $|X|$ is not divisible by p.

Now restrict the action on X to H. By Lemma 4.3.3 we have $|\mathrm{Fix}_X(H)| \equiv |X| \neq 0 \bmod p$, hence $\mathrm{Fix}_X(H) \neq \emptyset$. Let $P' = gPg^{-1} \in \mathrm{Fix}_X(H)$, and observe that we therefore have $H < N_G(P')$. Since P' is normal in $N_G(P')$, we see that $HP' < N_G(P')$, and by Theorem 3.6.11, the diamond isomorphism theorem, we have $HP'/P' \cong H/(P' \cap H)$. Consequently, $|HP'| = \frac{|H||P'|}{|P' \cap H|}$ (see also Exercise 3.6.6). So, $|HP'|$ is a power of p, and since P' is a Sylow p–subgroup, $HP' = P'$. But then $H < P'$. So $H < P' = gPg^{-1}$ or $g^{-1}Hg < P$, as required.

For the last statement, note that if P and P' are two Sylow p–subgroups, there exists $g \in G$ so that $gP'g^{-1} < P$. On the other hand, $|gP'g^{-1}| = |P'| = |P|$, so $gP'g^{-1} = P$. □

The final Sylow Theorem provides information on the *number* of Sylow p–subgroups of G, for any prime p. Specifically, for any prime $p \geq 2$, set

$$n_p(G) = |\{P < G \mid P \text{ is a Sylow } p\text{–subgroup }\}|.$$

If p^k divides $|G|$ but p^{k+1} does not divide $|G|$, then we will call p^k is the *maximal power of p dividing G*; this is precisely the order of any Sylow p–subgroup.

Theorem 4.3.9 (Third Sylow Theorem)**.** *Suppose G is a finite group, $p \geq 2$ is a prime, and p^k is the maximal power of p that divides $|G|$. Then $n_p(G)$ divides $\frac{|G|}{p^k}$ and $n_p(G) \equiv 1 \mod p$.*

Proof. Let P be a Sylow p–subgroup and X be the set of all conjugates of P. By Theorem 4.3.8 any two Sylow p–subgroups are conjugate, so $n_p(G) = |X|$. As we saw in the proof of Theorem 4.3.8, G acts transitively on X and $n_p(G) = |X| = |G|/|N_G(P)|$, and since $P < N_G(P)$, $|G|/|N_G(P)|$ divides $|G|/|P| = |G|/p^k$, proving the first claim.

To see that $n_p(G) \equiv 1 \mod p$, we consider the action of P by conjugation on X. Lemma 4.3.3 implies $|\operatorname{Fix}_X(P)| \equiv |X| = n_p(G) \mod p$, and as in the proof of Theorem 4.3.8 if $P' \in X$ is fixed by P, then $P < N_G(P')$ and arguing as in that proof, $P < P'$ so $P = P'$ (since $|P| = |P'|$). Consequently, $\operatorname{Fix}_X(P) = \{P\}$, and thus $n_p(G) \equiv |\operatorname{Fix}_X(P)| = 1 \mod p$. □

Although the following is fairly obvious, it is worth pointing out explicitly as this is one use of Theorem 4.3.9 we will take advantage of repeatedly.

Proposition 4.3.10. † *Suppose G is a finite group, $p \geq 2$ is a prime, and $P < G$ is a Sylow p–subgroup. Then P is normal if and only if $n_p(G) = 1$.*

Proof. Since any conjugate of a Sylow p–subgroup is a Sylow p–subgroup, $n_p(G) = 1$ means that all conjugates of P are equal, hence P is normal. Conversely, if P is normal, then there is only one conjugate of P. By Theorem 4.3.8, any two Sylow p–subgroups are conjugate, hence there is only one Sylow p–subgroup, i.e. $n_p(G) = 1$. □

4.3.1 Application to group structure

From Lagrange's Theorem, we know that for any prime p, a group of order p must be cyclic (see Corollary 3.5.8). In Exericse 4.3.4 you are asked to show that groups of order p^2 are abelian and in fact isomorphic to either $\mathbb{Z}_{p^2}$ or $\mathbb{Z}_p \times \mathbb{Z}_p$. The situation becomes much more complicated for groups whose order has at least two distinct prime factors. For example, $|S_3| = 6$, the product of the two smallest primes, and S_3 is already a fairly complicated (nonabelian) group. However, it turns out that all groups of order pq for distinct primes pq can be completely classified. To do this, we require a few lemmas.

Lemma 4.3.11. *Suppose G is a finite abelian group. If k is the least common multiple of the orders of all elements in G, then there exists an element $c \in G$ with $|c| = k$.*

Proof. It suffices to show that for any two elements $a, b \in G$ with $|a| = n$ and $|b| = m$, there exists an element $c \in G$ with $|c| = \operatorname{lcm}(n, m)$. To do this, first observe that using the prime factorizations of n, m, we can find $n_0 | n$ and $m_0 | m$ so that $\gcd(n_0, m_0) = 1$ and $n_0 m_0 = \operatorname{lcm}(n, m) = \operatorname{lcm}(n_0, m_0)$ (see Exercise 4.3.2). Set $a_0 = a^{n/n_0}$, $b_0 = b^{m/m_0}$, and $c = a_0 b_0$. Observe that $|a_0| = n_0$ and $|b_0| = m_0$. Since $\gcd(n_0, m_0) = 1$, we must have $\langle a_0 \rangle \cap \langle b_0 \rangle = \{e\}$. Indeed, any element in the intersection would have order dividing both n_0 and m_0, and hence must have order 1.

Now observe that since a and b commute, so do a_0 and b_0, and since $\operatorname{lcm}(n, m) = n_0 m_0$, we have

$$c^{\operatorname{lcm}(n,m)} = a_0^{n_0 m_0} b_0^{n_0 m_0} = (a^n)^{m_0} (b^m)^{n_0} = e.$$

Setting $|c| = k$, we have shown $k \leq \operatorname{lcm}(n, m)$. Because $e = c^k = a_0^k b_0^k$, it follows that

$$a_0^k = b_0^{-k} \in \langle a_0 \rangle \cap \langle b_0 \rangle.$$

Because $\langle a_0\rangle \cap \langle b_0\rangle = \{e\}$, we have $a_0^k = e = b_0^{-k}$, and therefore k must be a multiple of both n_0 and m_0, hence a multiple of $\operatorname{lcm}(n_0, m_0) = \operatorname{lcm}(n, m)$. Since $k \leq \operatorname{lcm}(n, m)$, it follows that $|c| = k = \operatorname{lcm}(n, m)$, as required. □

Using this lemma, we also have the following interesting fact (compare Exercise 4.3.3).

Proposition 4.3.12. *For any prime $p \geq 2$, the group of units $\mathbb{Z}_p^\times$ with respect to multiplication is a* **cyclic** *group, and hence $\mathbb{Z}_p^\times \cong \mathbb{Z}_{p-1}$.*

Proof. Since $\mathbb{Z}_p^\times$ is a finite abelian group, we may apply Lemma 4.3.11 to find an element $c \in \mathbb{Z}_p^\times$ so that $k = |c| \leq |\mathbb{Z}_p^\times| = p - 1$ is a multiple of the orders of every element of $\mathbb{Z}_p^\times$. Therefore, every element of $\mathbb{Z}_p^\times$ is a root of the polynomial $x^k - 1$. By Proposition 2.3.17, this polynomial has at most k distinct roots in $\mathbb{Z}_p$, and hence $p - 1 \leq k$. Therefore, $k = p - 1$, and hence $|c| = p - 1 = |\mathbb{Z}_p^\times|$, proving that $\langle c\rangle = \mathbb{Z}_p^\times$, and so $\mathbb{Z}_p^\times$ is cyclic. Proposition 3.3.20 implies $\mathbb{Z}_p^\times \cong \mathbb{Z}_{p-1}$. □

Example 4.3.13. Let $p > q \geq 2$ be distinct prime integers such that $q|(p-1)$. According to Proposition 4.3.12, $\mathbb{Z}_p^\times \cong \mathbb{Z}_{p-1}$ and since q divides $p - 1$, Theorem 3.2.10 guarantees the existence of a unique subgroup $H < \mathbb{Z}_p^\times$ with $H \cong \mathbb{Z}_q$. On the other hand, in Exercise 3.7.7 you showed that $\mathbb{Z}_p^\times \cong \operatorname{Aut}(\mathbb{Z}_p)$. Let $\alpha\colon \mathbb{Z}_q \to \operatorname{Aut}(\mathbb{Z}_p)$ be the composition of the isomorphism $\mathbb{Z}_q \cong H$, the inclusion $H < \mathbb{Z}_p^\times$, and the isomorphism $\mathbb{Z}_p^\times \cong \operatorname{Aut}(\mathbb{Z}_p)$. Define $G_{p,q} = \mathbb{Z}_p \rtimes_\alpha \mathbb{Z}_q$.

The group $G_{p,q}$ is a group of order pq, and since α is injective, $G_{p,q}$ is nonabelian. Observe that α is not unique, but since there is a unique subgroup of $\operatorname{Aut}(\mathbb{Z}_p)$ of order q, any other injective homomorphism $\beta\colon \mathbb{Z}_q \to \operatorname{Aut}(\mathbb{Z}_p)$ must have the same image. Consequently, α and β differ by precomposing with an automorphism $\phi \in \operatorname{Aut}(\mathbb{Z}_q)$. That is, for all $a \in \mathbb{Z}_q$, we have $\alpha_a = \beta_{\phi(a)}$. According to Exercise 4.3.7 (with $\psi\colon \mathbb{Z}_p \to \mathbb{Z}_p$ the identity), it follows that $\mathbb{Z}_p \rtimes_\alpha \mathbb{Z}_q \cong \mathbb{Z}_p \rtimes_\beta \mathbb{Z}_q$. Therefore, any two choices of injective homomorphism $\mathbb{Z}_q \to \operatorname{Aut}(\mathbb{Z}_p)$ give isomorphic groups, and hence $G_{p,q}$ is uniquely determined, up to isomorphism.

Theorem 4.3.14 (Groups of order pq)**.** *Suppose $p > q \geq 2$ are distinct primes and G is a group of order pq. If $q \nmid (p-1)$, then $G \cong \mathbb{Z}_{pq}$. If $q|(p-1)$, then either $G \cong \mathbb{Z}_{pq}$ or else $G \cong G_{p,q}$, the group constructed in Example 4.3.13.*

Proof. Suppose G is a group with $|G| = pq$. By Theorem 4.3.6 (or Theorem 4.3.2) there exist subgroups $P, Q < G$ of order p and q, respectively. Note that since P and Q have relatively prime orders, we must have $P \cap Q = \{e\}$. We first prove that P is necessarily a normal subgroup of G. By Theorem 4.3.9, $n_p(G) \equiv 1 \bmod p$ and $n_p(G)|q$. Because $p > q$, the only positive divisor of q which is also congruent to 1 mod p is 1, and hence $n_p(G) = 1$. By Proposition 4.3.10, $P \triangleleft G$.

Proposition 3.7.7 guarantees that $PQ < G$ and $PQ \cong P \rtimes_c Q$ where $c\colon Q \to \operatorname{Aut}(P)$ is given by conjugation in G. Moreover, $|PQ| = pq = |G|$, and so $G = PQ \cong P \rtimes_c Q$. We describe the possible semi-direct products, up to isomorphism, depending on p and q.

