Chapter 1: From integers to monoids

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

These techniques do not scale well to larger problems. Because of this, algebraists typically take a different turn, one that develops not just techniques, but structures and viewpoints that can be used to solve a vast array of problems, many of which are surprisingly different.

This chapter serves two purposes. First, we re-present ideas you have seen before, but state them in fairly precise terms, which we will then use repeatedly, and require you to use, so as to encourage you to reason with precise meanings of words. This is motivated by a desire for clarity and reproducibility; too often, people speak vaguely to each other, and words contain different meanings for different people.

On the other hand, we also try to introduce some very important algebraic ideas, but intuitively. We will use very concrete examples. True, these examples are probably not as concrete as you might like, but believe me when I tell you that the examples I will use are more concrete than the usual presentation. One goal is to get you to use these examples when thinking about the more general ideas later on. It will be important not only that you reproduce what you read here, but that you explore and play with the ideas and examples, specializing or generalizing them as needed to attack new problems.

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

1.1: Foundations

This chapter focuses on two familiar objects of study: the integers and the monomials. They share a number of important parallels that lay the foundation for the first algebraic structure that we will study. Before we investigate that in detail, let's turn to some general tools of mathematics that you should have seen before now.

Sets

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

Notation 1.1. If A is a subset of B, we write $A \subseteq B$. If A is a proper subset, we can still write $A \subsetneq B$, but if we want to emphasize that they are not equal, we write $A \subsetneq B$.

You should recognize these sets:

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

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

Notation 1.2. Notice that both $\mathbb{N} \subseteq \mathbb{Z}$ and $\mathbb{N} \subsetneq \mathbb{Z}$ are true.

We can put sets together in several ways.

Definition 1.3. Let S and T be two sets. The **Cartesian product of S and T** is the set of ordered pairs

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

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

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

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

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

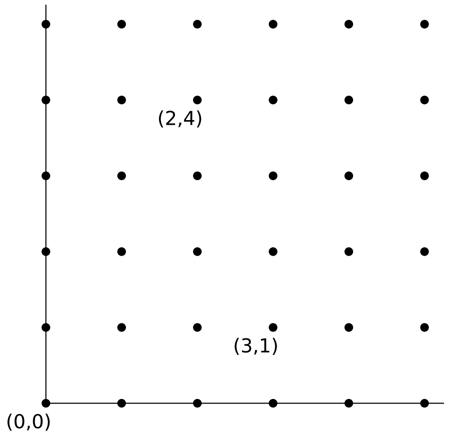
and the **difference** of S and T is the set

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

Example 1.4. Suppose $S = \{a, b\}$ and $T = \{x+1, y-1\}$. By definition,

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

Example 1.5. If we let $S = T = \mathbb{N}$, then $S \times T = \mathbb{N} \times \mathbb{N}$, the set of all ordered pairs whose entries are natural numbers. We can visualize this as a **lattice**, where points must have integer co-ordinates:



Relations

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

Definition 1.6. Any subset of $S \times T$ is **relation on the sets S and T**. A **function** is any relation f such that $(a,b) \in f$ implies $(a,c) \notin f$ for any $c \neq b$. An **equivalence relation on S** is a subset R of $S \times S$ that satisfies the properties

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

Notation 1.7. Even though relations and functions are sets, we usually write them in the manner to which you are accustomed.

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

Example 1.8. Define a relation $\sim$ on $\mathbb{Z}$ in the following way. We say that $a \sim b$ if $ab \in \mathbb{N}$. Is this an equivalence relation?

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

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

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

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

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

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

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

Binary operations

An important relation is defined by an operation.

Definition 1.9. Let S and T be sets. An **binary operation from S to T** is a function $f : S \times S \to T$. If $S = T$, we say that f is a binary operation **on** S. A binary operation f on S is **closed** if $f(a, b)$ is defined for all $a, b \in S$.

Example 1.10. Addition of the natural numbers is a function, $+ : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$; the sentence, $2 + 3 = 5$ can be thought of as $+(2,3) = 5$. Hence, addition is a binary operation on $\mathbb{N}$. Addition is defined for all natural numbers, so it is closed.

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

In each set described above, you can perform arithmetic: add, subtract, multiply, and (in most cases) divide. We need to make the meaning of these operations precise.[6]

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

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

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

- If $a = b$, then substitution implies that $a + (-b) = b + (-b) = 0$.
- Otherwise, let A be any set with a objects.
    - If I can remove a set with b objects from A, and have at least one object left over, let $c \in \mathbb{N}^+$ be the number of objects left over; then we define $a + (-b) = c$.
    - If I *cannot* remove a set with b objects from A, then let $c \in \mathbb{N}^+$ be the smallest number of objects I would need to add to A so that I could remove b objects. This satisfies the equation $a + c = b$; we then define $a + (-b) = -c$.

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

- $0 \cdot b = 0$ and $b \cdot 0 = 0$;

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

- $a \cdot b$ is the result of adding a copies of b, or

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

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

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

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

Notation 1.11. For convenience, we usually write $a-b$ instead of $a+(-b)$.

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

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

Orderings

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

Definition 1.12. For any two elements $a,b \in \mathbb{Z}$, we say that:

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

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

Remark 1.13. Do not yet assume certain "natural" properties of these orderings. For example, it is true that if $a \leq b$, then either $a<b$ or $a=b$. But why? You can reason to it from the definitions given here, so you should do so.

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

Some orderings enjoy special properties.

Definition 1.14. Let S be any set. A **linear ordering** on S is a relation $\sim$ where for any $a,b \in S$ one of the following holds:

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

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

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

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

Theorem 1.15. The relations $<$, $>$, $\leq$, and $\geq$ are linear orderings of $\mathbb{Z}$.

Our "proof" relies on some unspoken assumptions: in particular, the arithmetic on $\mathbb{Z}$ that we described before. Try to identify where these assumptions are used, because when you write your own proofs, you have to ask yourself constantly: Where am I using unspoken assumptions? In such places, either the assertion must be something accepted by the audience,[7] or you have to cite a reference your audience accepts, or you have to prove it explicitly. It's beyond the scope of this course to discuss these assumptions in detail, but you should at least try to find them.

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

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

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

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

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

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

Definition 1.16. Let S be a set and $\prec$ a linear ordering on S. We say that $\prec$ is a **well-ordering** if

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

Example 1.17. The relation $<$ is *not* a well-ordering of $\mathbb{Z}$, because $\mathbb{Z}$ itself has no smallest element under the ordering.

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

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

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

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

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

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

That is, any subset of $\mathbb{N}$, ordered by these orderings, has a smallest element. This may sound obvious, but it is very important, and what is remarkable is that *no one can prove it.*[8] It is an assumption about the natural numbers. This is why we state it as a principle (or axiom, if you prefer). In the future, if we talk about the well-ordering of $\mathbb{N}$, we mean the well-ordering $<$.

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

Theorem 1.18. Let $a_1 \geq a_2 \geq \cdots$ be a nonincreasing sequence of natural numbers. The sequence eventually stabilizes; that is, at some index i, $a_i = a_{i+1} = \cdots$.

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

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

Theorem 1.19 (Mathematical Induction)**.** Let P be a subset of $\mathbb{N}^+$. If P satisfies (IB) and (IS) where
(IB) $1 \in P$;
(IS) for every $i \in P$, we know that $i+1$ is also in P;
then $P = \mathbb{N}^+$.

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

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

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

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

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

[8] You might try to prove the well-ordering of $\mathbb{N}$ using induction. You would in fact succeed, because well-ordering is equivalent to induction: each implies the other.

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

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

Figure 1.1. Material for Exercise 1.20

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

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

Figure 1.2. Material for Exercise 1.21

Exercises.

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

Exercise 1.20.
(a) Fill in each blank of Figure 1.1 with the justification.
(b) Why would someone writing a proof of the claim think to look at $a - 0$?
(c) Why would that person start with $a + 0$ instead?

Exercise 1.21.
(a) Fill in each blank of Figure 1.2 with the justification.
(b) Why would someone writing a proof of this claim think to look at the values of $a - b$ and $b - a$?
(c) Why is the introduction of S essential, rather than a distraction?

Exercise 1.22. Let $a \in \mathbb{Z}$. Show that:
(a) $a < a + 1$;
(b) if $a \in \mathbb{N}$, then $0 \leq a$; and

Let S be a well-ordered set.

Claim: Every strictly decreasing sequence of elements of S is finite.

Proof:

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

Figure 1.3. Material for Exercise 1.30

(c) if $a \in \mathbb{N}^+$, then $1 \leq a$.

Exercise 1.23. Let $a, b, c \in \mathbb{Z}$.

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

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

(c) Prove that if $a \leq b$ and $b \leq c$, then $a \leq c$.

Exercise 1.24. Let $a, b \in \mathbb{N}$ and assume that $0 < a < b$. Let $d = b - a$. Show that $d < b$.

Exercise 1.25. Let $a, b, c \in \mathbb{Z}$ and assume that $a \leq b$. Prove that

(a) $a + c \leq b + c$;

(b) if $c \in \mathbb{N}^+$, then $a \leq ac$; and

(c) if $c \in \mathbb{N}^+$, then $ac \leq bc$.

Note: You may henceforth assume this for *all* the inequalities given in Definition 1.12.

Exercise 1.26. Let $S \subseteq \mathbb{N}$. We know from the well-ordering property that S has a smallest element. Prove that this smallest element is unique.

Exercise 1.27. Show that $>$ is not a well-ordering of $\mathbb{N}$.

Exercise 1.28. Show that the ordering $<$ of $\mathbb{Z}$ generalizes "naturally" to an ordering $<$ of $\mathbb{Q}$ that is also a linear ordering.

Exercise 1.29. By definition, a function is a relation. Can a function be an equivalence relation?

Exercise 1.30.

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

(b) Why would someone writing a proof of the claim think to write that $a_i < a_{i+1}$?

(c) Why would someone want to look at the smallest element of A?

1.2: The Division Theorem

Before proceeding to algebra, we need one more property of the integers. The last "arithmetic operation" that you know about is division, but this operation is… "interesting".

Theorem 1.31 (The Division Theorem for Integers)**.** Let $n, d \in \mathbb{Z}$ with $d \neq 0$. There exist unique $q \in \mathbb{Z}$ and $r \in \mathbb{Z}$ satisfying (D1) and (D2) where
(D1) $n = qd + r$;
(D2) $0 \leq r < |d|$.

One implication of this theorem is that division *is not an operation on* $\mathbb{Z}$! An operation on $\mathbb{Z}$ is a relation $f : \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$, but the quotient and remainder imply that division is a relation of the form $\div : (\mathbb{Z} \times (\mathbb{Z} \backslash \{0\})) \to \mathbb{Z} \times \mathbb{Z}$. That is not a binary operation *on* $\mathbb{Z}$. We explore this further in a moment, but for now let's turn to a proof of the theorem.

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

Existence of q and r: First we show the existence of q and r that satisfy (D1). Let $S = \{n - qd : q \in \mathbb{Z}\}$ and $M = S \cap \mathbb{N}$. You will show in Exercise 1.42 that M is non-empty. By the well-ordering of $\mathbb{N}$, M has a smallest element; call it r. By definition of S, there exists $q \in \mathbb{Z}$ such that $n - qd = r$. Properties of arithmetic imply that $n = qd + r$.

Does r satisfy (D2)? By way of contradiction, assume that it does not; then $|d| \leq r$. We had assumed that $d \in \mathbb{N}^+$, so Exercises 1.20 and 1.24 implies that $0 \leq r - d < r$. Rewrite property (D1) using properties of arithmetic:

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

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

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

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

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

Moreover, $r \leq r'$ implies that $r' - r \in \mathbb{N}$, so by substitution, $\left(q - q'\right) d \in \mathbb{N}$. Similarly, $0 \leq r \leq r'$ implies that $0 \leq r' - r \leq r'$. By substitution, $0 \leq \left(q - q'\right) d \leq r'$. Since $d \in \mathbb{N}^+$, it must be that $q-q'\in\mathbb{N}$ also (repeated addition of a negative giving a negative), so $0\leq q-q'$. If $0\neq q-q'$, then $1\leq q-q'$. By Exercise 1.25, $d\leq(q-q')\,d$. By Exercise 1.23, we see that $d\leq(q-q')\,d\leq r'<d$. This states that $d<d$, a contradiction. Hence $q-q'=0$, and by substitution, $r-r'=0$.

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

Definition 1.32 (terms associated with division)**.** Let $n,d\in\mathbb{Z}$ and suppose that $q,r\in\mathbb{Z}$ satisfy the Division Theorem. We call n the **dividend**, d the **divisor**, q the **quotient**, and r the **remainder**.

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

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

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

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

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

Remark 1.33. Why do we insist that $b\neq 0$? Basically, it doesn't make sense. The very idea of division means that if $a/b=c$, then $a=bc$. So, let $a/0=c$. In that case, $a=0c$. This is true only if $a=0$, so we can't have $b=0$. On the other hand, this reasoning doesn't apply to $0/0$, so what about allowing that to be in $\mathbb{Q}$? Actually, that offends our notion of an operation! Why? because if we put $0/0\in\mathbb{Q}$, it is not hard to show that both $0/0=1$ and $0/0=2$, which would imply that $1=2$!