First, we claim that either c is injective or c is trivial. To see this, observe that $c(Q) < \operatorname{Aut}(P)$, and hence $|c(Q)|$ must divide $p - 1 = |\operatorname{Aut}(P)|$ (we are using the fact that $P \cong \mathbb{Z}_p$). On the other hand, by the first isomorphism theorem (Theorem 3.6.5), $c(Q) \cong Q/\ker(c)$. Therefore, $|c(Q)| = [Q : \ker(c)]$, which divides $|Q| = q$. But q is prime, so either $|c(Q)| = q$ (and c is injective) or $|c(Q)| = 1$ (and c is trivial). This proves the claim.

Now if $q \nmid (p-1)$, then c is trivial. A trivial semi-direct product is a direct product (why?). Consequently, $G = PQ \cong P \times Q \cong \mathbb{Z}_p \times \mathbb{Z}_q$. Since $\gcd(p, q) = 1$, Theorem 1.5.8 implies $G \cong \mathbb{Z}_p \times \mathbb{Z}_q \cong \mathbb{Z}_{pq}$, as required.

Finally, suppose $q|(p-1)$. If c is trivial, we still have $G = PQ \cong \mathbb{Z}_{pq}$, as above. On the other hand, if c is nontrivial (hence injective), then $G \cong P \rtimes_c Q$ is a nonabelian group. From Proposition 3.3.20 we have isomorphisms $\phi\colon \mathbb{Z}_q \to Q$, and $\psi\colon \mathbb{Z}_p \to P$. Set $\alpha\colon \mathbb{Z}_q \to \operatorname{Aut}(\mathbb{Z}_p)$ to be $\alpha_a(k) = \psi^{-1}c_{\phi(a)}\psi(k)$. Observe that α_a is indeed an automorphism for all $a \in \mathbb{Z}_q$ (as the composition of isomorphisms). In addition, α defines a homomorphism:

$$\alpha_{aa'} = \psi^{-1}c_{\phi(aa')}\psi = \psi^{-1}c_{\phi(a)\phi(a')}\psi = \psi^{-1}c_{\phi(a)}c_{\phi(a')}\psi = \psi^{-1}c_{\phi(a)}\psi\psi^{-1}c_{\phi(a')}\psi = \alpha_a\alpha_{a'}.$$

Now note that by definition, $\psi(\alpha_a(k)) = c_{\phi(a)}(\psi(k))$ for all $a \in \mathbb{Z}_q$ and $k \in \mathbb{Z}_p$. By Exercise 4.3.7 we have an isomorphism $P \rtimes_c Q \cong \mathbb{Z}_p \rtimes_\alpha \mathbb{Z}_q$. Since α is injective, $\mathbb{Z}_p \rtimes_\alpha \mathbb{Z}_q \cong G_{p,q}$, completing the proof. □

Example 4.3.15. There are simpler applications of the Sylow Theorems than that of Theorem 4.3.14 which provide similar classification results. For example, we claim that any group G of order $14,161 = 7^2 17^2$ is necessarily abelian. To see this, let P be a Sylow 17–subgroup and Q a Sylow 7–subgroup. We have $|P| = 17^2 = 289$ and $|Q| = 7^2 = 49$. The number of Sylow 17 and Sylow 7 subgroups satisfy the following

$$\begin{array}{ll} n_{17}(G)|49 & n_{17}(G) \equiv 1 \bmod 17 \\ n_7(G)|289 & n_7(G) \equiv 1 \bmod 7 \end{array}$$

So, $n_{17}(G) = 1, 7$ or 49. On the other hand, $49 \equiv 15 \bmod 17$, so $n_{17}(G) = 1$ is the only possibility. Similarly, $n_7(G) = 1, 17$ or 289. Since $17 \equiv 3 \bmod 7$ and $289 \equiv 2 \bmod 7$, the only possibility is that $n_7(G) = 1$. By Proposition 4.3.10, P and Q are both normal in G and $PQ < G$. Any element in the intersection $g \in P \cap Q$ must have order dividing both 7 and 17. Such a $g \in G$ must be therefore be the identity, and so $P \cap Q = \{e\}$. By Proposition 3.7.1, we have $PQ \cong P \times Q$. As in the previous proof we see $|PQ| = 7^2 17^2 = |G|$, so $PQ = G$, and hence $G \cong P \times Q$.

Since every element of P commutes with every element of Q, all that remains is to show that both P and Q are themselves abelian. This follows from Exercise 4.3.4, and hence G is abelian. In fact, from that exercise, we see that G is isomorphic to one of the four groups

$$\mathbb{Z}_{289} \times \mathbb{Z}_{49},\ \mathbb{Z}_{17}^2 \times \mathbb{Z}_{49},\ \mathbb{Z}_{289} \times \mathbb{Z}_7^2,\ \text{or } \mathbb{Z}_{17}^2 \times \mathbb{Z}_7^2.$$

Exercises.

Exercise 4.3.1. Verify the claims in the proof of Theorem 4.3.2. Specifically, prove that **(i)** $a \cdot (g_1, \ldots, g_p) = (g_{a+1}, \ldots, g_{a+p})$ defines an action of $\mathbb{Z}_p$ on G^p and **(ii)** $X = \{(g_1, \ldots, g_p) \in G^p \mid g_1 g_2 \cdots g_p = e\}$ is invariant by $\mathbb{Z}_p$ and hence the $\mathbb{Z}_p$ action on G^p restricts to an action on X. *Hint: for the last part, you need to show that for all $a \in \mathbb{Z}_p$, $g_1 \cdots g_p = e$ if and only if $g_{a+1} \cdots g_{a+p} = e$.*

Exercise 4.3.2. Suppose $m, n \geq 1$ are two integers. Find $n_0|n$ and $m_0|m$ so that $\gcd(n_0, m_0) = 1$ and $n_0 m_0 = \operatorname{lcm}(n_0, m_0) = \operatorname{lcm}(n, m)$. *Hint: Writing $n = p_1^{k_1} \cdots p_r^{k_r}$ and $m = p_1^{j_1} \cdots p_r^{j_r}$ with $k_i, j_i \geq 0$ for all i and $p_1, \ldots, p_r$ distinct primes, prove that*

$$\operatorname{lcm}(n, m) = p_1^{\max\{k_1, j_1\}} \cdots p_r^{\max\{k_r, j_r\}}.$$

Define n_0 by dividing n by $p_i^{k_i}$ for each i such that $k_i \leq \max\{k_i, j_i\}$, then find the appropriate m_0 (careful: the construction is not entirely symmetric since we may have $k_i = j_i$ for some i).

Exercise 4.3.3. Find some $n \geq 2$ so that $\mathbb{Z}_n^\times$ is not cyclic (Proposition 4.3.12 implies that n is necessarily **not** prime).

Exercise 4.3.4. Let $p \geq 2$ be a prime and prove that any group G of order p^2 is isomorphic to either $\mathbb{Z}_{p^2}$ or $\mathbb{Z}_p \times \mathbb{Z}_p$. *Hint: If there exists an element of order p^2, then $G \cong \mathbb{Z}_{p^2}$, so assume that this is not the case and deduce that every nonidentity element of G has order p. Then prove that for any two elements $g, h \in G$, either $\langle g \rangle = \langle h \rangle$ or $\langle g \rangle \cap \langle h \rangle = \{e\}$. Finally, prove that there are two nonidentity elements $g, h \in G$ so that $\langle g \rangle \cap \langle h \rangle = \{e\}$, and apply Proposition 3.7.1.*

Exercise 4.3.5. Suppose $p \geq 2$ is a prime and G is a group with $|G| = p^3$. Prove that either $|Z(G)| = p$ or else G is abelian.

Exercise 4.3.6. Prove that any finite abelian group is isomorphic to a direct sum of its Sylow subgroups.

Exercise 4.3.7. Suppose A, B, H, K are groups and that $\alpha\colon A \to \mathrm{Aut}(H)$ and $\beta\colon B \to \mathrm{Aut}(K)$ are homomorphisms. Prove that if $\phi\colon A \to B$ and $\psi\colon H \to K$ are isomorphisms such that for all $a \in A$ and $h \in H$

$$\psi(\alpha_a(h)) = \beta_{\phi(a)}(\psi(h)),$$

then the formula $\eta(h, a) = (\psi(h), \phi(a))$ defines an isomorphism $\eta\colon H \rtimes_\alpha A \to K \rtimes_\beta B$.

Exercise 4.3.8. Let $p \geq 3$ be an odd prime. Prove that any nonabelian group of order $2p$ is isomorphic to the dihedral group D_p.

Exercise 4.3.9. Prove that any group of order $61,009$ is abelian.

Exercise 4.3.10. Prove that a group of order $4,199$ is cyclic.

Exercise 4.3.11. Prove that for all primes $p \geq 5$, a Sylow p–subgroup $P < S_p$ is **NOT** normal. Here S_p is the symmetric group on p elements.

The point of the previous exercise is to illustrate that the largest prime dividing the order of a group need not have associated Sylow subgroup being normal (in contrast to the case of the proof of Theorem 4.3.14).

Exercise 4.3.12. Classify (up to isomorphism) all groups of order 20. *Hints: There are* 5 *isomorphism classes. There are more possibilities for semidirect products than in Theorem 4.3.14. Also note that the Sylow* 2*–subgroups can be isomorphic to either* $\mathbb{Z}_4$ *or* $\mathbb{Z}_2 \times \mathbb{Z}_2$*. You may want to prove that for any surjective homomorphism* $\delta\colon \mathbb{Z}_2 \times \mathbb{Z}_2 \to \mathbb{Z}_2$*, that there exists an automorphism* $\phi\colon \mathbb{Z}_2 \times \mathbb{Z}_2 \to \mathbb{Z}_2 \times \mathbb{Z}_2$ *so that* $\delta = \pi \circ \phi$*, where* $\pi(a, b) = a$*, and that there exists a unique surjective homomorphism* $\mathbb{Z}_4 \to \mathbb{Z}_2$*. Exercise 4.3.7 will also be helpful.*

You should...

- Be able to do all the exercises from this section and know the proofs of statements marked by †.
- Know what Cauchy's Theorem says.
- Know what a Sylow subgroup of a finite group is and what the Sylow Theorems say (Theorems 4.3.6,4.3.8, 4.3.9).
- Be able to apply the Sylow theorems (e.g. as in Proposition 4.3.10 and essentially all of Subsection 4.3.1). Note that the Sylow Theorems are very often used in conjunction with Propositions 3.7.1 and 3.7.7.

Chapter 5

Rings, fields, and groups

In this final chapter, we bring rings and fields back into the discussion. This starts with some general discussion of these objects and some relationships between the two. After that, we apply this discussion to the problem of finding roots of polynomials. We end with an extremely brief introduction to a beautiful connection between rings, fields, and groups via Galois Theory.

5.1 Fields from rings

In this section we will describe two constructions of fields from commutative rings.

5.1.1 Field of fractions

The product of two nonzero integers is a nonzero integer. Likewise, the product of two nonzero polynomials is a nonzero polynomial. This is not true for an arbitrary ring. If R is a ring, a **zero divisor of** R is a nonzero element $a \in R$ such that for some nonzero element $b \in R$ we have $ab = 0$.

Example 5.1.1. Suppose $n > 1$ is non-prime integer. The ring $\mathbb{Z}_n$ has zero divisors: let $k, m > 1$ be such that $n = km$. Then $[k], [m] \neq [0]$ and $[k][m] = [km] = [n] = [0]$, so $[k], [m]$ are zero divisors.

A commutative ring with 1 with no zero divisors is called an **integral domain**. So, $\mathbb{Z}$ and $\mathbb{F}[x]$ are integral domains. Another general source of integral domains is the following.

Proposition 5.1.2. *If $\mathbb{F}$ is a field and $R \subset \mathbb{F}$ is a subring with 1, then R is an integral domain.*

Proof. Since $\mathbb{F}$ is a commutative ring, it follows that R is a commutative ring with 1. We need only verify that R has no zero divisors. So, suppose $a, b \in R$ with $a, b \neq 0$, but $ab = 1$. Then since $\mathbb{F}$ is a field, $a^{-1} \in \mathbb{F}$, and we have

$$0 = a^{-1}0 = a^{-1}ab = b$$

a contradiction. □

It turns out that any integral domain is a subring of a field. To prove this, we will explicitly construct a field from an integral domain R. First, let

$$\widetilde{Q} = \{(a, b) \in R^2 \mid b \neq 0\},$$

and define a relation $\sim$ by

$$(a, b) \sim (a', b') \Leftrightarrow ab' = a'b.$$

Lemma 5.1.3. *Suppose R is an integral domain and $\widetilde{Q}$, $\sim$ as above. Then $\sim$ is an equivalence relation.*

Proof. Exercise 5.1.2 □

We denote the set of equivalence classes of $\widetilde{Q}$ by $Q(R)$, and denote the equivalence class of $(a,b) \in \widetilde{Q}$ by a/b. Thus, $a/b = a'/b'$ if and only if $ab' = a'b$.

Theorem 5.1.4. *If R is an integral domain, then the following defines two operations on $Q(R)$:*

$$a/b + c/d = (ad + bc)/bd,$$

$$(a/b)(c/d) = (ac)/(bd).$$

Furthermore, with these operations, $Q(R)$ is a field, and the map $R \to Q(R)$ defined by $r \mapsto r/1$ is an injective ring homomorphism.

The last sentence means that we can view R as a subring of $Q(R)$, and so the theorem provides a converse to Proposition 5.1.2.

Proof. Note that $bd \neq 0$ so both definitions describe elements in $Q(R)$. To see that it is well-defined, let $a/b = a'/b'$ and $c/d = c'/d'$. Then we have $ab' = a'b$ and $cd' = c'd$, and hence

$$acb'd' = a'c'bd \Rightarrow (ac)/(bd) = (a'c')/(b'd'),$$

and

$$(bd)(a'd' + b'c') = a'bdd' + c'dbb' = ab'dd' + cd'bb' = (b'd')(ad + cb) \Rightarrow (a'd' + b'c')/(b'd') = (ad + bc)/(bd).$$

Therefore, the two operations are well-defined.

Associativity of multiplication is immediate from the definition. For addition (freely using associativity of addition and multiplication in R), we have

$$\begin{aligned}(a/b + c/d) + e/f &= (ad + bc)/bd + e/f = ((ad + bc)f + (bd)e)/((bd)f) \\ &= (a(df) + (b(cf) + b(de)))/(b(df)) = a/b + (cf + de)/df = a/b + (c/d + e/f).\end{aligned}$$

So both operations are associative. Likewise, commutivity of multiplication is clear, and for addition:

$$a/b + c/d = (ad + bc)/bd = (cb + da)/db = c/d + a/b.$$

The element $0/1$ $(= 0/b$ for any $b \neq 0)$ serves as 0, since

$$a/b + 0/1 = (a1 + 0b)/(1b) = a/b,$$

and every element $a/b \in Q(R)$ has an inverse:

$$(-a)/b + a/b = (-ab + ab)/b^2 = ((-a + a)b)/b^2 = (0b)/b^2 = 0/b^2 = 0/1$$

since $0(b^2) = 0(1)$. Thus addition makes $Q(R)$ into an abelian group.

The multiplicative identity is $1/1$ since

$$(a/b)(1/1) = (a1)/(b1) = a/b,$$

for all $a/b \in Q(R)$, and the multiplicative inverse of $a/b \in Q(R) \setminus \{0\}$ is b/a as $(b/a)(a/b) = (ab)/(ab) = 1/1$. To show that $Q(R)$ is a field, all that remains to prove that multiplication distributes over addition, and we leave this as Exercise 5.1.3.

To see that $r \mapsto r/1$ defines a homomorphism, note that for all $r, s \in R$, we have

$$r/1 + s/1 = (r1 + s1)/(1 \cdot 1) = (r + s)/1 \text{ and } (r/1)(s/1) = (rs)/(1 \cdot 1) = (rs)/1.$$

To prove that this homomorphism is injective, we note that if r is in the kernel, then $r/1 = 0/1$, which means $r = r1 = 0 \cdot 1 = 0$. So the kernel is $\{0\}$, and the homomorphism is injective. □

Example 5.1.5. Since $\mathbb{F}[x]$ is an integral domain which we are quite familiar with, we might wonder if $Q(\mathbb{F}[x])$ is also a familiar field. Indeed, when $\mathbb{F} = \mathbb{R}$, the elements of $Q(\mathbb{R}[x])$ are rational functions, as one studies in calculus, and earlier. We typically write $\mathbb{F}(x) = Q(\mathbb{F}[x])$ for this *field of rational functions.*

5.1.2 Quotient fields

There is another important method for constructing a field from a commutative ring R, though this construction requires some additional data. Recall from Exercise 3.3.13 that an ideal $\mathcal{I} \subset R$ is a subring such that for all $a \in \mathcal{I}$ and $r \in R$, we have $ar \in \mathcal{I}$. From Exercise 3.6.11 we know that the quotient additive group of cosets

$$\pi \colon R \to R/\mathcal{I} = \{r + \mathcal{I} \mid r \in R\}$$

admits a ring structure with multiplication defined by $(r + \mathcal{I})(s + \mathcal{I}) = sr + \mathcal{I}$, and that π is a quotient ring homomorphism.

For the ring $\mathbb{Z}$, every subgroup is in fact an ideal. Indeed, from Theorem 3.2.9 we know that every subgroup of $\mathbb{Z}$ is given by $\langle d \rangle$, for some $d \geq 0$. This consists of all integral multiples of d, and so given $md \in \langle d \rangle$ and $r \in \mathbb{Z}$, we have $r(md) = (rm)d \in \langle d \rangle$. The quotient ring is nothing but $\mathbb{Z}_d$. This illustrates that the quotient ring need not be an integral domain. On the other hand, when d is a prime, we know that it is actually a field.

One special property of a subgroup generated by a prime p is that it is *maximal*, meaning that any strictly larger subgroup must be all of $\mathbb{Z}$. We similarly say that an ideal $\mathcal{I} \subset R$ is *maximal*, if $\mathcal{I} \neq R$ and any ideal that strictly contains $\mathcal{I}$ must be all of R. Said differently, an ideal $\mathcal{I} \neq R$ is maximal if for any other ideal $\mathcal{J}$ with $\mathcal{I} \subset \mathcal{J} \subset R$, either $\mathcal{I} = \mathcal{J}$ or $\mathcal{I} = R$.

Theorem 5.1.6. *If R is a commutative ring with 1, then the ideal $\mathcal{I} \subset R$ is maximal if and only if $R/\mathcal{I}$ is a field.*

The proof appeals to the following two lemmas.

Lemma 5.1.7. *If $\mathbb{F}$ is a field, then it has no nontrivial ideals. That is, the only trivial ideals in $\mathbb{F}$ are $\{0\}$ and $\mathbb{F}$.*

Proof. Exercise 5.1.4. □

Suppose R is a commutative ring with 1 and $a \in R$. The *principal ideal generated by a*, denoted $(\!(a)\!)$, is the smallest ideal containing a (to see that this is well defined, see Exercise 5.1.5). The following lemma provides another useful description of $(\!(a)\!)$.

Lemma 5.1.8. † *Suppose R is a commutative ring with 1 and $a \in R$. Then $(\!(a)\!) = \{ra \mid r \in R\}$.*

Proof. Let $\mathcal{I} = \{ra \mid r \in R\}$. Since $1 \in R$, $a = 1a \in \mathcal{I}$. Since any ideal containing a must contain $\mathcal{I}$, it suffices to prove that $\mathcal{I}$ is an ideal. For this, let $ra, r'a \in \mathcal{I}$, we have

$$ra - r'a = (r - r')a \in \mathcal{I}$$

and if $s \in R$, then

$$s(ra) = (sr)a \in \mathcal{I}.$$

So, $\mathcal{I}$ is an ideal, completing the proof. □

Proof of Theorem 5.1.6. First, suppose $R/\mathcal{I}$ is a field. According to Exercise 3.6.14 (the correspondence theorem for surjective ring homomorphisms) the image by π of any ideal $\mathcal{J}$ with $\mathcal{I} \subset \mathcal{J} \subset R$ must be an ideal of $R/\mathcal{I}$. By Lemma 5.1.7, $\pi(\mathcal{J})$ is either the zero ideal or all of $R/\mathcal{I}$. Consequently, $\mathcal{J}$ is either $\mathcal{I}$ or R, and hence $\mathcal{I}$ is maximal.

Conversely, suppose that $\mathcal{I} \subset R$ is maximal. By Exercise 3.6.14 again it follows that $R/\mathcal{I}$ has no nontrivial ideals (i.e. every ideal is either zero or the entire quotient ring $R/\mathcal{I}$). Since R is commutative, so is $R/\mathcal{I}$. Since R has 1, $1+\mathcal{I}$ is 1 for $R/\mathcal{I}$. Let $a+\mathcal{I}$ be a nonzero element, which means that $a \notin \mathcal{I}$. The ideal generated $a + \mathcal{I}$ is an ideal that contains $a + \mathcal{I}$, and since there are no nontrivial ideals, we must have $(\!(a + \mathcal{I})\!) = R/\mathcal{I}$. By Lemma 5.1.8

$$(\!(a + \mathcal{I})\!) = \{(r + \mathcal{I})(a + \mathcal{I}) = ra + \mathcal{I} \mid r \in R\}.$$

Since $1 + \mathcal{I} \in (\!(a + \mathcal{I})\!)$, there exists an element $r \in R$ so that $(r + \mathcal{I})(a + \mathcal{I}) = ra + \mathcal{I} = 1 + \mathcal{I}$, and hence $a + \mathcal{I}$ is invertible, and hence $R/\mathcal{I}$ is a field. □

We record here another important application of Lemma 5.1.7.

Proposition 5.1.9. † *If $\mathbb{F}$ is a field and R any ring, then a homomorphism $\phi\colon \mathbb{F} \to R$ is either constant (equal to 0), or else ϕ is injective.*

Proof. The kernel $\ker(\phi)$ is an ideal of $\mathbb{F}$, so either $\ker(\phi) = \{0\}$ and ϕ is injective, or else $\ker(\phi) = \mathbb{F}$, and ϕ is identically zero. □

Corollary 5.1.10. *If $\mathbb{F}$ and $\mathbb{K}$ are fields and $\phi\colon \mathbb{F} \to \mathbb{K}$ is a nonconstant homomorphism, then ϕ restricts to an injective group homomorphism from $(\mathbb{F}^\times, \cdot)$ to $(\mathbb{K}^\times, \cdot)$. In particular, $\phi(1) = 1$, and so ϕ is unital.*

Proof. By Proposition 5.1.9, ϕ is injective, and hence $\phi(\mathbb{F}^\times) \subset \mathbb{K}^\times$. So, ϕ restricts to an injective map from $\mathbb{F}^\times$ to $\mathbb{K}^\times$. By definition of a field homomorphism, we have $\phi(ab) = \phi(a)\phi(b)$ for all a, b. In particular, ϕ must send the multiplicative identity to the multiplicative identity. □

Exercises.