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

- the **real numbers** contain a number for every possible measurement of distance along a line.[9]

We now have

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

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

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

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

Exercises.

Exercise 1.34. Identify the quotient and remainder when dividing:
(a) 10 by -5;
(b) -5 by 10;
(c) -10 by -4.

Exercise 1.35. Prove that if $a \in \mathbb{Z}$, $b \in \mathbb{N}^+$, and $a \mid b$, then $a \leq b$.

Exercise 1.36. Show that $a \leq |a|$ for all $a \in \mathbb{Z}$.

Exercise 1.37. Show that divisibility is transitive for the integers; that is, if $a, b, c \in \mathbb{Z}$, $a \mid b$, and $b \mid c$, then $a \mid c$.

Exercise 1.38. Extend the definition of $<$ so that we can order rational numbers. That is, find a criterion on $a, b, c, d \in \mathbb{Z}$ that tells us when $a/b < c/d$.

Definition 1.39. We define lcm, the **least common multiple** of two integers, as

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

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

Exercise 1.40.
(a) Fill in each blank of Figure 1.4 with the justification.
(b) One part of the proof claims that "A similar argument shows that $b \mid r$." State this argument in detail.

Exercise 1.41. Define a relation $\equiv$ on $\mathbb{Q}$, the set of real numbers, in the following way:

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

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

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

(d) Is every $a \in \mathbb{Q}$ an element of some class? Why?

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

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

Figure 1.4. Material for Exercise 1.40

(e) Show that if $S_a \neq S_b$, then $S_a \cap S_b = \emptyset$.

Exercise 1.42.
(a) Fill in each blank of Figure 1.5 with the justification.
(b) Why would someone writing a proof of the claim think to look at $n - qd$?
(c) Why would this person want to find a value of q?

Exercise 1.43. Let X and Yon the lattice $L = \mathbb{Z} \times \mathbb{Z}$. Let's say that addition is performed as with vectors:

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

multiplication is performed by this *very odd* definition:

$$X \cdot Y = (x_1y_1 - x_2y_2, x_1y_2 + x_2y_1),$$

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

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

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

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

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

Figure 1.5. Material for Exercise 1.42

1.3: Monomials and monoids

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

Monomials

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

$$\mathbb{M} = \{x^a : a \in \mathbb{N}\}, \tag{1}$$

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

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

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

$$\mathbb{M}_n = \left\{ \prod_{i=1}^{m} \left(x_1^{a_{i1}} x_2^{a_{i2}} \cdots x_n^{a_{in}} \right) : m, a_{ij} \in \mathbb{N} \right\}. \tag{2}$$

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

Example 1.44. The following are monomials:

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

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

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

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

The following are not monomials:

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

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

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

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

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

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

Lemma 1.45. Let $n \in \mathbb{N}^+$. Both $\mathbb{M}$ and $\mathbb{M}_n$ are closed under multiplication.

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

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

Since addition is closed in $\mathbb{N}$, the expression $a+b$ simplifies to a natural number. Call this number c. By substitution, $tu = x^c$. This has the form of a univariate monomial; compare it with the description of a monomial in equation (1). So, $tu \in \mathbb{M}$. Since we chose t and u to be arbitrary elements of $\mathbb{M}$, and found their product to be an element of $\mathbb{M}$, we conclude that $\mathbb{M}$ is closed under multiplication. □

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

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

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

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

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

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

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

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

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

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

$$tu = \prod_{i=1}^{m_t+m_u} \left(x_1^{c_{i1}} \cdots x_n^{c_{in}} \right) \tag{3}$$

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

In the list of products, the first few products come from t. How many? There are m_t from t.

The rest are from u. We can specify this precisely using a piecewise function:

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

Specifying c_{ij} this way justifies our claim that tu has the form shown in equation (3). That satisfies the requirements of $\mathbb{M}_n$, so we can say that $tu \in \mathbb{M}_n$. Since t and u were chosen arbitrarily from $\mathbb{M}_n$, it is closed under multiplication. □

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

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

Let's look at two more properties.

Lemma 1.46. Let $n \in \mathbb{N}^+$. Multiplication in $\mathbb{M}$ satifies the commutative property. Multiplication in both $\mathbb{M}$ and $\mathbb{M}_n$ satisfies the associative property.

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

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

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

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

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

□

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

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

with the parentheses as indicated, are absolutely crucial, and cannot be omitted from a good proof.[10]

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

Lemma 1.47. Both $\mathbb{M}$ and $\mathbb{M}_n$ have $1 = x^0 = x_1^0 x_2^0 \cdots x_n^0$ as a multiplicative identity.

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

Monoids

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

Definition 1.48. Let M be a set, and $\circ$ an operation on M. We say that the pair $(M, \circ)$ is a **monoid** if it satisfies the following properties:
(closed) for any $x, y \in M$, we have $x \circ y \in M$;
(associative) for any $x, y, z \in M$, we have $(x \circ y) \circ z = x \circ (y \circ z)$; and
(identity) there exists an **identity element** $e \in M$ such that for any $x \in M$, we have $e \circ x = x \circ e = x$.
We may also say that M is a **monoid under** $\circ$.

So far, then, we know the following:

Theorem 1.49. $\mathbb{N}$ is a monoid under addition and multiplication, while $\mathbb{M}$ and $\mathbb{M}_n$ are monoids under multiplication.

Proof. For $\mathbb{N}$, this is part of its definition. For $\mathbb{M}$ and $\mathbb{M}_n$, see Lemmas 1.45, 1.46, and 1.47. □

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

What other mathematical objects are examples of monoids?

Example 1.50. Let $m, n \in \mathbb{N}^+$. You should have seen in linear algebra that the set of matrices with integer entries $\mathbb{Z}^{m \times n}$ satisfies properties that make it a monoid under addition. The set

[10] Of course, a professional mathematician would not even prove these things in a paper, because they are well-known and easy. On the other hand, a good professional mathematician *would* feel compelled to include in a proof steps that include novel and/or difficult information.

of square matrices with integer entries $\mathbb{Z}^{m\times m}$ satisfies properties that make it a monoid under addition *and* multiplication. That said, your professor almost certainly didn't *call* it a monoid at the time.

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

Example 1.51. Let S be a set, and let F_S be the set of all functions mapping S to itself, with the proviso that for any $f \in F_S$, $f(s)$ is defined for every $s \in S$. We can show that F_S is a monoid under composition of functions, since

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

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

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

Here is one useful property that we can prove already. A natural question to ask about monoids is whether the identity of a monoid is unique. It isn't hard to show that it is.

Theorem 1.52. Suppose that M is a monoid, and there exist $e, i \in M$ such that $ex = x$ and $xi = x$ for all $x \in M$. Then $e = i$, so that the identity of a monoid is unique.

"Unique" in mathematics means *exactly one.* To prove uniqueness of an object x, you consider a generic object y that shares all the properties of x, then reason to show that $x = y$. This is not a contradiction, because we didn't assume that $x \neq y$ in the first place; we simply wondered about a generic y. We did the same thing with the Division Theorem (Theorem 1.31 on page 14).

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

$$e = ei.$$

Since e is a left identity, we know that

$$ei = i.$$

By substitution,

$$e = i.$$

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

Exercises.

Exercise 1.53. Is $\mathbb{N}$ a monoid under:
(a) subtraction?
(b) division?
Be sure to explain your answer.

Exercise 1.54. Is $\mathbb{Z}$ a monoid under:
(a) addition?
(b) subtraction?
(c) multiplication?
(d) division?
Be sure to explain your answer.

Exercise 1.55. Consider the set $B = \{F, T\}$ with the operation $\vee$ where

$$\begin{aligned} F \vee F &= F \\ F \vee T &= T \\ T \vee F &= T \\ T \vee T &= T. \end{aligned}$$

This operation is called **Boolean or**.

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

Exercise 1.56. Consider the set $B = \{F, T\}$ with the operation $\oplus$ where

$$\begin{aligned} F \oplus F &= F \\ F \oplus T &= T \\ T \oplus F &= T \\ T \oplus T &= F. \end{aligned}$$

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

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

Exercise 1.57. Suppose multiplication of x and y commutes. Show that multiplication in $\mathbb{M}_n$ is both closed and associative.

Exercise 1.58.
(a) Show that $\mathbb{N}[x]$, the ring of polynomials in one variable with integer coefficients, is a monoid under addition.
(b) Show that $\mathbb{N}[x]$ is also a monoid if the operation is multiplication.
(c) Explain why we can replace $\mathbb{N}$ by $\mathbb{Z}$ and the argument would remain valid. (*Hint:* think about the *structure* of these sets.)

Exercise 1.59. Recall the lattice L from Exercise 1.43.
(a) Show that L is a monoid under the addition defined in that exercise.
(b) Show that L is a monoid under the multiplication defined in that exercise.

Exercise 1.60. Let A be a set of symbols, and L the set of all finite sequences that can be constructed using elements of A. Let $\circ$ represent *concatenation of lists.* For example, $(a, b) \circ (c, d, e, f) = (a, b, c, d, e, f)$. Show that $(L, \circ)$ is a monoid.

Definition 1.61. For any set S, let $P(S)$ denote the set of all subsets of S. We call this the **power set** of S.

Exercise 1.62.
(a) Suppose $S = \{a, b\}$. Compute $P(S)$, and show that it is a monoid under $\cup$ (union).
(b) Let S be *any* set. Show that $P(S)$ is a monoid under $\cup$ (union).

Exercise 1.63.
(a) Suppose $S = \{a, b\}$. Compute $P(S)$, and show that it is a monoid under $\cap$ (intersection).
(b) Let S be *any* set. Show that $P(S)$ is a monoid under $\cap$ (intersection).

Exercise 1.64.
(a) Fill in each blank of Figure 1.6 with the justification.
(b) Is $(\mathbb{N}, \operatorname{lcm})$ also a monoid? If so, do we have to change anything about the proof? If not, which property fails?

Exercise 1.65. Recall the usual ordering $<$ on $\mathbb{M}$: $x^a < x^b$ if $a < b$. Show that this is a well-ordering.

Remark 1.66. While we can define a well-ordering on $\mathbb{M}_n$, it is a much more complicated proposition, which we take up in Section 11.2.

Exercise 1.67. In Exercise 1.37, you showed that divisibility is transitive in the integers.
(a) Show that divisibility is transitive in *any* monoid; that is, if M is a monoid, $a, b, c \in M$, $a \mid b$, and $b \mid c$, then $a \mid c$.
(b) In fact, you don't need all the properties of a monoid for divisibility to be transitive! Which properties *do* you need?

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

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

Figure 1.6. Material for Exercise 1.64

1.4: Isomorphism

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

Imagine two offices. How would you decide if the offices were equally suitable for a certain job? First, you would need to know what tasks have to be completed, and what materials you need for those tasks. For example, if your job required you to keep books for reference, you would want to find a bookshelf in the office. If it required you to write, you would need a desk, and perhaps a computer. If it required you to communicate with people in other locations, you might need a phone. Having made such a list, you would then want to compare the two offices. If they both had the equipment you needed, you'd think they were both suitable for the job at hand. It wouldn't really matter how the offices satisfied the requirements; if one had a desk by the window, and the other had it on the side opposite the window, that would be okay. If one office lacked a desk, however, it wouldn't be up to the required job.

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

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

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

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

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

Substituting xy for z suggests that we want the property

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

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

Definition 1.68. Let $(M, \times)$ and $(N, +)$ be monoids. If there exists a function $f : M \longrightarrow N$ such that

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

and

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

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

You may not remember the definitions of one-to-one and onto, or you may not understand how to prove them, so here is a precise definition, for reference.

Definition 1.69. Let $f : S \to U$ be a mapping of sets.

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

Another way of saying that a function $f : S \to U$ is onto is to say that $f(S) = U$; that is, the image of S is *all* of U, or that *every* element of U corresponds via f to some element of S.

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

Example 1.70. We claim that $(\mathbb{M}, \times)$ is isomorphic to $(\mathbb{N}, +)$. To see why, let $f : \mathbb{M} \longrightarrow \mathbb{N}$ by

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

First we show that f is a bijection.

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

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

So f is a bijection. Is it also an isomorphism? First we check that f preserves the operation. Let $t, u \in \mathbb{M}$.[11] By definition of $\mathbb{M}$, $t = x^a$ and $u = x^b$ for $a, b \in \mathbb{N}$. We now manipulate $f(tu)$ using definitions and substitutions to show that the operation is preserved:

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

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

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

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

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

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

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

Example 1.71. Suppose there *does* exist an isomorphism $f : (\mathbb{N}, +) \to (\mathbb{N}, \times)$. What would have to be true about f? Let $a \in \mathbb{N}$ such that $f(1) = a$; after all, f has to map 1 to *something!* An isomorphism must preserve the operation, so

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

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

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

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

Exercises.

Exercise 1.72. Show that the monoids "Boolean or" and "Boolean xor" from Exercises 1.55 and 1.56 are *not* isomorphic.

Exercise 1.73. Let $(M, \times)$, $(N, +)$, and $(P, \sqcap)$ be monoids.