Exercise 5.1.1. Prove that a finite integral domain is a field.

Exercise 5.1.2. Prove Lemma 5.1.3.

Exercise 5.1.3. Complete the proof of 5.1.4 by proving the distributive property.

Exercise 5.1.4. Prove Lemma 5.1.7. More generally, prove that if R is a commutative ring with 1 and $\mathcal{I} \subset R$ is an ideal containing an invertible element, then $\mathcal{I} = R$.

Exercise 5.1.5. Prove that if R is a commutative ring and $\{\mathcal{I}_\alpha\}_{\alpha \in J}$ is any collection of ideals in R, then

$$\bigcap_{\alpha \in J} \mathcal{I}_\alpha$$

is an ideal in R. In particular, the intersection of all ideals containing an element $a \in R$ is an ideal, and hence $(\!(a)\!)$ is well-defined.

You should...

- Be able to do all the exercises from this section and know the proofs of statements marked by †.
- Know the definition of an integral domain, maximal ideals, and principal ideals.
- Understand the construction of the field of fractions from an integral domain.
- Know how to construct the quotient field of a commutative ring with 1, given a maximal ideal.

5.2 Field extensions, polynomials, and roots

Throughout this section, suppose that $\mathbb{F}$ is a field. Recall that a field $\mathbb{K}$ containing $\mathbb{F}$ as a subfield is called an *extension* of $\mathbb{F}$. Recall that $\mathbb{K}$ can be viewed as a vector space over $\mathbb{F}$, and we will write $[\mathbb{K} : \mathbb{F}]$ to denote the dimension (over $\mathbb{F}$) of $\mathbb{K}$; see Section 2.2 and Example 2.4.6. When $[\mathbb{K} : \mathbb{F}] < \infty$, we say that $\mathbb{K}$ is a **finite extension** of $\mathbb{F}$. Further recall that $\mathbb{F}[x]$ denotes the ring of polynomials with coefficients in $\mathbb{F}$. A polynomial $f \in \mathbb{F}[x]$ is *irreducible* if whenever we write $f = uv$ with $u, v \in \mathbb{F}[x]$, then either u or v is constant. An element $\alpha \in \mathbb{F}$ is a *root* of $f \in \mathbb{F}[x]$ if $f(\alpha) = 0$; see Section 2.3.

If $\mathbb{F} \subset \mathbb{K}$ is an extension, then $\mathbb{F}[x] \subset \mathbb{K}[x]$, extending the inclusion $\mathbb{F} \subset \mathbb{K}$. It can happen that a polynomial $f \in \mathbb{F}[x]$ has a root in $\mathbb{K}$, but not in $\mathbb{F}$. For example, $x^2 + 1 \in \mathbb{R}[x]$ has a root in $\mathbb{C}$, but not in $\mathbb{R}$. It turns out, if $f \in \mathbb{F}[x]$, one can *always* find a field extension $\mathbb{F} \subset \mathbb{K}$, so that f has a root in $\mathbb{K}$. This basically follows from the discussion in the previous section and the next fact.

Proposition 5.2.1. *If $p \in \mathbb{F}[x]$ is a irreducible polynomial, then $\mathbb{F}[x]/(\!(p)\!)$ is a field. Consequently, $(\!(p)\!)$ is a maximal ideal.*

Proof. We must show that every nonzero element $f + (\!(p)\!)$ has a multiplicative inverse. Since $f + (\!(p)\!)$ is not the identity, it follows that p does not divide f. Since p is irreducible, this implies $\gcd(f, p) = 1$, and so by Proposition 2.3.10 there exists $u, v \in \mathbb{F}[x]$ so that $uf + vp = 1$. Now observe that

$$(f + (\!(p)\!))(u + (\!(p)\!)) = uf + (\!(p)\!) = (1 - vp) + (\!(p)\!) = 1 + (\!(p)\!),$$

where the last equation holds because $vp \in (\!(p)\!)$. Therefore, $\mathbb{F}[x]/(\!(p)\!)$ is a field. □

The homomorphism $\pi\colon \mathbb{F}[x] \to \mathbb{F}[x]/(\!(p)\!)$ from Proposition 5.2.1 restricts to a homomorphism on the field $\mathbb{F} \subset \mathbb{F}[x]$, consisting of the constant polynomials. The restriction $\pi|_{\mathbb{F}}$ cannot be constant zero (for then π would be constant zero), so it follows from Proposition 5.1.9 that $\pi|_{\mathbb{F}}$ must be injective. We use this to identify $\mathbb{F}$ with its image in $\mathbb{F}[x]/(\!(p)\!)$; that is, we think of the $\pi|_{\mathbb{F}}$ as an inclusion of $\mathbb{F}$, writing $a = \pi(a)$, for all $a \in \mathbb{F}$. Writing $\mathbb{K} = \mathbb{F}[x]/(\!(p)\!)$, this field $\mathbb{K}$ is an extension field of $\mathbb{F}$. We may consequently write $\mathbb{F}[x] \subset \mathbb{K}[x]$, viewing any polynomial with coefficients in $\mathbb{F}$ as a polynomial with coefficients in $\mathbb{K}$.

To simplify the notation, instead of $x + (\!(p)\!)$ for the image $\pi(x)$ of $x \in \mathbb{F}[x]$, we write $\alpha = \pi(x) \in \mathbb{K}$.

Theorem 5.2.2. *Suppose $\mathbb{F}$ is field, $p \in \mathbb{F}[x]$ is an irreducible polynomial, $\mathbb{K} = \mathbb{F}[x]/(\!(p)\!)$, and $\alpha \in \mathbb{K}$ are as above. Then α is a root of $p \in \mathbb{K}[x]$.*

Moreover, $[\mathbb{K} : \mathbb{F}] = \deg(p)$ and every element of $\mathbb{K}$ can be uniquely expressed in the form

$$\sum_{i=0}^{n-1} c_i \alpha^i, \tag{5.1}$$

where $n = \deg(p)$ and $c_0, \ldots, c_{n-1} \in \mathbb{F}$.

Proof. The first part of the theorem is just a matter of unravelling the notation.

We write

$$p(x) = \sum_{i=0}^{n} a_i x^i \in \mathbb{F}[x].$$

But when we view $p \in \mathbb{K}[x]$, because we are identifying $\mathbb{F}$ as a subfield of $\mathbb{K}$ via π, we have

$$p(x) = \sum_{i=0}^{n} \pi(a_i) x^i \in \mathbb{K}[x].$$

So, what is $p(\alpha) \in \mathbb{K}$? Since $\alpha = \pi(x)$, and π is a homomorphism, we have

$$p(\alpha) = \sum_{i=0}^{n} \pi(a_i)\alpha^i = \sum_{i=0}^{n} \pi(a_i)\pi(x)^i = \pi(p(x)) = 0$$

since $p \in \mathbb{F}[x]$ is in the kernel of π.

To prove the second part, we claim that every element of $\mathbb{K} = \mathbb{F}[x]/(\!(p)\!)$ can be uniquely represented as $f + (\!(p)\!)$ where $\deg(f) < \deg(p)$. To see this, note that by Proposition 2.3.9 there exists $q, r \in \mathbb{F}[x]$ so that $f = pq + r$ where $\deg(r) < \deg(p)$. Consequently,

$$f + (\!(p)\!) = r + pq + (\!(p)\!) = r + (\!(p)\!).$$

Therefore, every element of $\mathbb{F}[x]/(\!(p)\!)$ can be expressed in the form $f + (\!(p)\!)$ where $\deg(f) < \deg(p)$. For the uniqueness, suppose $\deg(f), \deg(h) < \deg(p)$ and $f + (\!(p)\!) = h + (\!(p)\!)$. Then $f - h \in (\!(p)\!)$, which means that there exists $u \in \mathbb{F}[x]$ so that $f - h = up$. In this case we have $u = 0$, for otherwise, since $\deg(f - h) < \deg(p)$, we would have

$$\deg(f - h) < \deg(p) \leq \deg(u) + \deg(p) = \deg(up) = \deg(f - h),$$

a contradiction. So, $u = 0$ and hence $f = h$.

Given an arbitrary element $f + (\!(p)\!)$ with $\deg(f) < n = \deg(p)$ and writing $f = \sum_{i=0}^{n-1} c_i x^i$, we see that in terms of α, this element becomes

$$f + (\!(p)\!) = f(\alpha) = \sum_{i=0}^{n-1} c_i \alpha^i,$$

proving that the elements of $\mathbb{K}$ have the form claimed in (5.1). Thus $1, \alpha, \alpha^2, \ldots, \alpha^{n-1}$ is a basis for $\mathbb{K}$ over $\mathbb{F}$, and consequently $[\mathbb{K} : \mathbb{F}] = n = \deg(p)$. □

Example 5.2.3. Consider the construction above in the case $p = x^2+1 \in \mathbb{R}[x]$. Then the field $\mathbb{K}$ constructed in Theorem 5.2.2 consists of elements

$$\mathbb{K} = \{a + b\alpha \mid a, b \in \mathbb{R}\}.$$

What is the product $(a+b\alpha)(c+d\alpha)$? We should multiply the polynomials, then apply the division algorithm (dividing by $x^2 + 1$), then "evaluate" the remainder at α. First we multiply:

$$(a + bx)(c + dx) = ac + (ad + bc)x + bdx^2.$$

Applying the division algorithm by $x^2 + 1$, we obtain

$$ac + (ad + bc)x + bdx^2 = bd(x^2 + 1) - bd + ac + (ad + bc)x = bd(x^2 + 1) + (ac - bd) + (ad + bc)x.$$

Therefore, evaluating the remainder $(ac - bd) + (ad + bc)x$ at $x = \alpha$ we get

$$(a + b\alpha)(c + d\alpha) = (ac - bd) + (ad + bc)\alpha.$$

Note that if instead of calling the element α, we call it i, then all we have

$$\mathbb{K} = \{a + bi \mid a, b \in \mathbb{R}\}$$

and the product is given by $(a + bi)(c + di) = ac - bd + (ad + bc)i$, which is to say, the construction of $\mathbb{K}$ exactly produces $\mathbb{C}$. The next subsection provides a more general context for this fact.

In general, if $\mathbb{K}$ is obtained from $p \in \mathbb{F}[x]$ as in Theorem 5.2.2, then an alternative to applying the Euclidean algorithm when computing products is to do repeated substitutions using the fact that $p(\alpha) = 0$. Specifically, if $p = a_0 + a_1 x + \cdots + a_{n-1}x^{n-1} + x^n$, then in $\mathbb{K}$ we have

$$\alpha^n = -(a_0 + a_1\alpha + \cdots + a_{n-1}\alpha^{n-1}),$$

and consequently, we can formally multiply then repeatedly substitute using this equation until all powers of α are less than n.

Example 5.2.4. In the previous example, the above equation becomes $\alpha^2 = -1$, and thus the product can be computed by multiplying and substituting:

$$(a + b\alpha)(c + d\alpha) = ac + (ad + bc)\alpha + bd\alpha^2 = ac + (ad + bc)\alpha + bd(-1) = (ac - bd) + (ad + bc)\alpha.$$

Example 5.2.5. Consider the polynomial $x^3 + 3 \in \mathbb{Q}[x]$. This has one real root, $-\sqrt[3]{3}$ and one can easily check that this is not in $\mathbb{Q}$ (compare with Example 2.2.4). On the other hand, the other two roots in $\mathbb{C}$ are nonreal (given by $\sqrt[3]{3}e^{\pi i/3}$ and $\sqrt[3]{3}e^{-\pi i/3}$), and in particular are not in $\mathbb{Q}$. By Exercise 2.3.5, $x^3 + 3$ is irreducible in $\mathbb{Q}[x]$. Letting $\mathbb{K}$ denote the field obtained from Theorem 5.2.2 with $\alpha \in \mathbb{K}$ and $\alpha^3 + 3 = 0$, we have

$$\mathbb{K} = \{a + b\alpha + c\alpha^2 \mid a, b, c \in \mathbb{Q}\}.$$

The product is given by multiplying formally, and substituting $\alpha^3 = -3$:

$$\begin{aligned}(a + b\alpha + c\alpha^2)(u + v\alpha + w\alpha^2) &= au + av\alpha + aw\alpha^2 + bu\alpha + bv\alpha^2 + bw\alpha^3 + cu\alpha^2 + cv\alpha^3 + cw\alpha^4 \\ &= (au - 3bw - 3cv) + (av + bu - 3cw)\alpha + (aw + bv + cu)\alpha^2.\end{aligned}$$

Imagine trying to prove directly that this defines a multiplication making $\mathbb{K}$ into a field!

5.2.1 Extensions and subfields

Suppose $\mathbb{F}$ is a field and $p \in \mathbb{F}[x]$ is an irreducible polynomial. From Theorem 5.2.2 we see that the quotient field construction produces an extension field $\mathbb{K}$ in which p has a root. On the other hand, suppose we already have *some* extension field $\mathbb{F} \subset \mathbb{L}$ and there exists $\beta \in \mathbb{L}$ which is a root of p in $\mathbb{L}$. Let $\mathbb{F}(\beta)$ denote the smallest subfield of $\mathbb{L}$ containing $\mathbb{F}$ and β. More precisely, $\mathbb{F}(\beta)$ is the intersection of all subfields of $\mathbb{L}$ containing $\mathbb{F}$ and β. By Exercise 5.2.1, $\mathbb{F}(\beta)$ is indeed a subfield of $\mathbb{L}$. Below we will see that $\mathbb{F}(\beta)$ is essentially the same as $\mathbb{K}$. To make this precise, we first need another definition and a proposition.

Proposition 5.2.6. *Suppose $\mathbb{F} \subset \mathbb{L}$, $p \in \mathbb{F}[x]$ is an irreducible polynomial, such that p has a root β in $\mathbb{L}$. If $\mathbb{K} = \mathbb{F}[x]/\mathcal{I}$ is the field constructed in Theorem 5.2.2 so that p has a root α in $\mathbb{K}$, then the inclusion $\mathbb{F} \subset \mathbb{L}$ extends to an injective field homomorphism $\mathbb{K} \to \mathbb{L}$ given by*

$$a_0 + a_1\alpha + a_2\alpha^2 + \cdots + a_{n-1}\alpha^{n-1} \mapsto a_0 + a_1\beta + a_2\beta^2 + \cdots + a_{n-1}\beta^{n-1}.$$

The image of $\mathbb{K}$ is exactly $\mathbb{F}(\beta)$, and consequently, $\mathbb{K} \cong \mathbb{F}(\beta)$.

Proof. Evaluating a polynomial in $\mathbb{F}[x]$ at β defines a homomorphism $\phi\colon \mathbb{F}[x] \to \mathbb{L}$, given by $f \mapsto f(\beta)$. We claim that $\ker(\phi) = (\!(p)\!)$. To see this, suppose $f \in \ker(\phi)$. Then $f(\beta) = 0$, and we claim that p divides f in $\mathbb{F}[x]$. If this were not the case, then because p is irreducible, we would have $\gcd(f, p) = 1$, and hence there would be $u, v \in \mathbb{F}[x]$ so that $uf + vp = 1$. Evaluating these on β gives

$$1 = u(\beta)f(\beta) + v(\beta)p(\beta) = 0 + 0 = 0,$$

a contradiction (since $0 \neq 1$ in a field). So, p divides f, and hence $f \in (\!(p)\!)$. On the other hand, if $f \in (\!(p)\!)$, then $f = up$ and $f(\beta) = u(\beta)p(\beta) = 0$, so $f \in \ker(\phi)$.

Therefore, by the first isomorphism theorem (see Exercise 3.6.12), there exists a unique injective homomorphism $\widetilde{\phi}\colon \mathbb{K} \to \mathbb{L}$. This is given by $\widetilde{\phi}(f + (\!(p)\!)) = f(\beta)$, and hence $\widetilde{\phi}(\alpha) = \widetilde{\phi}(x + (\!(p)\!)) = \beta$. For $a \in \mathbb{F}$, we have $\widetilde{\phi}(a) = \phi(a) = a$, so this extends the inclusion $\mathbb{F} \subset \mathbb{L}$. Since $\widetilde{\phi}(\alpha) = \beta$ and $\widetilde{\phi}$ is a field homomorphism, it is given by the formula from the statement of the proposition.

Finally, note that $\phi(\mathbb{K})$ is a subfield of $\mathbb{L}$ containing both $\mathbb{F}$ and β. Observe that $\mathbb{F}(\beta)$ must contain all $\mathbb{F}$–linear combinations of powers of β (because it is a subfield), and so for any $f \in \mathbb{F}[x]$, we must have $\phi(f) = f(\beta) \in \mathbb{F}(\beta)$, and so $\mathbb{F}(\beta) = \widetilde{\phi}(\mathbb{K})$, as required. □

More generally, if $\beta_1, \ldots, \beta_n \in \mathbb{L}$, we write $\mathbb{F}(\beta_1, \ldots, \beta_n) \subset \mathbb{L}$ to denote the smallest subfield of $\mathbb{L}$ containing $\mathbb{F}, \beta_1, \ldots, \beta_n$.

Example 5.2.7. In Example 5.2.5, we constructed an extension $\mathbb{K}$ from $x^3 + 3 \in \mathbb{Q}[x]$ with $[\mathbb{K} : \mathbb{Q}] = 3$. On the other hand, $-\sqrt[3]{3} \in \mathbb{R}$ is a root of $x^3 + 3$, so $\mathbb{Q}(\sqrt[3]{3}) = \mathbb{Q}(-\sqrt[3]{3}) \cong \mathbb{K}$.

The next proposition applies to the field $\mathbb{K}$ constructed in Theorem 5.2.2, but more generally, to any finite extension field.

Proposition 5.2.8. † *Suppose $\mathbb{F} \subset \mathbb{K}$ is any extension field with $[\mathbb{K} : \mathbb{F}] = n < \infty$. Then any $\beta \in \mathbb{K}$ is the root of a polynomial $f \in \mathbb{F}[x]$ with $\deg(f) \leq n$.*

Proof. Since $[\mathbb{K} : \mathbb{F}] = n < \infty$, the elements $1, \beta, \beta^2, \ldots, \beta^n$ must be linearly dependent. Therefore there exists $a_0, a_1, \ldots, a_n \in \mathbb{F}$ such that

$$a_0 + a_1\beta + a_2\beta^2 + \cdots + a_n\beta^n = 0.$$

But then $f(x) = a_0 + a_1x + \cdots + a_nx^n \in \mathbb{F}[x]$ has β as a root. □

Definition 5.2.9. *Suppose $\mathbb{F} \subset \mathbb{K}$ is an extension field. Then $\beta \in \mathbb{K}$ is said to be* **algebraic over** $\mathbb{F}$ *if β is the root of a polynomial $f \in \mathbb{F}[x]$. If every element of $\mathbb{K}$ is algebraic over $\mathbb{F}$, then we say that $\mathbb{K}$* **is algebraic over** $\mathbb{F}$. *The previous proposition therefore says that finite extensions are always algebraic.*

Proposition 5.2.10. *Suppose $\mathbb{F} \subset \mathbb{K}$ is an extension and $\beta \in \mathbb{K}$ is algebraic over $\mathbb{F}$. Then there exists a unique monic, irreducible polynomial $p \in \mathbb{F}[x]$ having β as a root. Moreover, p divides any $f \in \mathbb{F}[x]$ having β as a root.*

Proof. Let $p \in \mathbb{F}[x]$ be any monic polynomial of smallest degree having β as a root. We claim that p is irreducible. Suppose $p = ab$ for $a, b \in \mathbb{F}[x]$. Viewing $\mathbb{F}[x] \subset \mathbb{K}[x]$, we know that $x - \beta$ divides p, and hence $x - \beta$ divides either a or b (see Corollary 2.3.13). But then, if $x - \beta$ divides a we see that β is a root of a, and by minimality of the degree of p (and since $a|p$) we must have $\deg(a) = \deg(p)$, which implies $\deg(b) = 0$, and hence b is a constant. It follows that p is irreducible.

If $f, g \in \mathbb{F}[x]$ are any two polynomials having β as a root, then note that $h = \gcd(f, g)$ also has β as a root. Indeed, by Proposition 2.3.10 there exists $a, b \in \mathbb{F}[x]$ so that $h = af + bg$. But then $h(\beta) = a(\beta)f(\beta) + b(\beta)g(\beta) = 0 + 0 = 0$. If $f \in \mathbb{F}[x]$ is any polynomial with β as a root, then this implies $p = \gcd(f, p)$, and hence $f = ap$ for some $a \in \mathbb{F}[x]$. In particular, if f is also monic irreducible, then $f = p$. □

The polynomial p from Proposition 5.2.10 is called the (monic) **minimal polynomial of β over $\mathbb{F}$**.

5.2.2 Galois groups re-introduced

Recall that an automorphism of a field $\mathbb{K}$ is a field isomorphism $\sigma\colon \mathbb{K} \to \mathbb{K}$ (see Example 2.6.16). Further recall that if $\mathbb{F} \subset \mathbb{K}$ is a field extension, then $\mathrm{Aut}(\mathbb{K}, \mathbb{F})$ denotes the *group* of field automorphisms of $\mathbb{K}$ that fix every element of $\mathbb{F}$ (see Example 2.6.19). That is,

$$\mathrm{Aut}(\mathbb{K}, \mathbb{F}) = \{\sigma\colon \mathbb{K} \to \mathbb{K} \mid \sigma \text{ is an automorphism, and } \sigma(\alpha) = \alpha \text{ for all } \alpha \in \mathbb{F}\}.$$

The group $\mathrm{Aut}(\mathbb{K}, \mathbb{F})$ is called the **Galois group of $\mathbb{K}$ over $\mathbb{F}$**.

In Exercise 5.2.3 you are asked to prove that for any $\sigma \in \mathrm{Aut}(\mathbb{K}, \mathbb{F})$, $\sigma\colon \mathbb{K} \to \mathbb{K}$ is a linear transformation (necessarily invertible), when we view $\mathbb{K}$ as a vector space over $\mathbb{F}$. Thus $\mathrm{Aut}(\mathbb{K}, \mathbb{F}) < \mathrm{GL}(\mathbb{K})$ (though this notation is somewhat ambiguous since $\mathbb{K}$ is a vector space over *any* subfield, and $\mathrm{GL}(\mathbb{K})$ could refer to linear maps over any of these subfields). As we shall see, $\mathrm{Aut}(\mathbb{K}, \mathbb{F})$ is typically a very small subgroup of $\mathrm{GL}(\mathbb{K})$.