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

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

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

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

1.5: Direct products

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

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

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

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

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

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

We can do this with other sets, as well.

Definition 1.74. Let $r\in\mathbb{N}^+$ and $S_1,S_2,\ldots,S_r$ be sets. The **Cartesian product** of S_1, …, S_r is the set of all lists of r elements where the ith entry is an element of S_i; that is,

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

Example 1.75. We already mentioned a Cartesian product of two sets in the introduction to this chapter. Another example would be $\mathbb{N}\times\mathbb{M}$; elements of $\mathbb{N}\times\mathbb{M}$ include $\left(2,x^3\right)$ and $\left(0,x^5\right)$. In general, $\mathbb{N}\times\mathbb{M}$ is the set of all ordered pairs where the first entry is a natural number, and the second is a monomial.

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

Definition 1.76. Let $r\in\mathbb{N}^+$ and $M_1,M_2,\ldots,M_r$ be monoids. The **direct product** of M_1, …, M_r is the pair

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

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

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

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

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

$$x\otimes y=(x_1y_1,x_2y_2,\ldots,x_ry_r)$$

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

Example 1.77. Recall that $\mathbb{N}\times\mathbb{M}$ is a Cartesian product; if we consider the monoids $(\mathbb{N},\times)$ and $(\mathbb{M},\times)$, we can show that the direct product is a monoid, much like $\mathbb{N}$ and $\mathbb{M}$! To see why, we check each of the properties.

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

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

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

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

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

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

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

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

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

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

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

Interestingly, if we had used $(\mathbb{N}, +)$ *instead* of $(\mathbb{N}, \times)$ in the previous example, we *still* would have obtained a direct product! Indeed, the direct product of monoids is *always* a monoid!

Theorem 1.78. The direct product of monoids $M_1, \ldots, M_r$ is itself a monoid. Its identity element is $(e_1, e_2, \ldots, e_r)$, where each e_i denotes the identity of the corresponding monoid M_i.

Proof. You do it! See Exercise 1.81. □

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

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

That leads to the following result.

Theorem 1.79. The variables of $\mathbb{M}_n$ are commutative if and only if $\mathbb{M}_n \cong \mathbb{M}^n$.

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

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

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

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

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

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

Conversely, suppose $\mathbb{M}_n \cong \mathbb{M}^n$. By Exercise 1.73, $\mathbb{M}^n \cong \mathbb{M}_n$. By definition, there exists a bijection $f : \mathbb{M}^n \longrightarrow \mathbb{M}_n$ satisfying Definition 1.68. Let $t, u \in \mathbb{M}^n$; by definition, we can find $a_i, b_j \in \mathbb{N}$ such that $t = x_1^{a_1} \cdots x_n^{a_n}$ and $u = x_1^{b_1} \cdots x_n^{b_n}$. Since f preserves the operation, $f(tu) = f(t) \otimes f(u)$. Now, $f(t)$ and $f(u)$ are elements of $\mathbb{M}^n$, which is commutative by Exercise 1.82 (with the $S_i = \mathbb{M}$ here). Hence $f(t) \otimes f(u) = f(u) \otimes f(t)$, so that $f(tu) = f(u) \otimes f(t)$. Using the fact that f preserves the operation again, only in reverse, we see that $f(tu) = f(ut)$. Recall that f, as a bijection, is one-to-one! Thus $tu = ut$, and $\mathbb{M}^n$ is commutative. □

Notation 1.80. Although we used $\otimes$ in this section to denote the operation in a direct product, this is not standard; I was trying to emphasize that the product is different for the direct product than for the monoids that created it. In general, the product $x \otimes y$ is written simply as xy. Thus, the last line of the proof above would have $f(t) f(u)$ instead of $f(t) \otimes f(u)$.

Exercises.

Exercise 1.81. Prove Theorem 1.78. Use Example 1.77 as a guide.

Exercise 1.82. Suppose M_1, M_2, ..., and M_n are *commutative* monoids. Show that the direct product $M_1 \times M_2 \times \cdots \times M_n$ is also a commutative monoid.

Exercise 1.83. Show that $\mathbb{M}^n \cong \mathbb{N}^n$. What does this imply about $\mathbb{M}_n$ and $\mathbb{N}^n$?

Exercise 1.84. Recall the lattice L from Exercise 1.43. Exercise 1.59 shows that this is both a monoid under addition and a monoid under multiplication, as defined in that exercise. Is either monoid isomorphic to $\mathbb{N}^2$?

Exercise 1.85. Let $\mathbb{T}_S^n$ denote the set of terms in n variables whose coefficients are elements of the set S. For example, $2xy \in \mathbb{T}_{\mathbb{Z}}^2$ and $\pi x^3 \in \mathbb{T}_{\mathbb{R}}^1$.

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

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

Exercise 1.86. We define the **kernel** of a monoid homomorphism $\varphi : M \to N$ as

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

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

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

(b) Fill in each blank of Figure 1.7 with the justification.

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

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

Show that

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

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

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

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

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

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

Obviously, we'd rather that not happen, so

(iii) Fill in each blank of Figure 1.8 with the justification.

Once you've shown that the operation is well defined, show that

(iv) M/K is a monoid with this operation.

This means that we can use monoid morphisms to create new monoids.

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

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

Figure 1.7. Material for Exercise 1.86(b)

1.6: Absorption and the Ascending Chain Condition

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

Absorption

Definition 1.87. Let M be a monoid, and $A \subseteq M$. If $ma \in A$ for every $m \in M$ and $a \in A$, then A **absorbs from** M. We also say that A is an **absorbing subset**, or that satisfies the **absorption property**.

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

Example 1.88. Write $2\mathbb{Z}$ for the set of even integers. By definition, $2\mathbb{Z} \subsetneq \mathbb{Z}$. Notice that $2\mathbb{Z}$ is *not* a monoid, since $1 \notin 2\mathbb{Z}$. On the other hand, any $a \in 2\mathbb{Z}$ has the form $a = 2z$ for some $z \in \mathbb{Z}$. Thus, for any $m \in \mathbb{Z}$, we have

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

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

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

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

Figure 1.8. Material for Exercise 1.86

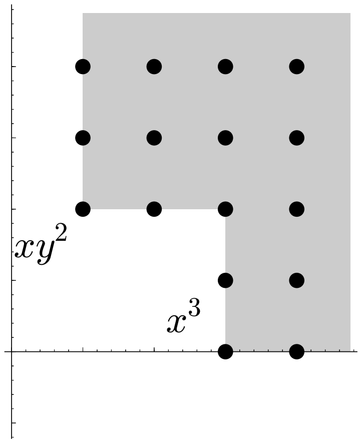
The set of integer multiples of an integer is important enough that it inspires notation.

Notation 1.89. We write $d\mathbb{Z}$ for the set of integer multiples of d.

So $2\mathbb{Z} = \{\ldots, -2, 0, 2, 4, \ldots\}$ is the set of integer multiples of 2; $5\mathbb{Z}$ is the set of integer multiples of 5; and so forth. You will show in Exercise 1.99 that $d\mathbb{Z}$ absorbs multiplication from $\mathbb{Z}$, but *not* addition.

The monomials provide another important example of absorption.

Example 1.90. Let A be an absorbing subset of $\mathbb{M}_2$. Suppose that $xy^2, x^3 \in A$, but none of their factors is in A. Since A absorbs from $\mathbb{M}_2$, all the monomial multiples of xy^2 and x^3 are also in A. We can illustrate this with a **monomial diagram**:



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

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

Definition 1.91. Suppose A is an absorbing subset of $\mathbb{M}_n$, and $t \in A$. If no other $u \in A$ divides t, then we call t a **generator** of A.

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

This leads us to a remarkable result.

Dickson's Lemma and the Ascending Chain Condition

Theorem 1.92 (Dickson's Lemma)**.** Every absorbing subset of $\mathbb{M}_n$ has a finite number of generators.

(Actually, Dickson proved a similar result for a similar set, but is more or less the same.) The proof is a little complicated, so we'll illustrate it using some monomial diagrams. In Figure 1.9(A), we see an absorbing subset A. (The same as you saw before.) Essentially, the argument *projects* A down one dimension, as in Figure 1.9(B). In this smaller dimension, an argument by induction allows us to choose a finite number of generators, which correspond to elements of A, illustrated in Figure 1.9(C). These corresponding elements of A are always generators of A, but they might not be *all* the generators of A, shown in Figure 1.9(C) by the red circle. In that case, we take the remaining generators of A, use them to construct a new absorbing subset, and project again to obtain new generators, as in Figure 1.9(D). The thing to notice is that, in Figures 1.9(C) and 1.9(D), the y-values of the new generators decrease with each projection. This cannot continue indefinitely, since $\mathbb{N}$ is well-ordered, and we are done.

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

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

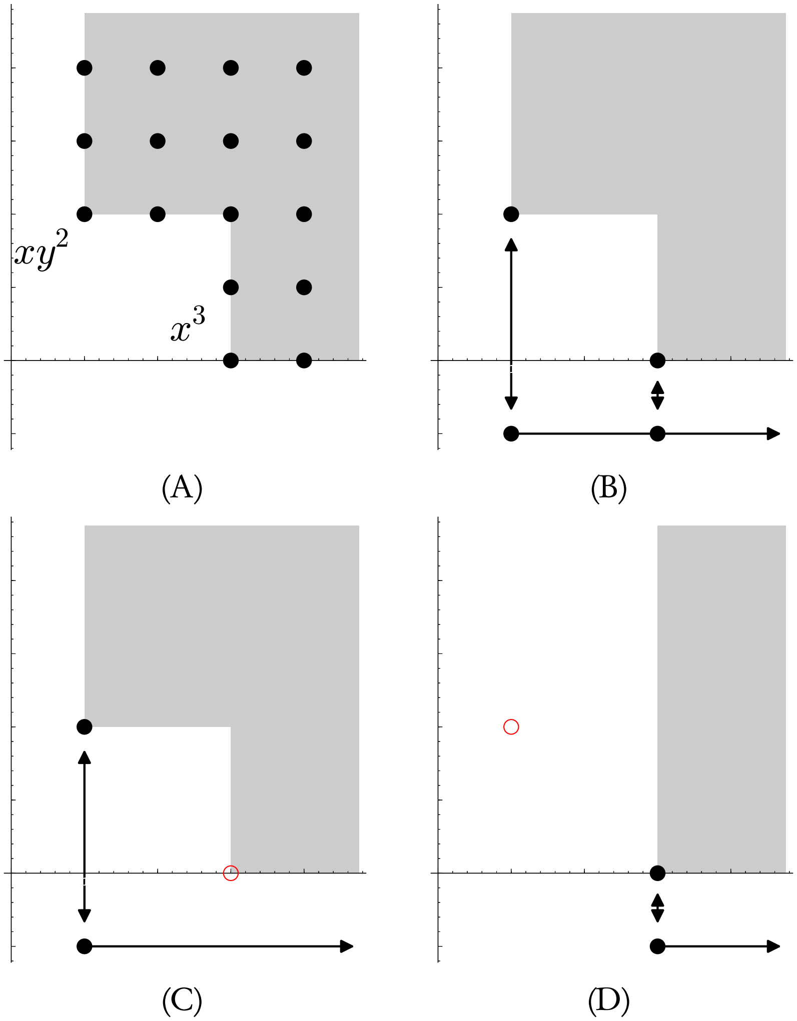


Figure 1.9. Illustration of the proof of Dickson's Lemma.

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

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

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

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

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

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

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

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

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

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

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

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

Definition 1.93. Let M be a monoid. Suppose that, for any ideals A_1, A_2, … of M, we can guarantee that if $A_1 \subseteq A_2 \subseteq \cdots$, then there is some $n \in \mathbb{N}^+$ such that $A_n = A_{n+1} = \cdots$. In this case, we say that M satisfies the **ascending chain condition**,, or that M is **Noetherian**.

A look back at the Hilbert-Dickson game

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

Definition 1.94. Let A be an absorbing subset of $\mathbb{M}_n$. The **Hilbert Function** $H_A(d)$ counts the number of monomials of total degree d and *not* in A. The **Affine Hilbert Function** $H_A^{\text{aff}}(d)$ is the sum of the Hilbert Function for degree no more than d; that is, $H_A^{\text{aff}}(d) = \sum_{i=0}^{d} H_A(d)$.

Example 1.95. In the diagram of Example 1.90, $H(0) = 1$, $H(1) = 2$, $H(2) = 3$, $H(3) = 2$, and $H(d) = 1$ for all $d \geq 4$. On the other hand, $H^{\text{aff}}(4) = 9$.

The following result is immediate.

Theorem 1.96. Suppose that A is the absorbing subset generated by the moves chosen in a Hilbert-Dickson game, and let $d \in \mathbb{N}$. The number of moves (a, b) possible in a Hilbert-Dickson game with $a + b \leq d$ is $H_A^{\text{aff}}(d)$.

Corollary 1.97. Every Hilbert-Dickson game must end in a finite number of moves.