Proposition 5.2.11. † *Suppose $\mathbb{F} \subset \mathbb{K}$ is a field extension, $\beta \in \mathbb{K}$ is algebraic over $\mathbb{F}$, and that $p \in \mathbb{F}[x]$ is the minimal polynomial of β over $\mathbb{F}$. Then for any $\sigma \in \mathrm{Aut}(\mathbb{K}, \mathbb{F})$, $\sigma(\beta)$ is also a root of p. Consequently, $\mathrm{Aut}(\mathbb{K}, \mathbb{F})$ acts on the roots of p.*

Compare the proof below with the proof of Corollary 2.1.3.

Proof. Write $p = a_0 + a_1 x + \cdots + a_n x^n$. Then since σ is an automorphism *fixing every element of* $\mathbb{F}$, we have

$$0 = \sigma(0) = \sigma(p(\beta)) = \sigma(a_0) + \sigma(a_1)\sigma(\beta) + \cdots + \sigma(a_n)\sigma(\beta)^n = a_0 + a_1\sigma(\beta) + \cdots + a_n\sigma(\beta)^n = p(\sigma(\beta)).$$

□

Since the number of roots of a polynomial provides a lower bound on the degree (see Proposition 2.3.17), we have the following immediate corollary of the previous proposition.

Corollary 5.2.12. *Suppose $\mathbb{F} \subset \mathbb{K}$ is a field extension and $\alpha \in \mathbb{K}$ is algebraic over $\mathbb{F}$. If the orbit of α by $\mathrm{Aut}(\mathbb{K}, \mathbb{F})$ has n elements, then the minimal polynomial of α over $\mathbb{F}$ has degree at least n.*

Proof. Every element of the orbit of α is a root of the minimal polynomial of α according to Proposition 5.2.11. Consequently, the degree of this polynomial is at least the number of elements in the orbit. □

Corollary 5.2.13. *If $\mathbb{F} \subset \mathbb{K}$ is a finite extension, then $\mathrm{Aut}(\mathbb{K}, \mathbb{F})$ is finite.*

Proof. Let $\alpha_1, \ldots, \alpha_k \in \mathbb{K}$ be a basis for $\mathbb{K}$ over $\mathbb{F}$, and let $p_1, \ldots, p_k$ be the minimal polynomials of these elements. The Galois group $G = \mathrm{Aut}(\mathbb{K}, \mathbb{F})$ acts on the *finite* set $X \subset \mathbb{K}$ consisting of the roots of $p_1, \ldots, p_k$, which includes $\alpha_1, \ldots, \alpha_k$. The intersection of the stabilizers $H = \mathrm{stab}_G(\alpha_1) \cap \cdots \cap \mathrm{stab}_G(\alpha_k)$ must be the identity, $H = \{e\}$, since any element of $\mathrm{Aut}(\mathbb{K}, \mathbb{F})$ is determined by what it does to the basis $\alpha_1, \ldots, \alpha_k$ (since automorphisms are linear; see Exercise 5.2.3).

Finally, by the orbit stabilizer theorem (Theorem 4.1.12) and Corollary 5.2.12,

$$[G : \mathrm{stab}_G(\alpha_i)] = |G \cdot \alpha_i| \leq \deg(p_i) < \infty.$$

Therefore, $\mathrm{stab}_G(\alpha_i)$ has finite index in G, and hence so does $H = \{e\}$, so that $|G| = [G : H] < \infty$. □

Example 5.2.14. Note that $\mathbb{R} \subset \mathbb{C}$ is a finite extension: $[\mathbb{C} : \mathbb{R}] = 2 < \infty$. As a basis of $\mathbb{C}$ over $\mathbb{R}$ we have $1, i$. The minimal polynomial of 1 is $x - 1$ and the minimal polynomial of i is $x^2 + 1$. We have already seen that $x^2 + 1$ is irreducible (since it has no roots in $\mathbb{R}$ and it has degree 2). We have already observed that $\sigma(a + bi) = \overline{a + bi} = a - bi$, defines an element $\sigma \in \mathrm{Aut}(\mathbb{C}, \mathbb{R})$ of order 2 (see Example 2.6.19). On the other hand $G = \mathrm{Aut}(\mathbb{C}, \mathbb{R})$ acts on the set of roots of these polynomials $\{1, i, -i\}$. The orbit of 1 is $G \cdot 1 = \{1\}$ since there is no other root of $x - 1$, while the orbit of i is $G \cdot i = \{i, -i\}$ since $\sigma \cdot i = \sigma(i) = -i$. From this we also see that $\mathrm{stab}_G(1) = G$ and $[G : \mathrm{stab}_G(i)] = 2$. Since $\{e\} = \mathrm{stab}_G(1) \cap \mathrm{stab}_G(i) = G \cap \mathrm{stab}_G(i) = \mathrm{stab}_G(i)$, it follows that $|G| = [G : \{e\}] = [G : \mathrm{stab}_G(i)] = 2$. That is,

$$G = \mathrm{Aut}(\mathbb{C}, \mathbb{R}) = \{e, \sigma\} \cong \mathbb{Z}_2.$$

Example 5.2.15. In Example 5.2.5 and 5.2.7 we constructed a field $\mathbb{K} \cong \mathbb{Q}(\sqrt[3]{3})$ of degree 3 over $\mathbb{Q}$ from the irreducible polynomial $x^3 + 3$. We have $1, \sqrt[3]{3}, \sqrt[3]{9}$ as a basis for $\mathbb{Q}(\sqrt[3]{3})$ over $\mathbb{Q}$ (why?). Since the other roots of $x^3 + 3$ are **not** in $\mathbb{Q}(\sqrt[3]{3})$, any element $\sigma \in G = \mathrm{Aut}(\mathbb{Q}(\sqrt[3]{3}), \mathbb{Q})$ must act trivially on $\sqrt[3]{3}$ and hence also on $\sqrt[3]{9} = (\sqrt[3]{3})^2$ (and of course on 1). That is, σ must be the identity, and we see that $\mathrm{Aut}(\mathbb{Q}(\sqrt[3]{3}), \mathbb{Q}) = \{e\}$.

Exercises.

Exercise 5.2.1. Prove that if $\mathbb{L}$ is a field and $\mathcal{H}$ is a collection of subfields of $\mathbb{L}$, then

$$\bigcap_{\mathbb{K} \in \mathcal{H}} \mathbb{K}$$

is a subfield of $\mathbb{L}$.

Exercise 5.2.2. Prove that for any field $\mathbb{F}$ and any nonconstant polynomial $f \in \mathbb{F}[x]$, there exists a field $\mathbb{L}$ such that f factors into linear factors over $\mathbb{L}$. *Hints: Write f as a product of irreducible factors, and apply Theorem 5.2.2 to one of those factors to produce $\mathbb{F} \subset \mathbb{K}_1$. Repeat and induct appropriately.*

Exercise 5.2.3. Suppose $\mathbb{F} \subset \mathbb{K}$ is a field extension. Prove that if $\sigma \in \mathrm{Aut}(\mathbb{K}, \mathbb{F})$, then $\sigma \colon \mathbb{K} \to \mathbb{K}$ is a linear transformation, when we view $\mathbb{K}$ as a vector space over $\mathbb{F}$.

Exercise 5.2.4. Suppose $\mathbb{F} \subset \mathbb{K} \subset \mathbb{L}$ are field extensions. Prove that $[\mathbb{L} : \mathbb{F}] = [\mathbb{L} : \mathbb{K}][\mathbb{K} : \mathbb{F}]$.

Exercise 5.2.5. Consider the subfield $\mathbb{Q}(\sqrt{2}, \sqrt{3}) \subset \mathbb{R}$.
a. Prove that $[\mathbb{Q}(\sqrt{2}, \sqrt{3}) : \mathbb{Q}] = 4$ by proving that $[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 2$, and that $x^2 - 3$ is irreducible in $\mathbb{Q}(\sqrt{2})[x]$, then appeal to Proposition 5.2.6 and Exercise 5.2.4. *Hint: to prove $x^2 - 3$ is irreducible, suppose there is $a + b\sqrt{2} \in \mathbb{Q}(\sqrt{2})$ so that $(a + b\sqrt{2})^2 - 3 = 0$ and derive a contradiction.*
b. Prove that $1, \sqrt{2}, \sqrt{3}, \sqrt{6}$ is a basis for $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ over $\mathbb{Q}$.
c. Prove that $\mathrm{Aut}(\mathbb{Q}(\sqrt{2}, \sqrt{3}), \mathbb{Q}) \cong \mathbb{Z}_2 \times \mathbb{Z}_2$. *Hint:* $\mathrm{Aut}(\mathbb{Q}(\sqrt{2}, \sqrt{3}), \mathbb{Q})$ *acts on the roots* $\{\sqrt{2}, -\sqrt{2}, \sqrt{3}, -\sqrt{3}\}$ *of* $x^2 - 2$ *and* $x^2 - 3$.

Exercise 5.2.6. Let $\zeta_8 = e^{i\pi/4} \in \mathbb{C}$ and consider the subfield $\mathbb{Q}(\zeta_8)$.
a. Prove $\mathbb{Q}(\zeta_8) = \mathbb{Q}(\sqrt{2}, i)$.
b. Prove that $\mathrm{Aut}(\mathbb{Q}(\sqrt{2}, i), \mathbb{Q}) \cong \mathbb{Z}_2 \times \mathbb{Z}_2$.
c. Prove that $x^4 + 1 \in \mathbb{Q}[x]$ is the minimal polynomial of ζ_8.

You should...

- Be able to do all the exercises from this section and know the proofs of statements marked by †.
- Know how the field $\mathbb{K}$ is constructed in Theorem 5.2.2 to produce an extension in which p has a root. Specifically, know that this the quotient group of $\mathbb{F}[x]$ by $\mathcal{I} = \{ap \mid a \in \mathbb{F}[x]\}$.
- Understand the relationship between the abstract construction of Theorem 5.2.2 and concrete extensions (see 5.2.6).
- Know what an algebraic extension is and what the minimal polynomial of element in an algebraic extension is.
- Know what the Galois group is. Be able to study this using the action on the roots.
- Be familiar with the examples in this section and in the exercises.

5.3 Fundamental theorem of Galois Theory

In Example 5.2.14 we saw that $|\mathrm{Aut}(\mathbb{C}, \mathbb{R})| = 2$, and $[\mathbb{C} : \mathbb{R}] = 2$. Similarly, in Exercises 5.2.5 and 5.2.6, you find extensions $\mathbb{F} \subset \mathbb{K}$ such that $[\mathbb{K} : \mathbb{F}] = 4 = |\mathrm{Aut}(\mathbb{K}, \mathbb{F})|$. Closer inspection of these examples would reveal a much deeper connection between the Galois group and the extension. This is explained by the *Fundamental Theorem of Galois Theory*, which we state after providing a few required definitions.