Proof. Every ith move in a Hilbert-Dickson game corresponds to the creation of a new absorbing subset A_i of $\mathbb{M}_2$. Let A be the union of these A_i; you will show in Exercise 1.100 that A also absorbs from $\mathbb{M}_2$. By Dickson's Lemma, A has finitely many generators; call them $t_1, \ldots, t_m$. Each t_j appears in A, and the definition of union means that each t_j must appear in some A_{i_j}. Let k be the largest such i_j; that is, $k = \max\{i_1, \ldots, i_m\}$. Practically speaking, "largest" means "last chosen", so each t_i has been chosen at this point. Another way of saying this in symbols is that $t_1, \ldots, t_m \in \bigcup_{i=1}^{k} A_i$. All the generators of A are in this union, so no element of A can be absent! So $A = \bigcup_{i=1}^{k} A_i$; in other words, the ideal is generated after finitely many moves. □

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

Exercises.

Exercise 1.98. Is $2\mathbb{Z}$ an absorbing subset of $\mathbb{Z}$ under addition? Why or why not?

Exercise 1.99. Let $d \in \mathbb{Z}$ and $A = d\mathbb{Z}$. Show that A is an absorbing subset of $\mathbb{Z}$.

Exercise 1.100. Fill in each blank of Figure 1.10 with its justification.

Exercise 1.101. Let L be the lattice defined in Exercise 1.43. Exercise 1.59 shows that L is a monoid under its strange multiplication. Let $P = (3, 1)$ and A be the absorbing subset generated by P. Sketch L and P, distinguishing the elements of P from those of L using different colors, or an X, or some similar distinguishing mark.

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

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

Figure 1.10. Material for Exercise 1.100

Part II

Groups

Chapter 2:
Groups

In Chapter 1, we described monoids. In this chapter, we study a *group,* which is a special kind of monoid. Groups are special in that every element in the group has an *inverse element.*

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

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

2.1: Groups

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

Precise definition, first examples

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Order of a group, Cayley tables

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

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

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

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

Example 2.7. The sets

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

are all groups under multiplication:

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

I will now make an extraordinary claim:

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

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

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

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

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

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

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

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

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

We did *not* say that e represents the *only* identity. For all we know, a might also be an identity; is that possible? In fact, it is not possible; why? Remember that a group is a monoid. We showed in Proposition 2.12 that the identity of a monoid is unique; thus, the identity of a group is unique; thus, there can be only one identity, e.

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

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

So the Cayley table of our group looks like:

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

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

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

Other elementary properties of groups

Notation 2.10. We adopt the following convention:

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

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

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

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

Proof. You prove it! See Exercise 2.15. □

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

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

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

$$xy = e.$$

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

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

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

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

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

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

$$xy = e.$$

Since z is an inverse of x,

$$xz = e.$$

By substitution,

$$xy = xz.$$

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

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

By the associative property,

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

and by the inverse property,

$$ey = ez.$$

Since e is the identity of G,

$$y = z.$$

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

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

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

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

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

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

Let G be a group, and $x \in G$.

Claim: $\left(x^{-1}\right)^{-1} = x$; or, if the operation is addition, $-(-x) = x$.

Proof:

1. By _____, $x \cdot x^{-1} = e$ and $x^{-1} \cdot x = e$.
2. By _____, $\left(x^{-1}\right)^{-1} = x$.
3. Negative are merely how we express opposites when the operation is addition, so $-(-x) = x$.

Figure 2.1. Material for Exercise 2.15

Since $a = a$, substitution implies that $bc = bd$. Thus

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

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

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

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

$$(g_1, \ldots, g_n) \otimes (h_1, \ldots, h_n) = (g_1h_1, \ldots, g_nh_n),$$

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

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

Exercises.

Exercise 2.15.

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

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

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

Exercise 2.17. Is $\left(\mathbb{N}^+, \operatorname{lcm}\right)$ a group? (See Exercise 1.64.)

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

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

Exercise 2.20. Recall the Boolean-or monoid $(B, \vee)$ from Exercise 1.55. Is it a group? If so, is it abelian? Explain how it justifies each property. If not, explain why not.

Exercise 2.21. Recall the Boolean-xor monoid $(B, \oplus)$ from Exercise 1.56. Is it a group? If so, is it abelian? Explain how it justifies each property. If not, explain why not.

Exercise 2.22. In Section 1.3, we showed that F_S, the set of all functions, is a monoid for any S.

(a) Show that $F_{\mathbb{R}}$, the set of all functions on the real numbers $\mathbb{R}$, is *not* a group.

(b) Describe a subset of $F_{\mathbb{R}}$ that *is* a group. Another way of looking at this question is: what restriction would you have to impose on any function $f \in F_S$ to fix the problem you found in part (a)?

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

Exercise 2.24. Show that the Cartesian product

$$\mathbb{Z} \times \mathbb{Z} := \{(a, b) : a, b \in \mathbb{Z}\}$$

is a group under the direct product's notion of addition; that is,

$$x + y = (a + c, b + d).$$

Exercise 2.25. Let $(G, \circ)$ and $(H, *)$ be groups, and define

$$G \times H = \{(a, b) : a \in G,\ b \in H\}.$$

Define an operation † on $G \times H$ in the following way. For any $x, y \in G \times H$, write $x = (a, b)$ and $y = (c, d)$; we say that

$$x \dagger y = (a \circ c, b * d).$$

(a) Show that $(G \times H, \dagger)$ is a group.

(b) Show that if G and H are both abelian, then so is $G \times H$.

Exercise 2.26. Let $n \in \mathbb{N}^+$. Let G_1, G_2, …, G_n be groups, and consider

$$\begin{aligned}\prod_{i=1}^{n} G_i &= G_1 \times G_2 \times \cdots \times G_n \\ &= \{(a_1, a_2, \ldots, a_n) : a_i \in G_i\ \forall i = 1, 2, \ldots, n\}\end{aligned}$$

with the operation † where if $x = (a_1, a_2, \ldots, a_n)$ and $y = (b_1, b_2, \ldots, b_n)$, then

$$x \dagger y = (a_1 b_1, a_2 b_2, \ldots, a_n b_n),$$

Claim: Any two elements a, b of any group G satisfy $(ab)^{-1} = b^{-1}a^{-1}$.
Proof:

1. Let _____.
2. By the _____, _____, and _____ properties of groups,

$$(ab)\,b^{-1}a^{-1} = a\left(b \cdot b^{-1}\right)a^{-1} = aea^{-1} = aa^{-1} = e.$$

3. We chose _____ arbitrarily, so this holds for all elements of all groups, as claimed.

Figure 2.2. Material for Exercise 2.34

where each product $a_i b_i$ is performed according to the operation of the group G_i. Show that $\prod_{i=1}^{n} G_i$ is a group, and notice that this shows that the direct product of groups is a group, as claimed above. (We used ⊗ instead of † there, though.)

Exercise 2.27. Let $m \in \mathbb{N}^+$.
(a) Show in detail that $\mathbb{R}^{m\times m}$ is a group under addition.
(b) Show by counterexample that $\mathbb{R}^{m\times m}$ is *not* a group under multiplication.

Exercise 2.28. Let $m \in \mathbb{N}^+$. Explain why $\mathrm{GL}_m(\mathbb{R})$ satisfies the identity and inverse properties of a group.

Exercise 2.29. Let $\mathbb{R}^+ = \{x \in \mathbb{R} : x > 0\}$, and $\times$ the ordinary multiplication of real numbers. Show that $\left(\mathbb{R}^+, \times\right)$ is a group.

Exercise 2.30. Define $\mathbb{Q}^*$ to be the set of non-zero rational numbers; that is,

$$\mathbb{Q}^* = \left\{ \frac{a}{b} : a, b \in \mathbb{Z} \text{ where } a \neq 0 \text{ and } b \neq 0 \right\}.$$

Show that $\mathbb{Q}^*$ is a multiplicative group.

Exercise 2.31. Show that every group of order 3 has the same structure.

Exercise 2.32. *Not* every group of order 4 has the same structure, because there are two Cayley tables with different structures. One of these groups is the **Klein four-group**, where each element is its own inverse; the other is called a **cyclic group** of order 4, where not every element is its own inverse. Determine the Cayley tables for each group.

Exercise 2.33. Let G be a group, and $x, y \in G$. Show that $xy^{-1} \in G$.

Exercise 2.34.

(a) Let $m \in \mathbb{N}^+$ and $G = \mathrm{GL}_m(\mathbb{R})$. Show that there exist $a, b \in G$ such that $(ab)^{-1} \neq a^{-1}b^{-1}$.
(b) Suppose that H is an arbitrary group.
  (i) Explain why we cannot assume that for every $a, b \in H$, $(ab)^{-1} = a^{-1}b^{-1}$.
  (ii) Fill in the blanks of Figure 2.2 with the appropriate justification or statement.

Exercise 2.35. Let $\circ$ denote the ordinary composition of functions, and consider the following functions that map any point $P = (x, y) \in \mathbb{R}^2$ to another point in $\mathbb{R}^2$:

$$\begin{aligned} I(P) &= P, \\ F(P) &= (y, x), \\ X(P) &= (-x, y), \\ Y(P) &= (x, -y). \end{aligned}$$

(a) Let $P = (2, 3)$. Label the points P, $I(P)$, $F(P)$, $X(P)$, $Y(P)$, $(F \circ X)(P)$, $(X \circ Y)(P)$, and $(F \circ F)(P)$ on an x-y axis. (Some of these may result in the same point; if so, label the point twice.)

(b) Show that $F \circ F = X \circ X = Y \circ Y = I$.

(c) Show that $G = \{I, F, X, Y\}$ is *not* a group.

(d) Find the smallest group $\overline{G}$ such that $G \subset \overline{G}$. While you're at it, construct the Cayley table for $\overline{G}$.

(e) Is $\overline{G}$ abelian?

Exercise 2.36. Let i be a number such that $i^2 = -1$, and let Q_8 be the set of **quaternions**, defined by the matrices $\{\pm\mathbf{1}, \pm\mathbf{i}, \pm\mathbf{j}, \pm\mathbf{k}\}$ where

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

(a) Show that $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -\mathbf{1}$.

(b) Show that $\mathbf{ij} = \mathbf{k}$, $\mathbf{jk} = \mathbf{i}$, and $\mathbf{ik} = -\mathbf{j}$.

(c) Use (a) and (b) to build the Cayley table of Q_8. (In this case, the Cayley table is the multiplication table.)

(c) Show that Q_8 is a group under matrix multiplication.

(d) Explain why Q_8 is not an abelian group.

Definition 2.37. Let G be any group.

1. For all $x, y \in G$, define the **commutator of** x **and** y to be $x^{-1}y^{-1}xy$. We write $[x, y]$ for the commutator of x and y.
2. For all $z, g \in G$, define the **conjugation of** g **by** z to be zgz^{-1}. We write g^z for the conjugation of g by z.

Exercise 2.38. (a) Explain why $[x, y] = e$ iff x and y commute.

(b) Show that $[x, y]^{-1} = [y, x]$; that is, the inverse of $[x, y]$ is $[y, x]$.

(c) Show that $(g^z)^{-1} = (g^{-1})^z$; that is, the inverse of conjugation of g by z is the conjugation of the inverse of g by z.

(d) Fill in each blank of Figure 2.3 with the appropriate justification or statement.

2.2: The symmetries of a triangle

Claim: $[x,y]^z = [x^z, y^z]$ for all $x,y,z \in G$.

Proof:

1. Let _____.
2. By _____, $[x^z, y^z] = \left[zxz^{-1}, zyz^{-1}\right]$.
3. By _____, $\left[zxz^{-1}, zyz^{-1}\right] = \left(zxz^{-1}\right)^{-1}\left(zyz^{-1}\right)^{-1}\left(zxz^{-1}\right)\left(zyz^{-1}\right)$.
4. By Exercise _____,

$$\begin{aligned}\left(zxz^{-1}\right)^{-1}\left(zyz^{-1}\right)^{-1}\left(zxz^{-1}\right)\left(zyz^{-1}\right) &= \\ = \left(zx^{-1}z^{-1}\right)\left(zy^{-1}z^{-1}\right)\left(zxz^{-1}\right)\left(zyz^{-1}\right)&.\end{aligned}$$

5. By _____,

$$\begin{aligned}\left(zx^{-1}z^{-1}\right)\left(zy^{-1}z^{-1}\right)\left(zxz^{-1}\right)\left(zyz^{-1}\right) &= \\ \left(zx^{-1}\right)\left(z^{-1}z\right)y^{-1}\left(z^{-1}z\right)x\left(z^{-1}z\right)\left(yz^{-1}\right)&.\end{aligned}$$

6. By _____,

$$\begin{aligned}\left(zx^{-1}\right)\left(z^{-1}z\right)y^{-1}\left(z^{-1}z\right)x\left(z^{-1}z\right)\left(yz^{-1}\right) &= \\ = \left(zx^{-1}\right)ey^{-1}exe\left(yz^{-1}\right)&.\end{aligned}$$

7. By _____, $\left(zx^{-1}\right)ey^{-1}exe\left(yz^{-1}\right) = \left(zx^{-1}\right)y^{-1}x\left(yz^{-1}\right)$.
8. By _____, $\left(zx^{-1}\right)y^{-1}x\left(yz^{-1}\right) = z\left(x^{-1}y^{-1}xy\right)z^{-1}$.
9. By _____, $z\left(x^{-1}y^{-1}xy\right)z^{-1} = z[x,y]z^{-1}$.
10. By _____, $z[x,y]z^{-1} = [x,y]^z$.
11. By _____, $[x^z, y^z] = [x,y]^z$.

Figure 2.3. Material for Exercise 2.38(c)

In this section, we show that the symmetries of an equilateral triangle form a group. We call this group D_3. This group *is not abelian.* You already know that groups of order 2, 3, and 4 are abelian; in Section 3.3 you will learn why a group of order 5 must also be abelian. Thus, D_3 is the smallest non-abelian group.

Intuitive development of D_3

To describe D_3, start with an equilateral triangle in $\mathbb{R}^2$, with its center at the origin. We want to look at its group of symmetries. Intuitively, a "symmetry" is a transformation of the plane that leaves the *triangle* in the same location, even if its *points* are in different locations. "Transformations" include actions like rotation, reflection (flip), and translation (shift). Translating the plane in some direction certainly won't leave the triangle intact, but rotation and reflection can. Two obvious symmetries of an equilateral triangle are a 120° rotation through the origin, and a reflection through the y-axis. We'll call the first of these ρ, and the second φ. See Figure 2.4.

It is helpful to observe two important properties.

Theorem 2.39. If φ and ρ are as specified, then $\varphi\rho = \rho^2\varphi$.

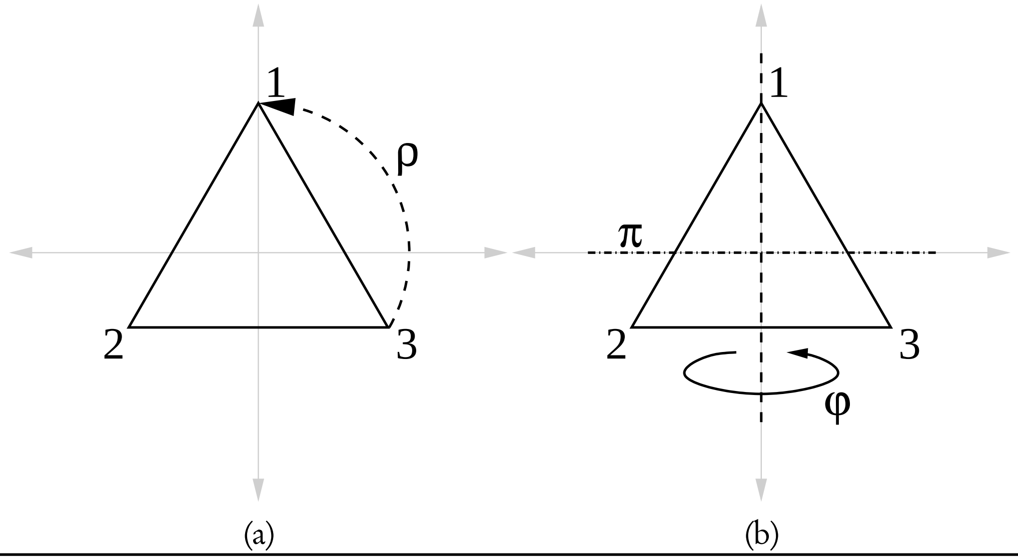


Figure 2.4. Rotation and reflection of the triangle

For now, we consider intuitive proofs only. Detailed proofs appear later in the section.

Intuitive proof. The expression $\varphi\rho$ means to apply ρ first, then φ. It'll help if you sketch what takes place here. Rotating 120° moves vertex 1 to vertex 2, vertex 2 to vertex 3, and vertex 3 to vertex 1. Flipping through the y-axis leaves the top vertex in place; since we performed the rotation first, the top vertex is now vertex 3, so vertices 1 and 2 are the ones swapped. Thus, vertex 1 has moved to vertex 3, vertex 3 has moved to vertex 1, and vertex 2 is in its original location.

On the other hand, $\rho^2\varphi$ means to apply φ first, then apply ρ twice. Again, it will help to sketch what follows. Flipping through the y-axis swaps vertices 2 and 3, leaving vertex 1 in the same place. Rotating twice then moves vertex 1 to the lower right position, vertex 3 to the top position, and vertex 2 to the lower left position. This is the same arrangement of the vertices as we had for $\varphi\rho$, which means that $\varphi\rho = \rho^2\varphi$. □

You might notice that there's a gap in our reasoning: we showed that the *vertices* of the triangle ended up in the same place, but not the *points in between*. That requires a little more work, which is why we provide detailed proofs later.

By the way, did you notice something interesting about Corollary 2.39? It implies that the operation in D_3 is non-commutative! We have $\varphi\rho = \rho^2\varphi$, and a little logic shows that $\rho^2\varphi \neq \rho\varphi$: thus $\varphi\rho \neq \rho\varphi$. After all, $\rho\varphi$

Another "obvious" symmetry of the triangle is the transformation where you *do nothing* – or, if you prefer, where you effectively *move every point back to itself,* as in a 360° rotation, say. We'll call this symmetry ι. It gives us the last property we need to specify the group, D_3.

Theorem 2.40. In D_3, $\rho^3 = \varphi^2 = \iota$.

Intuitive proof. Rotating 120° three times is the same as rotating 360°, which is the same as not rotating at all! Likewise, φ moves any point (x,y) to $(x,-y)$, and applying φ again moves $(x,-y)$ back to (x,y), which is the same as not flipping at all!

We are now ready to specify D_3. □

Definition 2.41. Let $D_3 = \{\iota, \varphi, \rho, \rho^2, \rho\varphi, \rho^2\varphi\}$.

Theorem 2.42. D_3 is a group under composition of functions.

Proof. To prove this, we will show that all the properties of a group are satisfied. We will start the proof, and leave you to finish it in Exercise 2.46.

Closure: In Exercise 2.46, you will compute the Cayley table of D_3. There, you will see that every composition is also an element of D_3.

Associative: Way back in Section 1.3, we showed that F_S, the set of functions over a set S, was a monoid under composition for *any* set S. To do that, we had to show that composition of functions was associative. There's no point in repeating that proof here; doing it once is good enough for a sane person. Symmetries are functions; after all, they map any point in $\mathbb{R}^2$ to another point in $\mathbb{R}^2$, with no ambiguity about where the point goes. So, we've already proved this.

Identity: We claim that ι is the identity function. To see this, let $\sigma \in D_3$ be any symmetry; we need to show that $\iota\sigma = \iota$ and $\sigma\iota = \sigma$. For the first, apply σ to the triangle. Then apply ι. Since ι effectively leaves everything in place, all the points are in the same place they were after we applied σ. In other words, $\iota\sigma = \sigma$. The proof that $\sigma\iota = \sigma$ is similar.

Alternately, you could look at the result of Exercise 2.46; you will find that $\iota\sigma = \sigma\iota = \sigma$ for every $\sigma \in D_3$.

Inverse: Intuitively, rotation and reflection are one-to-one-functions: after all, if a point P is mapped to a point R by either, it doesn't make sense that another point Q would also be mapped to R. Since one-to-one functions have inverses, every element σ of D_3 must have an inverse function σ^{-1}, which undoes whatever σ did. But is $\sigma^{-1} \in D_3$, also? Since σ maps every point of the triangle onto the triangle, σ^{-1} will undo that map: every point of the triangle will be mapped back onto itself, as well. So, yes, $\sigma^{-1} \in D_3$.

Here, the intuition is a little too imprecise; it isn't *that* obvious that rotation is a one-to-one function. Fortunately, the result of Exercise 2.46 shows that ι, the identity, appears in every row and column. That means that every element has an inverse. □

Detailed proof that D_3 contains all symmetries of the triangle

To prove that D_3 contains *all* symmetries of the triangle, we need to make some notions more precise. First, what is a symmetry? A **symmetry** of *any* polygon is a distance-preserving function on $\mathbb{R}^2$ that maps points of the polygon back onto itself. Notice the careful wording: the *points* of the polygon can change places, but since they have to be mapped back onto the polygon, the polygon itself has to remain in the same place.

Let's look at the specifics for our triangle. What functions are symmetries of the triangle? To answer this question, we divide it into two parts.

1. What are the distance-preserving functions that map $\mathbb{R}^2$ to itself, and leave the origin undisturbed? Here, distance is measured by the usual metric,

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

(You might wonder why we don't want the origin to move. Basically, if a function α preserves both distances between points and a figure centered at the origin, then the origin *cannot* move, since then its distance to points on the figure would change.)

2. Not all of the functions identitifed by question (1) map points on the triangle back onto the triangle; for example a 45° degree rotation does not. Which ones do?

Lemma 2.43 answers the first question.

Lemma 2.43. Let $\alpha : \mathbb{R}^2 \to \mathbb{R}^2$. If

- α does not move the origin; that is, $\alpha(0,0) = (0,0)$, and
- the distance between $\alpha(P)$ and $\alpha(R)$ is the same as the distance between P and R for every $P, R \in \mathbb{R}^2$,

then α has one of the following two forms:

$$\rho = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \quad \exists t \in \mathbb{R}$$

or

$$\varphi = \begin{pmatrix} \cos t & \sin t \\ \sin t & -\cos t \end{pmatrix} \quad \exists t \in \mathbb{R}.$$

The two values of t may be different.

Proof. Assume that $\alpha(0,0) = (0,0)$ and for every $P, R \in \mathbb{R}^2$ the distance between $\alpha(P)$ and $\alpha(R)$ is the same as the distance between P and R. We can determine α precisely merely from how it acts on two points in the plane!

First, let $P = (1,0)$. Write $\alpha(P) = Q = (q_1, q_2)$; this is the point where α moves Q. The distance between P and the origin is 1. Since $\alpha(0,0) = (0,0)$, the distance between Q and the origin is $\sqrt{q_1^2 + q_2^2}$. Because α preserves distance,

$$1 = \sqrt{q_1^2 + q_2^2},$$

or

$$q_1^2 + q_2^2 = 1.$$

The only values for Q that satisfy this equation are those points that lie on the circle whose center is the origin. Any point on this circle can be parametrized as

$$(\cos t, \sin t)$$

where $t \in [0, 2\pi)$ represents an angle. Hence, $\alpha(P) = (\cos t, \sin t)$.

Let $R = (0,1)$. Write $\alpha(R) = S = (s_1, s_2)$. An argument similar to the one above shows that S also lies on the circle whose center is the origin. Moreover, the distance between P and R is $\sqrt{2}$, so the distance between Q and S is also $\sqrt{2}$. That is,

$$\sqrt{(\cos t - s_1)^2 + (\sin t - s_2)^2} = \sqrt{2},$$

or

$$(\cos t - s_1)^2 + (\sin t - s_2)^2 = 2. \tag{4}$$

We can simplify (4) to obtain

$$-2\left(s_1 \cos t + s_2 \sin t\right) + \left(s_1^2 + s_2^2\right) = 1. \tag{5}$$

To solve this, recall that the distance from S to the origin must be the same as the distance from R to the origin, which is 1. Hence

$$\begin{aligned} \sqrt{s_1^2 + s_2^2} &= 1 \\ s_1^2 + s_2^2 &= 1. \end{aligned}$$

Substituting this into (5), we find that

$$\begin{aligned} -2\left(s_1 \cos t + s_2 \sin t\right) + s_1^2 + s_2^2 &= 1 \\ -2\left(s_1 \cos t + s_2 \sin t\right) + 1 &= 1 \\ -2\left(s_1 \cos t + s_2 \sin t\right) &= 0 \\ s_1 \cos t &= -s_2 \sin t. \end{aligned} \tag{6}$$

At this point we can see that $s_1 = \sin t$ and $s_2 = -\cos t$ would solve the problem; so would $s_1 = -\sin t$ and $s_2 = \cos t$. Are there any other solutions?

Recall that $s_1^2 + s_2^2 = 1$, so $s_2 = \pm\sqrt{1 - s_1^2}$. Likewise $\sin t = \pm\sqrt{1 - \cos^2 t}$. Substituting into equation (6) and squaring (so as to remove the radicals), we find that

$$\begin{aligned} s_1 \cos t &= -\sqrt{1 - s_1^2} \cdot \sqrt{1 - \cos^2 t} \\ s_1^2 \cos^2 t &= \left(1 - s_1^2\right)\left(1 - \cos^2 t\right) \\ s_1^2 \cos^2 t &= 1 - \cos^2 t - s_1^2 + s_1^2 \cos^2 t \\ s_1^2 &= 1 - \cos^2 t \\ s_1^2 &= \sin^2 t \\ \therefore s_1 &= \pm \sin t. \end{aligned}$$

Along with equation (6), this implies that $s_2 = \mp \cos t$. Thus there are *two* possible values of s_1 and s_2.

It can be shown (see Exercise 2.49) that α is a linear transformation on the vector space $\mathbb{R}^2$ with the basis $\left\{\vec{P}, \vec{R}\right\} = \{(1,0), (0,1)\}$. Linear algebra tells us that we can describe any linear transformation as a matrix. If $s = (\sin t, -\cos t)$ then

$$\alpha = \begin{pmatrix} \cos t & \sin t \\ \sin t & -\cos t \end{pmatrix};$$

otherwise

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

The lemma names the first of these forms φ and the second ρ. □

Before answering the second question, let's consider an example of what the two basic forms of α do to the points in the plane.