Definition 5.3.1. *Suppose that $\mathbb{F} \subset \mathbb{K}$ is a finite extension field and let $G = \mathrm{Aut}(\mathbb{K}, \mathbb{F})$. For any subgroup $H < G$, the* **fixed field** *of H is defined to be*

$$H' = \mathrm{Fix}_{\mathbb{K}}(H) = \{k \in \mathbb{K} \mid h(k) = k \text{ for all } h \in H\}.$$

For any subfield $\mathbb{L} \subset \mathbb{K}$, set

$$\mathbb{L}' = \bigcap_{k \in \mathbb{L}} \mathrm{stab}_G(k) = \{g \in G \mid g(k) = k \text{ for all } k \in \mathbb{L}\}.$$

Finally, we say that $\mathbb{K}$ is a **Galois extension of** *$\mathbb{F}$, if $G' = \mathbb{F}$. That is, $\mathbb{K}$ is a Galois extension of $\mathbb{F}$ if $\mathbb{F}$ is precisely the set of elements fixed by G.*[1]

Lemma 5.3.2. *Suppose $\mathbb{F} \subset \mathbb{K}$ and $G = \mathrm{Aut}(\mathbb{K}, \mathbb{F})$. For any subfield $\mathbb{L} \subset \mathbb{K}$, $\mathbb{L}'$ is a subgroup of G and for any subgroup $H < G$, $H' \subset \mathbb{K}$ is a subfield.*

Proof. Exercise. □

Theorem 5.3.3 (Fundamental Theorem of Galois Theory)**.** *Suppose $\mathbb{K}$ is a Galois extension of $\mathbb{F}$ and $G = \mathrm{Aut}(\mathbb{K}, \mathbb{F})$. Then the "priming operations" determine inverse bijections between the set of subfields of $\mathbb{K}$ and subgroups of G:*

$$\begin{array}{ccc} \{\mathbb{L} \subset \mathbb{K}\} & \rightleftarrows & \{H < G\} \\ \mathbb{L} & \longmapsto & \mathbb{L}' \\ \| & & \| \\ H' & \longmapsfrom & H \end{array}$$

Furthermore, for all $\mathbb{L} \subset \mathbb{M} \subset \mathbb{K}$ we have

[1] Our definition of a Galois extension assumes that it is a finite extension. This is not completely standard in the literature.

(i) $\mathbb{M}' < \mathbb{L}'$,

(ii) $[\mathbb{M} : \mathbb{L}] = [\mathbb{L}' : \mathbb{M}']$,

(iii) $\mathrm{Aut}(\mathbb{K}, \mathbb{L}) = \mathbb{L}'$,

(iv) $\mathbb{M}$ *is Galois over* $\mathbb{L}$ *if and only if* $\mathbb{M}' \triangleleft \mathbb{L}'$. *In this case* $\mathrm{Aut}(\mathbb{M}, \mathbb{L}) \cong \mathbb{L}'/\mathbb{M}'$.

We note that the assumption that $\mathbb{K}$ is Galois over $\mathbb{F}$ cannot be omitted. Indeed, in Example 5.2.15, we saw that $\mathrm{Aut}(\mathbb{Q}(\sqrt[3]{3}), \mathbb{Q})$ is trivial, while $[\mathbb{Q}(\sqrt[3]{3}) : \mathbb{Q}] = 3$. It is clear that the Galois assumption is not satisfied since $G = \mathrm{Aut}(\mathbb{Q}(\sqrt[3]{3}), \mathbb{Q})$ being trivial means that $G' = \mathbb{Q}(\sqrt[3]{3})$, instead of being equal to $\mathbb{Q}$ (as is required for the extension to be Galois).

We would like to look at some examples to illustrate the theorem. However, we must first be able to decide when an extension is Galois. The next theorem provides just such a criterion.

Theorem 5.3.4. *Suppose* $f \in \mathbb{Q}[x]$ *is any nonconstant polynomial, and* $\mathbb{K} \subset \mathbb{C}$ *is the smallest subfield containing all the roots of* f. *Then* $\mathbb{K}$ *is Galois over* $\mathbb{Q}$.

The restriction to polynomials in $\mathbb{Q}[x]$ is only for convenience, though some additional assumptions must be made. In fact, they key property is that $\mathbb{Q}$ does not contain $\mathbb{Z}_p$ as a subfield, for any p. This is sometimes expressed by saying that $\mathbb{Q}$ has **characteristic zero**. A similar theorem is true for any polynomial $f \in \mathbb{F}[x]$, where $\mathbb{F}$ has characteristic zero. In this case, we consider the smallest field extension $\mathbb{K}$ in which f factors into linear factors (see Exercise 5.2.2).

With Theorem 5.3.4 in hand, we look at an example.

Example 5.3.5. Consider the field $\mathbb{Q}(\sqrt[3]{3}, \zeta_3) \subset \mathbb{C}$, where ζ_3 is a primitive third root of 1 (that is, $\zeta_3 \in C_3$, a third root of 1 with order 3). Specifically, we can take $\zeta_3 = \frac{-1+i\sqrt{3}}{2}$. Note that $\mathbb{Q}(\sqrt[3]{3}, \zeta_3)$ is a degree two extension of $\mathbb{Q}(\sqrt[3]{3})$, which is itself a degree three extension over $\mathbb{Q}$. Consequently (from Exercise 5.2.4),

$$[\mathbb{Q}(\sqrt[3]{3}, \zeta_3) : \mathbb{Q}] = [\mathbb{Q}(\sqrt[3]{3}, \zeta_3), \mathbb{Q}(\sqrt[3]{3})][\mathbb{Q}(\sqrt[3]{3}) : \mathbb{Q}] = 2 \cdot 3 = 6.$$

Explicitly, $\mathcal{B} = \{1, \sqrt[3]{3}, \sqrt[3]{9}, \zeta_3, \sqrt[3]{3}\zeta_3, \sqrt[3]{9}\zeta_3\}$ is a basis for $\mathbb{Q}(\sqrt[3]{3})$ over $\mathbb{Q}$ (check this!). We note that $\mathbb{Q}(\sqrt[3]{3}, \zeta_3)$ is the smallest subfield of $\mathbb{C}$ containing the three roots of $x^3 + 3$, namely $-\sqrt[3]{3}, -\sqrt[3]{3}\zeta_3, -\sqrt[3]{3}\zeta_3^2$. By Theorem 5.3.4, $\mathbb{Q}(\sqrt[3]{3}, \zeta_3)$ is Galois over $\mathbb{Q}$ and by Theorem 5.3.3 the Galois group $G = \mathrm{Aut}(\mathbb{Q}(\sqrt[3]{3}, \zeta_3), \mathbb{Q})$ has order 6.

On the other hand, G acts on the set of roots $X = \{-\sqrt[3]{3}, -\sqrt[3]{3}\zeta_3, -\sqrt[3]{3}\zeta_3^2\}$ of $x^3 + 3$, which defines a homomorphism $\alpha \colon G \to \mathcal{S}ym(X)$. Note that if $\sigma \in G$ and $\alpha(\sigma) = \mathrm{id}_X$, then $\sigma(-\sqrt[3]{3}) = -\sqrt[3]{3}$ and $\sigma(-\sqrt[3]{3}\zeta_3) = -\sqrt[3]{3}\zeta_3$ and consequently,

$$\sigma(\zeta_3) = \sigma(\frac{1}{\sqrt[3]{3}}\sqrt[3]{3}\zeta_3) = \sigma(\frac{1}{\sqrt[3]{3}})\sigma(\sqrt[3]{3}\zeta_3) = \frac{1}{\sqrt[3]{3}}\sqrt[3]{3}\zeta_3 = \zeta_3.$$

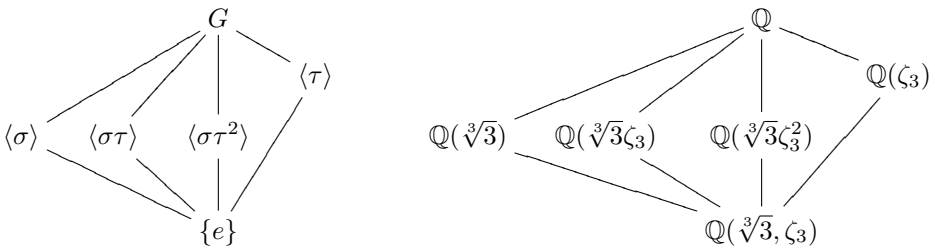
Examining the effect on the basis $\mathcal{B}$, we see that $\sigma = e$, the identity in G. That is, α is injective. Since $|X| = 3$ and $|\mathcal{S}ym(X)| = 6$, it follows that α is an isomorphism, proving $G \cong \mathcal{S}ym(X) \cong S_3$.

From this we see that any permutation of X extends uniquely to an element of G. For example, the element σ that fixes $-\sqrt[3]{3}$ and interchanges $-\sqrt[3]{3}\zeta_3$ and $-\sqrt[3]{3}\zeta_3^2$ is the restriction of complex conjugation to $\mathbb{Q}(\sqrt[3]{3}, \zeta_3)$ (indeed, $\zeta_3^2 = \overline{\zeta_3}$). Let $\tau \in G$ denote the unique element that cyclically permutes the roots as

$$-\sqrt[3]{3} \mapsto -\sqrt[3]{3}\zeta_3 \mapsto -\sqrt[3]{3}\zeta_3^2 \mapsto -\sqrt[3]{3}.$$

Then $G = \{e, \tau, \tau^2, \sigma, \sigma\tau, \sigma\tau^2\}$, with each of $\sigma, \sigma\tau, \sigma\tau^2$ elements of order two, while τ and τ^2 are the elements of order 3.

To visualize Theorem 5.3.3, we can draw the subgroup lattice and next to it, draw the corresponding fixed subfields of $\mathbb{Q}(\sqrt[3]{3}, \zeta_3)$:



(Verify that the fixed fields are as shown.) The only nontrivial normal subgroup here is $\langle\tau\rangle \triangleleft G$. Consequently, the theorem tells us that

$$\mathrm{Aut}(\mathbb{Q}(\zeta_3), \mathbb{Q}) \cong G/\langle\tau\rangle \cong \mathbb{Z}_2.$$

This order two Galois group is nothing but complex conjugation restricted to $\mathbb{Q}(\zeta_3)$, which we can view as the coset $\sigma\langle\tau\rangle \in G/\langle\tau\rangle$.

Another very interesting family of examples are the **cyclotomic extensions**, which we now describe. For any n, let ζ_n denote a primitive n^{th} root of 1, that is, an element $\zeta_n \in C_n$ with order n. Consider the subfield $\mathbb{Q}(\zeta_n) \subset \mathbb{C}$. Note that $\mathbb{Q}(\zeta_n) = \mathbb{Q}$ for $n = 1$ and 2. We already saw $\mathbb{Q}(\zeta_3)$ in the previous example. Since the roots of $x^n - 1$ in $\mathbb{C}$ are $1, \zeta_n, \zeta_n^2, \cdots, \zeta_n^{n-1}$, it follows that all the roots of $x^n - 1$ are in $\mathbb{Q}(\zeta_n)$ and hence $\mathbb{Q}(\zeta_n)$ is Galois over $\mathbb{Q}$ by Theorem 5.3.4.

Theorem 5.3.6. $[\mathbb{Q}(\zeta_n) : \mathbb{Q}] = \varphi(n)$, *the Euler phi-function evaluated at* n.