Example 2.44. Consider the set of points

$$\mathcal{S} = \{(0,2),(\pm 2,1),(\pm 1,-2)\};$$

these form the vertices of a (non-regular) pentagon in the plane. Let $t = \pi/4$; then

$$\rho = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} \quad \text{and} \quad \varphi = \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix}.$$

If we apply ρ to every point in the plane, then the points of $\mathcal{S}$ move to

$$\begin{aligned} \rho(\mathcal{S}) &= \{\rho(0,2),\rho(-2,1),\rho(2,1),\rho(-1,-2),\rho(1,-2)\} \\ &= \left\{\left(-\sqrt{2},\sqrt{2}\right),\left(-\sqrt{2}-\frac{\sqrt{2}}{2},-\sqrt{2}+\frac{\sqrt{2}}{2}\right),\right. \\ &\qquad \left(\sqrt{2}-\frac{\sqrt{2}}{2},\sqrt{2}+\frac{\sqrt{2}}{2}\right), \\ &\qquad \left(-\frac{\sqrt{2}}{2}+\sqrt{2},-\frac{\sqrt{2}}{2}-\sqrt{2}\right), \\ &\qquad \left.\left(\frac{\sqrt{2}}{2}+\sqrt{2},\frac{\sqrt{2}}{2}-\sqrt{2}\right)\right\} \\ &\approx \{(-1.4,1.4),(-2.1,-0.7),(0.7,2.1), \\ &\qquad (0.7,-2.1),(2.1,-0.7)\}. \end{aligned}$$

This is a 45° ($\pi/4$) counterclockwise rotation in the plane.

If we apply φ to every point in the plane, then the points of $\mathcal{S}$ move to

$$\begin{aligned} \varphi(\mathcal{S}) &= \{\varphi(0,2),\varphi(-2,1),\varphi(2,1),\varphi(-1,-2),\varphi(1,-2)\} \\ &\approx \{(1.4,-1.4),(-0.7,-2.1),(2.1,0.7), \\ &\qquad \downarrow(-2.1,0.7),(-0.7,2.1)\}. \end{aligned}$$

This is shown in Figure 2.5 . The line of reflection for φ has slope $\left(1-\cos\frac{\pi}{4}\right)/\sin\frac{\pi}{4}$. (You will show this in Exercise 2.51)

The second questions asks which of the matrices described by Lemma 2.43 also preserve the triangle.

- The first solution (ρ) corresponds to a rotation of degree t of the plane. To preserve the triangle, we can only have $t = 0, 2\pi/3, 4\pi/3$ (0°, 120°, 240°). (See Figure 2.4(a).) Let ι

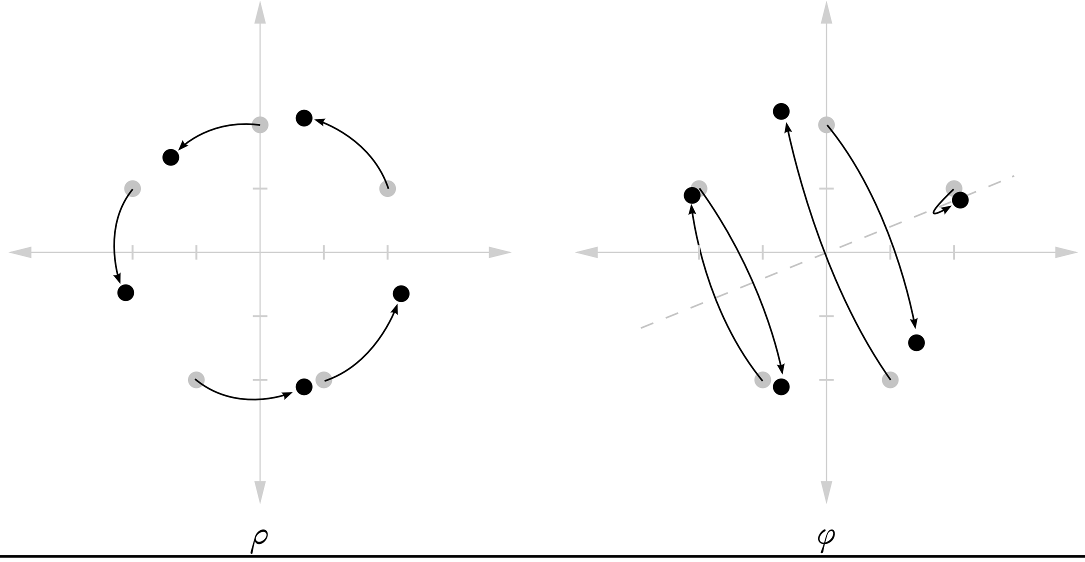


Figure 2.5. Actions of ρ and φ on a pentagon, with $t = \pi/4$

correspond to $t = 0$, the identity rotation; notice that

$$\iota = \begin{pmatrix} \cos 0 & -\sin 0 \\ \sin 0 & \cos 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

which is what we would expect for the identity. We can let ρ correspond to a counterclockwise rotation of 120°, so

$$\rho = \begin{pmatrix} \cos\frac{2\pi}{3} & -\sin\frac{2\pi}{3} \\ \sin\frac{2\pi}{3} & \cos\frac{2\pi}{3} \end{pmatrix} = \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}.$$

A rotation of 240° is the same as rotating 120° twice. We can write that as $\rho \circ \rho$ or ρ^2; matrix multiplication gives us

$$\begin{aligned} \rho^2 &= \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix} \\ &= \begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}. \end{aligned}$$

- The second solution (φ) corresponds to a flip along the line whose slope is

  $$m = (1 - \cos t) / \sin t.$$

  One way to do this would be to flip across the y-axis (see Figure 2.4(b)). For this we need the slope to be undefined, so the denominator needs to be zero and the numerator needs to be non-zero. One possibility for t is $t = \pi$ (but not $t = 0$). So

  $$\varphi = \begin{pmatrix} \cos\pi & \sin\pi \\ \sin\pi & -\cos\pi \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}.$$

There are two other flips, but we can actually ignore them, because they are combinations of φ and ρ. (Why? See Exercise 2.48.)

We can now give more detailed proofs of Theorems 2.39 and 2.40. We'll prove the first here, and you'll prove the second in the exercises.

Detailed proof of Theorem 2.39. Compare

$$\varphi\rho = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}$$

and

$$\begin{aligned} \rho^2\varphi &= \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}. \end{aligned}$$

□

Exercises.

Unless otherwise specified ρ and φ refer to the elements of D_3.

Exercise 2.45. Show explicitly (by matrix multiplication) that $\rho^3 = \varphi^2 = \iota$.

Exercise 2.46. The multiplication table for D_3 has at least this structure:

$\circ$	ι	φ	ρ	ρ^2	$\rho\varphi$	$\rho^2\varphi$
ι	ι	φ	ρ	ρ^2	$\rho\varphi$	$\rho^2\varphi$
φ	φ		$\rho^2\varphi$			
ρ	ρ	$\rho\varphi$				
ρ^2	ρ^2					
$\rho\varphi$	$\rho\varphi$					
$\rho^2\varphi$	$\rho^2\varphi$					

Complete the multiplication table, writing every element in the form $\rho^m\varphi^n$, never with φ before ρ. *Do not use matrix multiplication;* instead, use Theorems 2.39 and 2.40.

Exercise 2.47. Find a geometric figure (not a polygon) that is preserved by at least one rotation, at least one reflection, *and* at least one translation.

Exercise 2.48. Two other values of t allow us to define flips for the triangle. Find these values of t, and explain why their matrices are equivalent to the matrices $\rho\varphi$ and $\rho^2\varphi$.

Exercise 2.49. Show that any function α satisfying the requirements of Theorem 2.43 is a linear transformation; that is, for all $P, Q \in \mathbb{R}^2$ and for all $a, b \in \mathbb{R}$, $\alpha(aP + bQ) = a\alpha(P) + b\alpha(Q)$. Use the following steps.

(a) Prove that $\alpha(P)\cdot\alpha(Q) = P\cdot Q$, where $\cdot$ denotes the usual dot product (or inner product) on $\mathbb{R}^2$.
(b) Show that $\alpha(1,0)\cdot\alpha(0,1) = 0$.
(c) Show that $\alpha((a,0)+(0,b)) = a\alpha(1,0) + b\alpha(0,1)$.
(d) Show that $\alpha(aP) = a\alpha(P)$.
(e) Show that $\alpha(P+Q) = \alpha(P) + \alpha(Q)$.

Exercise 2.50. Show that the only stationary point in $\mathbb{R}^2$ for the general ρ is the origin. That is, if $\rho(P) = P$, then $P = (0,0)$. (By "general", we mean any ρ, not just the one in D_3.)

Exercise 2.51. Fill in each blank of Figure 2.6 with the appropriate justification.

Claim: The only stationary points of φ lie along the line whose slope is $(1-\cos t)/\sin t$, where $t \in [0,2\pi)$ and $t \neq 0, \pi$. If $t = 0$, only the x-axis is stationary, and for $t = \pi$, only the y-axis.
Proof:

1. Let $P \in \mathbb{R}^2$. By _____, there exist $x, y \in \mathbb{R}$ such that $P = (x,y)$.
2. Assume φ leaves P stationary. By _____,
$$\begin{pmatrix}\cos t & \sin t\\ \sin t & -\cos t\end{pmatrix}\begin{pmatrix}x\\ y\end{pmatrix} = \begin{pmatrix}x\\ y\end{pmatrix}.$$
3. By linear algebra,
$$\begin{pmatrix}\underline{\hspace{1.5cm}}\\ \underline{\hspace{1.5cm}}\end{pmatrix} = \begin{pmatrix}x\\ y\end{pmatrix}.$$
4. By the principle of linear independence, _____ $= x$ and _____ $= y$.
5. For each equation, collect x on the left hand side, and y on the right, to obtain
$$\begin{cases} x(\underline{\hspace{1cm}}) = -y(\underline{\hspace{1cm}})\\ x(\underline{\hspace{1cm}}) = y(\underline{\hspace{1cm}})\end{cases}.$$
6. If we solve the first equation for y, we find that $y =$ _____.
   (a) This, of course, requires us to assume that _____ $\neq 0$.
   (b) If that *was* in fact zero, then $t =$ _____, _____ (remembering that $t \in [0,2\pi)$).
7. Put these values of t aside. If we solve the second equation for y, we find that $y =$ _____.
   (a) Again, this requires us to assume that _____ $\neq 0$.
   (b) If that *was* in fact zero, then $t =$ _____. We already put this value aside, so ignore it.
8. Let's look at what happens when $t \neq$ _____ and _____.
   (a) Multiply numerator and denominator of the right hand side of the first solution by the denominator of the second to obtain $y =$ _____.
   (b) Multiply right hand side of the second with denominator of the first: $y =$ _____.
   (c) By _____, $\sin^2 t = 1 - \cos^2 t$. Substitution into the second solution gives the first!
   (d) That is, points that lie along the line $y =$ _____ are left stationary by φ.
9. Now consider the values of t we excluded.
   (a) If $t =$ _____, then the matrix simplifies to $\varphi =$ _____.
   (b) To satisfy $\varphi(P) = P$, we must have _____ $= 0$, and _____ free. The points that satisfy this are precisely the _____-axis.
   (c) If $t =$ _____, then the matrix simplifies to $\varphi =$ _____.
   (d) To satisfy $\varphi(P) = P$, we must have _____ $= 0$, and _____ free. The points that satisfy this are precisely the _____-axis.

Figure 2.6. Material for Exercise 2.51

2.3: Cyclic groups and order of elements

Here we re-introduce the familiar notation of exponents, in a manner consistent with what you learned for exponents of real numbers. We use this to describe an important class of groups that recur frequently.

Cyclic groups and generators

Notation 2.52. Let G be a group, and $g \in G$. If we want to perform the operation on g ten times, we could write

$$\prod_{i=1}^{10} g = g \cdot g \cdot g \cdot g \cdot g \cdot g \cdot g \cdot g \cdot g \cdot g$$

but this grows tiresome. Instead we will adapt notation from high-school algebra and write

$$g^{10}.$$

We likewise define g^{-10} to represent

$$\prod_{i=1}^{10} g^{-1} = g^{-1} \cdot g^{-1} \cdot g^{-1} \cdot g^{-1} \cdot g^{-1} \cdot g^{-1} \cdot g^{-1} \cdot g^{-1} \cdot g^{-1} \cdot g^{-1}.$$

Indeed, for any $n \in \mathbb{N}^+$ and any $g \in G$ we adopt the following convention:

- g^n means to perform the operation on n copies of g, so $g^n = \prod_{i=1}^n g$;
- g^{-n} means to perform the operation on n copies of g^{-1}, so $g^{-n} = \prod_{i=1}^n g^{-1} = \left(g^{-1}\right)^n$;
- $g^0 = e$, and if I want to be annoying I can write $g^0 = \prod_{i=1}^0 g$.

In additive groups we write instead $ng = \sum_{i=1}^n g$, $(-n)\, g = \sum_{i=1}^n (-g)$, and $0g = 0$.

Notice that this definition assume n is *positive*.

Definition 2.53. Let G be a group. If there exists $g \in G$ such that every element $x \in G$ has the form $x = g^n$ for some $n \in \mathbb{Z}$, then G is a **cyclic group** and we write $G = \langle g \rangle$. We call g a **generator** of G.

The idea of a cyclic group is that it has the form

$$\left\{\ldots, g^{-2}, g^{-1}, e, g^1, g^2, \ldots\right\}.$$

If the group is additive, we would of course write

$$\{\ldots, -2g, -g, 0, g, 2g, \ldots\}.$$

Example 2.54. $\mathbb{Z}$ is cyclic, since any $n \in \mathbb{Z}$ has the form $n \cdot 1$. Thus $\mathbb{Z} = \langle 1 \rangle$. In addition, n has the form $(-n) \cdot (-1)$, so $\mathbb{Z} = \langle -1 \rangle$ as well. Both 1 and -1 are generators of $\mathbb{Z}$.

You will show in the exercises that $\mathbb{Q}$ is not cyclic.

In Definition 2.53 we referred to g as *a* generator of G, not as *the* generator. There could in fact be more than one generator; we see this in Example 2.54 from the fact that $\mathbb{Z} = \langle 1 \rangle = \langle -1 \rangle$. Here is another example.

Example 2.55. Let

$$G = \left\{ \begin{array}{c} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \\ \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \end{array} \right\} \subsetneq \mathrm{GL}_m(\mathbb{R}).$$

It turns out that G is a group; both the second and third matrices generate it. For example,

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^2 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$$
$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^3 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$
$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^4 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

An important question arises here. Given a group G and an element $g \in G$, define

$$\langle g \rangle = \left\{ \ldots, g^{-2}, g^{-1}, e, g, g^2, \ldots \right\}.$$

We know that every cyclic group has the form $\langle g \rangle$ for some $g \in G$. Is the converse also true that $\langle g \rangle$ is a group for any $g \in G$? As a matter of fact, yes!

Theorem 2.56. For every group G and for every $g \in G$, $\langle g \rangle$ is an abelian group.

To prove Theorem 2.56, we need to make sure we can perform the usual arithmetic on exponents.

Lemma 2.57. Let G be a group, $g \in G$, and $m, n \in \mathbb{Z}$. Each of the following holds:
(A) $g^m g^{-m} = e$; that is, $g^{-m} = (g^m)^{-1}$.
(B) $(g^m)^n = g^{mn}$.
(C) $g^m g^n = g^{m+n}$.

The proof will justify this argument by applying the notation described at the beginning of this chapter. We have to be careful with this approach, because in the lemma we have $m, n \in \mathbb{Z}$, but the notation was given under the assumption that $n \in \mathbb{N}^+$. To make this work, we'll have to consider the cases where m and n are positive or negative separately. We call this a *case analysis*.

Proof. Each claim follows by case analysis.

(A) If $m = 0$, then $g^{-m} = g^0 = e = e^{-1} = \left(g^0\right)^{-1} = (g^m)^{-1}$.

Otherwise, $m \neq 0$. First assume that $m \in \mathbb{N}^+$. By notation, $g^{-m} = \prod_{i=1}^{m} g^{-1}$. Hence

$$\begin{aligned} g^m g^{-m} &\underset{\text{def.}}{=} \left(\prod_{i=1}^{m} g\right)\left(\prod_{i=1}^{m} g^{-1}\right) \\ &\underset{\text{ass.}}{=} \left(\prod_{i=1}^{m-1} g\right)\left(g \cdot g^{-1}\right)\left(\prod_{i=1}^{m-1} g^{-1}\right) \\ &\underset{\text{id.}}{=} \left(\prod_{i=1}^{m-1} g\right) e \left(\prod_{i=1}^{m-1} g^{-1}\right) \\ &\underset{\text{inv.}}{=} \left(\prod_{i=1}^{m-1} g\right)\left(\prod_{i=1}^{m-1} g^{-1}\right) \\ &\vdots \\ &= e. \end{aligned}$$

Since the inverse of an element is unique, $g^{-m} = (g^m)^{-1}$.

Now assume that $m \in \mathbb{Z}\backslash\mathbb{N}$. Since m is negative, we cannot express the product using m; the notation discussed on page 60 requires a *positive* exponent. Consider instead $\widehat{m} = |m| \in \mathbb{N}^+$. Since the opposite of a negative number is positive, we can write $-m = \widehat{m}$ and $-\widehat{m} = m$. Since $\widehat{m}$ is positive, we can apply the notation to it directly; $g^{-m} = g^{\widehat{m}} = \prod_{i=1}^{\widehat{m}} g$, while $g^m = g^{-\widehat{m}} = \prod_{i=1}^{\widehat{m}} g^{-1}$. (To see this in a more concrete example, try it with an actual number. If $m = -5$, then $\widehat{m} = |-5| = 5 = -(-5)$, so $g^m = g^{-5} = g^{-\widehat{m}}$ and $g^{-m} = g^5 = g^{\widehat{m}}$.) As above, we have

$$g^m g^{-m} \underset{\text{subs.}}{=} g^{-\widehat{m}} g^{\widehat{m}} \underset{\text{not.}}{=} \left(\prod_{i=1}^{\widehat{m}} g^{-1}\right)\left(\prod_{i=1}^{\widehat{m}} g\right) = e.$$

Hence $g^{-m} = (g^m)^{-1}$.

(B) If $n = 0$, then $(g^m)^n = (g^m)^0 = e$ because *anything* to the zero power is e. Assume first that $n \in \mathbb{N}^+$. By notation, $(g^m)^n = \prod_{i=1}^{n} g^m$. We split this into two subcases.

(B1) If $m \in \mathbb{N}$, we have

$$(g^m)^n \underset{\text{not.}}{=} \prod_{i=1}^{n}\left(\prod_{i=1}^{m} g\right) \underset{\text{ass.}}{=} \prod_{i=1}^{mn} g \underset{\text{not.}}{=} g^{mn}.$$

(B2) Otherwise, let $\widehat{m} = |m| \in \mathbb{N}^+$ and we have

$$\begin{aligned} (g^m)^n &\underset{\text{subs.}}{=} \left(g^{-\widehat{m}}\right)^n \underset{\text{not.}}{=} \prod_{i=1}^{n}\left(\prod_{i=1}^{\widehat{m}} g^{-1}\right) \\ &\underset{\text{ass.}}{=} \prod_{i=1}^{\widehat{m}n} g^{-1} \underset{\text{not.}}{=} \left(g^{-1}\right)^{\widehat{m}n} \\ &\underset{\text{not.}}{=} g^{-\widehat{m}n} \underset{\text{subs.}}{=} g^{mn}. \end{aligned}$$

What if n is negative? Let $\widehat{n} = -n$; by notation, $(g^m)^n = (g^m)^{-\widehat{n}} = \prod_{i=1}^{\widehat{n}} (g^m)^{-1}$. By (A), this becomes $\prod_{i=1}^{\widehat{n}} g^{-m}$. By notation, we can rewrite this as $\left(g^{-m}\right)^{\widehat{n}}$. Since $\widehat{n} \in \mathbb{N}^+$, we can apply case (B1) or (B2) as appropriate, so

$$\begin{aligned}(g^m)^n = \left(g^{-m}\right)^{\widehat{n}} &\underset{\text{(B1) or (B2)}}{=} g^{(-m)\widehat{n}}\\ &\underset{\text{integers!}}{=} g^{m(-\widehat{n})} \underset{\text{subst}}{=} g^{mn}.\end{aligned}$$

(C) We consider three cases.
If $m = 0$ or $n = 0$, then $g^0 = e$, so $g^{-0} = g^0 = e$.
If m, n have the same sign (that is, $m, n \in \mathbb{N}^+$ or $m, n \in \mathbb{Z}\backslash\mathbb{N}$), then write $\widehat{m} = |m|$, $\widehat{n} = |n|$, $g_m = g^{\frac{\widehat{m}}{m}}$, and $g_n = g^{\frac{\widehat{n}}{n}}$. This effects a really nice trick: if $m \in \mathbb{N}^+$, then $g_m = g$, whereas if m is negative, $g_m = g^{-1}$. This notational trick allows us to write $g^m = \prod_{i=1}^{\widehat{m}} g_m$ and $g^n = \prod_{i=1}^{\widehat{n}} g_n$, where $g_m = g_n$ and $\widehat{m}$ and $\widehat{n}$ are both positive integers. Then

$$\begin{aligned}g^m g^n = \prod_{i=1}^{\widehat{m}} g_m \prod_{i=1}^{\widehat{n}} g_n &= \prod_{i=1}^{\widehat{m}} g_m \prod_{i=1}^{\widehat{n}} g_m\\ &= \prod_{i=1}^{\widehat{m}+\widehat{n}} g_m = (g_m)^{\widehat{m}+\widehat{n}} = g^{m+n}.\end{aligned}$$

Since g and n were arbitrary, the induction implies that $g^n g^{-n} = e$ for all $g \in G$, $n \in \mathbb{N}^+$. Now consider the case where m and n have different signs. In the first case, suppose m is negative and $n \in \mathbb{N}^+$. As in (A), let $\widehat{m} = |m| \in \mathbb{N}^+$; then

$$g^m g^n = \left(g^{-1}\right)^{-m} g^n = \left(\prod_{i=1}^{\widehat{m}} g^{-1}\right)\left(\prod_{i=1}^{n} g\right).$$

If $\widehat{m} \geq n$, we have more copies of g^{-1} than g, so after cancellation,

$$g^m g^n = \prod_{i=1}^{\widehat{m}-n} g^{-1} = g^{-(\widehat{m}-n)} = g^{m+n}.$$

Otherwise, $\widehat{m} < n$, and we have more copies of g than of g^{-1}. After cancellation,

$$g^m g^n = \prod_{i=1}^{n-\widehat{m}} g = g^{n-\widehat{m}} = g^{n+m} = g^{m+n}.$$

The remaining case ($m \in \mathbb{N}^+$, $n \in \mathbb{Z}\backslash\mathbb{N}$) is similar, and you will prove it for homework.

□

These properties of exponent arithmetic allow us to show that $\langle g \rangle$ is a group.

Proof of Theorem 2.56. We show that $\langle g \rangle$ satisfies the properties of an abelian group. Let $x, y, z \in \langle g \rangle$. By definition of $\langle g \rangle$, there exist $a, b, c \in \mathbb{Z}$ such that $x = g^a$, $y = g^b$, and $z = g^c$. We will use Lemma 2.57 implicitly.

- By substitution, $xy = g^a g^b = g^{a+b} \in \langle g \rangle$. So $\langle g \rangle$ is closed.
- By substitution, $x(yz) = g^a \left(g^b g^c\right)$. These are elements of G by inclusion (that is, $\langle g \rangle \subseteq G$ so $x, y, z \in G$), so the associative property *in* G gives us
$$x(yz) = g^a \left(g^b g^c\right) = \left(g^a g^b\right) g^c = (xy)z.$$
- By definition, $e = g^0 \in \langle g \rangle$.
- By definition, $g^{-a} \in \langle g \rangle$, and $x \cdot g^{-a} = g^a g^{-a} = e$. Hence $x^{-1} = g^{-a} \in \langle g \rangle$.
- Using the fact that $\mathbb{Z}$ is commutative under addition,
$$xy = g^a g^b = g^{a+b} = g^{b+a} = g^b g^a = yx.$$

□

The order of an element

Given an element and an operation, Theorem 2.56 links them to a group. It makes sense, therefore, to link an element to the order of the group that it generates.

Definition 2.58. Let G be a group, and $g \in G$. We say that the **order** of g is $\operatorname{ord}(g) = |\langle g \rangle|$. If $\operatorname{ord}(g) = \infty$, we say that g has **infinite order**.

If the order of a group is finite, then we can write an element in different ways.

Example 2.59. Recall Example 2.55; we can write

$$\begin{aligned}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^0 &= \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^4 \\ &= \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^8 = \cdots .\end{aligned}$$

Since multiples of 4 give the identity, let's take any power of the matrix, and divide it by 4. The Division Theorem allows us to write any power of the matrix as $4q + r$, where $0 \leq r < 4$. Since there are only four possible remainders, and multiples of 4 give the identity, positive powers of this matrix can generate only four possible matrices:

$$\begin{aligned}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^{4q} &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \\ \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^{4q+1} &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \\ \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^{4q+2} &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \\ \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^{4q+3} &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.\end{aligned}$$

We can do the same with negative powers; the Division Theorem still gives us only four possible remainders. Let's write

$$g = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Thus

$$\langle g \rangle = \left\{ I_2, g, g^2, g^3 \right\}.$$

The example suggests that if the order of an element G is $n \in \mathbb{N}$, then we can write

$$\langle g \rangle = \left\{ e, g, g^2, \ldots, g^{n-1} \right\}.$$

This explains why we call $\langle g \rangle$ a *cyclic* group: once they reach $\operatorname{ord}(g)$, the powers of g "cycle".To prove this in general, we have to show that for a generic cyclic group $\langle g \rangle$ with $\operatorname{ord}(g) = n$,

- n is the smallest positive power that gives us the identity; that is, $g^n = e$, and
- for any two integers between 0 and n, the powers of g are different; that is, if $0 \le a < b < n$, then $g^a \ne g^b$.