Using this theorem (and Exercise 3.7.7) we can describe the Galois group $G_n = \mathrm{Aut}(\mathbb{Q}(\zeta_n), \mathbb{Q})$. Specifically, observe that C_n is a *subgroup* of the multiplicative group of units $C_n < \mathbb{Q}(\zeta_n)^\times$. Since every element of G_n induces a group homomorphism of $\mathbb{Q}(\zeta_n)^\times$ and permutes the roots of $x^n - 1$, we obtain a homomorphism $\psi\colon G_n \to \mathrm{Aut}(C_n)$. This is clearly injective, while on the other hand, $C_n \cong \mathbb{Z}_n$ means $\mathrm{Aut}(C_n) \cong \mathrm{Aut}(\mathbb{Z}_n) \cong \mathbb{Z}_n^\times$ (see Exercise 3.7.7). Therefore $|G_n| = \varphi(n)$, and hence ψ is an isomorphism.

Any $\sigma \in G_n$ is determined by $\sigma(\zeta_n)$, and since σ induces an automorphism of C_n, $\sigma(\zeta_n)$ must be another generator of C_n, which is an element ζ_n^a for some $a \in \mathbb{Z}$ such that $\gcd(a, n) = 1$. This provides an explicit isomorphism $\alpha\colon \mathbb{Z}_n^\times \to G_n$ defined by $\alpha_a(\zeta_n) = \zeta_n^a$ (note that with our usual abuse of notation, a is a representative of a congruence class modulo n, but it is easy to see that α_a is independent of the choice).

For every divisor $d|n$, we have $\mathbb{Q}(\zeta_d) \subset \mathbb{Q}(\zeta_n)$ since $\zeta_n^{n/d}$ is a primitive d^{th} root of unity. There are other subfields, in general, though. For example, in Exercise 5.2.6 we see that $\mathbb{Q}(\sqrt{2}) \subset \mathbb{Q}(\zeta_8)$.

We can find all subfields by listing all subgroups. This is particularly easy in the case that $n = p$, a prime, since in this case $G_p \cong \mathbb{Z}_p^\times$ is a cyclic group of order $p-1$ (hence isomorphic to $\mathbb{Z}_{p-1}$; see Proposition 4.3.12). We also note that the minimal polynomial of ζ_p over $\mathbb{Q}$, which we denote $\Phi_p(x)$, has a very simple form:

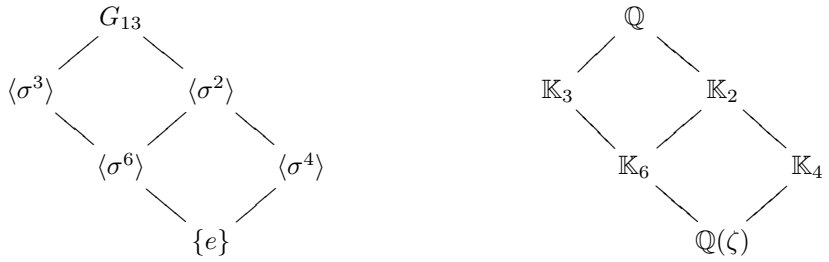
$$\Phi_p(x) = (x^p - 1)/(x - 1) = 1 + x + x^2 + \cdots + x^{p-1}.$$

Example 5.3.7. Let's consider the case $p = 13$. For simplicity of notation, set $\zeta = \zeta_{13}$, and consider $G_{13} = \mathrm{Aut}(\mathbb{Q}(\zeta), \mathbb{Q}) \cong \mathbb{Z}_{13}^\times \cong \mathbb{Z}_{12}$. We may verify that $2 \in \mathbb{Z}_{13}^\times$ is a generator (or more precisely, $[2]$ is a generator), and therefore $\sigma \in G_{13}$ defined by $\sigma(\zeta) = \zeta^2$ is a generator for $G_{13} \cong \mathbb{Z}_{13}^\times$.

According to Theorem 3.2.10, the subgroups of $\mathbb{Z}_{12}$ are precisely the cyclic subgroups $\langle d\rangle$ for every divisor d of 12. The isomorphism $\eta\colon \mathbb{Z}_{12} \to G_{13}$ is given by $\eta(a) = \sigma^a$ (or more precisely, $\eta([a]) = \sigma^a$). Therefore the subgroups of G_p are precisely the images $\eta(\langle d\rangle) = \langle\sigma^d\rangle < G_{13}$ for $d = 1, 2, 3, 4, 6, 12$ (the divisors of 12).

With this we can compute the subgroup lattice of G_{13} (compare Example 3.2.12) with the objective of

finding the corresponding fixed fields on the right.



To compute the fixed fields $\mathbb{K}_j$ of $\langle\sigma^j\rangle$ for each of $j = 2, 3, 4, 6$, it is convenient to write down what the generator σ does to the 12 roots $\zeta, \zeta^2, \cdots, \zeta^{12}$ of the minimal polynomial $\Phi_{13}(x) = 1 + x + x^2 + \cdots + x^{12}$. Doing this, we see that σ permutes the roots as follows:

$$\zeta \mapsto \zeta^2 \mapsto \zeta^4 \mapsto \zeta^8 \mapsto \zeta^3 \mapsto \zeta^6 \mapsto \zeta^{12} \mapsto \zeta^{11} \mapsto \zeta^9 \mapsto \zeta^5 \mapsto \zeta^{10} \mapsto \zeta^7 \mapsto \zeta.$$

Alternatively, we can simply record the permutation of the exponents $\sigma = (1\ 2\ 4\ 8\ 3\ 6\ 12\ 11\ 9\ 5\ 10\ 7) \in S_{12}$. Then we also have

$$\sigma^2 = (1\ 4\ 3\ 12\ 9\ 10)(2\ 8\ 6\ 11\ 5\ 7) \text{ and } \sigma^4 = (1\ 3\ 9)(4\ 12\ 10)(2\ 6\ 5)(8\ 11\ 7)$$

$$\sigma^3 = (1\ 8\ 12\ 5)(2\ 3\ 11\ 10)(4\ 6\ 9\ 7) \text{ and } \sigma^6 = (1\ 12)(8\ 5)(2\ 11)(3\ 10)(4\ 9)(6\ 7).$$

Note that the sum of any of the roots which are cyclically permuted by σ^j is **fixed** by σ^j. For example, $\zeta+\zeta^4+\zeta^3+\zeta^{12}+\zeta^9+\zeta^{10}$ is fixed by σ^2. Because the minimal polynomial of ζ is $\Phi_{13}(x) = 1+x+x^2+\cdots+x^{12}$, the sum of all the roots is -1, and so for example, we have

$$\zeta^2 + \zeta^8 + \zeta^6 + \zeta^{11} + \zeta^5 + \zeta^7 = -1 - (\zeta + \zeta^4 + \zeta^3 + \zeta^{12} + \zeta^9 + \zeta^{10})$$

and hence

$$\mathbb{Q}(\zeta + \zeta^4 + \zeta^3 + \zeta^{12} + \zeta^9 + \zeta^{10}) = \mathbb{Q}(\zeta + \zeta^4 + \zeta^3 + \zeta^{12} + \zeta^9 + \zeta^{10}, \zeta^2 + \zeta^8 + \zeta^6 + \zeta^{11} + \zeta^5 + \zeta^7)$$

We claim that this is precisely the fixed field $\mathbb{K}_2$ of $\langle\sigma^2\rangle$. Indeed, this field is clearly fixed by σ^2, while

$$\sigma(\zeta + \zeta^4 + \zeta^3 + \zeta^{12} + \zeta^9 + \zeta^{10}) = \zeta^2 + \zeta^8 + \zeta^6 + \zeta^{11} + \zeta^5 + \zeta^7,$$

(check it!). We also note that $\zeta + \zeta^4 + \zeta^3 + \zeta^{12} + \zeta^9 + \zeta^{10}$ cannot be a rational number $b \in \mathbb{Q}$ since otherwise ζ would be a root $x + x^4 + x^3 + x^{12} + x^9 + x^{10} - b$, which is not a multiple of $\Phi_p(x)$ (see Proposition 5.2.10). Therefore, σ does not fix the field (and hence it is not equal to $\mathbb{Q}$), proving

$$\mathbb{K}_2 = \mathbb{Q}(\zeta + \zeta^4 + \zeta^3 + \zeta^{12} + \zeta^9 + \zeta^{10}).$$

We likewise find

$$\mathbb{K}_4 = \mathbb{Q}(\zeta + \zeta^3 + \zeta^9, \zeta^4 + \zeta^{12} + \zeta^{10}, \zeta^2 + \zeta^6 + \zeta^5),$$

$$\mathbb{K}_3 = \mathbb{Q}(\zeta + \zeta^8 + \zeta^{12} + \zeta^5, \zeta^2 + \zeta^3 + \zeta^{11} + \zeta^{10}),$$

and

$$\mathbb{K}_6 = \mathbb{Q}(\zeta + \zeta^{12}, \zeta^8 + \zeta^5, \zeta^2 + \zeta^{11}, \zeta^3 + \zeta^{10}, \zeta^4 + \zeta^9).$$

Since G_{13} is abelian, all subgroups are normal. Consequently, all extensions are Galois. For example, $\mathbb{K}_6$ is Galois over $\mathbb{K}_3$ and $\mathrm{Aut}(\mathbb{K}_6, \mathbb{K}_3) \cong \langle\sigma^3\rangle/\langle\sigma^6\rangle \cong \mathbb{Z}_2$.

5.3.1 Final thoughts

At the very end of Section 2.3 we discussed the quadratic formula and mentioned related formulas for degree 3 and 4 polynomials. At that time we also noted that for the polynomial $f = x^5 - x - 1 \in \mathbb{Q}[x]$, we cannot solve for the roots simply using arithmetic and taking radicals, that is applying $\sqrt[n]{\ }$. Theorem 5.3.3 is the key ingredient to understanding this. The point is that being able to solve for the roots using radicals means that the roots lie in a field of the form $\mathbb{Q}(\zeta_n, \beta_1, \beta_2, \ldots, \beta_k) \subset \mathbb{C}$ where β_i is a root of the polynomial $x^{m_i} - d_i \in \mathbb{Q}(\beta_1, \ldots, \beta_{i-1})[x]$ and n is an integer making $\mathbb{Q}(\zeta_n, \beta_1, \beta_2, \ldots, \beta_k)$ into a Galois extension of $\mathbb{Q}$. In this case, the Galois group $G = \mathrm{Aut}(\mathbb{Q}(\zeta_n, \beta_1, \ldots, \beta_k), \mathbb{Q})$ enjoys a nice property called **solvability**, which means that it has a chain of normal subgroups

$$\{e\} \triangleleft G_1 \triangleleft G_2 \triangleleft \cdots \triangleleft G_{n-1} \triangleleft G_n = G$$

where G_k/G_{k-1} is abelian (this is not too difficult to see, since the action on roots of polynomials of the form $x^{m_i} - d_i$ are easy to understand). That G has such a chain of subgroups is not too difficult to prove using Theorems 5.3.3 and 5.3.4. It turns out that any Galois extension of $\mathbb{Q}$ containing the roots of $f = x^5 - x - 1$, cannot have this property. In fact, if we take the smallest subfield in which f can be factored into linear factors, this subfield has degree 5!, and the Galois group is isomorphic to S_5 (the reason for this involves the finite fields $\mathbb{Z}_p$!). The group S_5 has only one nontrivial normal subgroup, namely A_5, the subgroup of even permutations, and A_5 has **no** nontrivial normal subgroups. One must study chains of subgroups as described above in more detail to see why this implies that *no* extension in which this has all its roots can have solvable Galois group, but at this point, all proofs have been abandoned and we are only telling a story. To truly understand, the interested student will need to delve deeper into *abstract algebra*.....

You should...

- Be able to understand this section well enough to do the take home problem for the final exam.