Theorem 2.60 accomplishes that, and a bit more as well.

Theorem 2.60. Let G be a group, $g \in G$, and $\operatorname{ord}(g) = n$. Then
(A) for all $a, b \in \mathbb{N}$ such that $0 \le a < b < n$, we have $g^a \ne g^b$.
In addition, if $n < \infty$, each of the following holds:
(B) $g^n = e$;
(C) n is the smallest positive integer d such that $g^d = e$; and
(D) if $a, b \in \mathbb{Z}$ and $n \mid (a - b)$, then $g^a = g^b$.

Proof. The fundamental assertion of the theorem is (A). The remaining assertions turn out to be corollaries.

(A) By way of contradiction, suppose that there exist $a, b \in \mathbb{N}$ such that $0 \le a < b < n$ and $g^a = g^b$; then $e = (g^a)^{-1} g^b$. By Exercise 2.63, we can write

$$e = g^{-a} g^b = g^{-a+b} = g^{b-a}.$$

Let $S = \left\{ m \in \mathbb{N}^+ : g^m = e \right\}$. By the well-ordering property of $\mathbb{N}$, there exists a smallest element of S; call it d. Recall that $a < b$, so $b - a \in \mathbb{N}^+$, so $g^{b-a} \in S$. By the choice of d, we know that $d \le b - a$. By Exercise 1.24, $d \le b - a < b$, so $0 < d < b < n$.
We can now list d distinct elements of $\langle g \rangle$:

$$g, g^2, g^3, \ldots, g^d = e. \tag{7}$$

Since $d < n$, this list omits $n - d$ elements of $\langle g \rangle$. (If $\operatorname{ord}(g) = \infty$, then it omits infinitely many elements of $\langle g \rangle$.) Let x be one such element. By definition of $\langle g \rangle$, we can write $x = g^c$ for some $c \in \mathbb{Z}$. Choose q, r that satisfy the Division Theorem for division of c by d; that is,

$$c = qd + r \quad \text{such that} \quad q, d \in \mathbb{Z} \text{ and } 0 \le r < d.$$

We have $g^c = g^{qd+r}$. By Lemma 2.57,

$$g^c = \left(g^d\right)^q \cdot g^r = e^q \cdot g^r = e \cdot g^r = g^r.$$

Recall that $0 \leq r < d$, so we listed g^r above when we listed the powers of g less than d. Since $g^r = g^c$, we have already listed g^c. This contradicts the assumption that $g^c = g^r$ was not listed. Hence if $0 \leq a < b < n$, then $g^a \neq g^b$.

For the remainder of the proof, we assume that $n < \infty$.

(B) Let $S = \left\{m \in \mathbb{N}^+ : g^m = e\right\}$. Is S non-empty? Since $\langle g \rangle < \infty$, there must exist $a, b \in \mathbb{N}^+$ such that $a < b$ and $g^a = g^b$. Using the inverse property and substitution, $g^0 = e = g^b \left(g^a\right)^{-1}$. By Lemma 2.57, $g^0 = g^{b-a}$. By definition, $b - a \in \mathbb{N}^+$. Hence S is non-empty.
By the well-ordering property of $\mathbb{N}$, there exists a smallest element of S; call it d. Since $\langle g \rangle$ contains n elements, $1 < d \leq n$. If $d < n$, that would contradict assertion (A) of this theorem (with $a = 0$ and $b = d$). Hence $d = n$, and $g^n = e$, and we have shown (A).

(C) In (B), S is the set of all positive integers m such that $g^m = e$; we let the smallest element be d, and thus $d \leq n$. On the other hand, (A) tells us that we cannot have $d < n$; otherwise, $g^d = g^0 = e$. Hence, $n \leq d$. We already had $d \leq n$, so the two must be equal.

(D) Let $a, b \in \mathbb{Z}$. Assume that $n \mid (a - b)$. Let $q \in \mathbb{Z}$ such that $nq = a - b$. Then

$$\begin{aligned} g^b &= g^b \cdot e = g^b \cdot e^q \\ &= g^b \cdot (g^n)^q = g^b \cdot g^{nq} \\ &= g^b \cdot g^{a-b} = g^{b+(a-b)} = g^a. \end{aligned}$$

□

We conclude therefore that, at least when they are finite, cyclic groups are aptly named: increasing powers of g generate new elements until the power reaches n, in which case $g^n = e$ and we "cycle around".

Exercises.

Exercise 2.61. Recall from Example 2.55 the matrix

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

Express A as a power of the other non-identity matrices of the group.

Exercise 2.62. In Exercise 2.36 you showed that the quaternions form a group under matrix multiplication. Verify that $H = \{\mathbf{1}, -\mathbf{1}, \mathbf{i}, -\mathbf{i}\}$ is a cyclic group. What elements generate H?

Exercise 2.63. Complete the proof of Lemma 2.57(C).

Exercise 2.64. Fill in each blank of Figure 2.7 with the justification or statement.

Let G be a group, and $g \in G$. Let $d, n \in \mathbb{Z}$ and assume $\operatorname{ord}(g) = d$.
Claim: $g^n = e$ if and only if $d \mid n$.
Proof:

1. Assume that $g^n = e$.
   (a) By _____, there exist $q, r \in \mathbb{Z}$ such that $n = qd + r$ and $0 \leq r < d$.
   (b) By _____, $g^{qd+r} = e$.
   (c) By _____, $g^{qd} g^r = e$.
   (d) By _____, $\left(g^d\right)^q g^r = e$.
   (e) By _____, $e^q g^r = e$.
   (f) By _____, $eg^r = e$. By the identity property, $g^r = e$.
   (g) By _____, d is the *smallest* positive integer such that $g^d = e$.
   (h) Since _____, it cannot be that r is positive. Hence, $r = 0$.
   (i) By _____, $g = qd$. By definition, then $d \mid n$.
2. Now we show the converse. Assume that _____.
   (a) By definition of divisibility, _____.
   (b) By substitution, $g^n =$_____.
   (c) By Lemma 2.57, the right hand side of that equation can be rewritten as to _____.
   (d) Recall that $\operatorname{ord}(g) = d$. By definition of order, $g^d = e$, so we can rewrite the right hand side again as _____.
   (e) A little more simplification turns the right hand side into _____, which obviously simplifies to e.
   (f) By _____, then, $g^n = e$.
3. We showed first that if $g^n = e$, then $d|n$; we then showed that _____. This proves the claim.

Figure 2.7. Material for Exercise 2.64

Exercise 2.65. Show that any group of 3 elements is cyclic.

Exercise 2.66. Is the Klein 4-group (Exercise 2.32 on page 49) cyclic? What about the cyclic group of order 4?

Exercise 2.67. Show that Q_8 is not cyclic.

Exercise 2.68. Show that $\mathbb{Q}$ is not cyclic.

Exercise 2.69. Use a fact from linear algebra to explain why $GL_m(\mathbb{R})$ is not cyclic.

2.4: The roots of unity

One of the major motivations in the development of group theory was to study roots of polynomials. A polynomial, of course, has the form

$$ax + b, \quad ax^2 + bx + c, \quad ax^3 + bx^2 + cx + d, \quad \ldots$$

A **root** of a polynomial $f(x)$ is any a such that $f(a) = 0$. For example, if $f(x) = x^4 - 1$, then 1 and -1 are both roots of f. However, they are not the *only* roots of f! For the full explanation, you'll need to read about polynomial rings and ideals in Chapters 7 and 8, but we can take some first steps in that direction already.

Imaginary and complex numbers

First, notice that f factors as $f(x) = (x-1)(x+1)(x^2+1)$. The roots 1 and -1 show up in the linear factors, and they're the only possible roots of those factors. So, if f has other roots, we would expect them to be roots of x^2+1. However, the square of a real number is nonnegative; adding 1 forces it to be positive. So, x^2+1 has no roots in $\mathbb{R}$.

Let's make a root up, anyway. If it doesn't make sense, we should find out soon enough. Let's call this polynomial $g(x) = x^2+1$, and say that g has a root, which we'll call i, for "imaginary". Since i is a root of g, we have the equation

$$0 = g(i) = i^2+1,$$

or $i^2 = -1$.

We'll create a new set of numbers by adding i to the set $\mathbb{R}$. Since $\mathbb{R}$ is a monoid under multiplication and a group under addition, we'd like to preserve those properties as well. This means we have to define multiplication and addition for our new set, and maybe add more objects, too.

We start with $\mathbb{R}\cup\{i\}$. Does multiplication add any new elements? Since $i^2 = -1$, and $-1 \in \mathbb{R}$ already, we're okay there. On the other hand, for any $b \in \mathbb{R}$, we'd like to multiply b and i. Since bi is not already in our new set, we'll have to add it if we want to keep multiplication closed. Our set has now expanded to $\mathbb{R}\cup\{bi : b \in \mathbb{R}\}$.

Let's look at addition. Our new set has real numbers like 1 and "imaginary" numbers like $2i$; if addition is to satisfy closure, we need $1+2i$ to be in the set, too. That's not the case yet, so we have to extend our set by $a+bi$ for any $a, b \in \mathbb{R}$. That gives us

$$\mathbb{R}\cup\{bi : b\in\mathbb{R}\}\cup\{a+bi : a,b\in\mathbb{R}\}.$$

If you think about it, the first two sets are in the third; just let $a = 0$ or $b = 0$ and you get bi or a, respectively. So, we can simplify our new set to

$$\{a+bi : a,b\in\mathbb{R}\}.$$

Do we need anything else?

We haven't checked closure of addition. In fact, we still haven't *defined* addition of complex numbers. We will borrow an idea from polynomials, and add complex numbers by adding like terms; that is, $(a+bi)+(c+di) = (a+c)+(b+d)i$. Closure implies that $a+c \in \mathbb{R}$ and $b+d \in \mathbb{R}$, so this is just another expression in the form already described. In fact, we can also see what additive inverses look like; after all, $(a+bi)+(-a-bi) = 0$. We don't have to add any new objects to our set to maintain the group structure of addition.

We also haven't checked closure of multiplication in this larger set — or even defined it, really. Again, let's borrow an idea from polynomials, and multiply complex numbers using the distributive property; that is,

$$(a+bi)(c+di) = ac+adi+bci+bdi^2.$$

Remember that $i^2 = -1$, and we can combine like terms, so the expression above simplifies to

$$(a+bi)(c+di) = (ac-bd) + (ad+bc)i.$$

Since $ac - bd \in \mathbb{R}$ and $ad + bc \in \mathbb{R}$, this is just another expression in the form already described. Again, we don't have to add any new objects to our set.

Definition 2.70. The **complex numbers** are the set

$$\mathbb{C} = \left\{a + bi : a, b \in \mathbb{R}, \ i^2 = -1\right\}.$$

The **real part** of $a + bi$ is a, and the **imaginary part** is b.

We can now state with confidence that we have found what we wanted to obtain.

Theorem 2.71. $\mathbb{C}$ is a monoid under multiplication, and an abelian group under addition.

Proof. Let $x, y, z \in \mathbb{C}$. Write $x = a + bi$, $y = c + di$, and $z = e + fi$, for some $a, b, c, d, e, f \in \mathbb{R}$. Let's look at multiplication first.

closure? We built $\mathbb{C}$ to be closed under multiplication, so the discussion above suffices.

associative? We need to show that

$$(xy)z = x(yz). \tag{8}$$

Expanding the product on the left, we have

$$[(a+bi)(c+di)](e+fi) = [(ac-bd)+(ad+bc)i](e+fi).$$

Expand again, and we get

$$\begin{aligned}[(a+bi)(c+di)](e+fi) = {} & [(ac-bd)e-(ad+bc)f] \\ & + [(ac-bd)f+(ad+bc)e]i.\end{aligned}$$

Now let's look at the product on the right of equation (8). Expanding it, we have

$$(a+bi)[(c+di)(e+fi)] = (a+bi)[(ce-df)+(cf+de)i].$$

Expand again, and we get

$$\begin{aligned}(a+bi)[(c+di)(e+fi)] = {} & [a(ce-df)-b(cf+de)] \\ & + [a(cf+de)+b(ce-df)]i.\end{aligned}$$

If you look carefully, you will see that both expansions resulted in the same complex number:

$$(ace-bde-adf-bcf)+(acf-bdf+ade+bce)i.$$

Thus, multiplication is $\mathbb{C}$ is associative.

identity? We claim that $1 \in \mathbb{R}$ is the multiplicative identity even for $\mathbb{C}$. Recall that we can write $1 = 1 + 0i$. Then,

$$1x = (1 + 0i)(a + bi) = (1a - 0b) + (1b + 0a)i = a + bi = x.$$

Since x was arbitrary in $\mathbb{C}$, it must be that 1 is, in fact, the identity.

We have shown that $\mathbb{C}$ is a monoid under multiplication. What about addition; it is a group? We leave that to the exercises. □

There are a *lot* of wonderful properties of $\mathbb{C}$ that we could discuss. For example, you can see that the roots of $x^2 + 1$ lie in $\mathbb{C}$, but what of the roots of $x^2 + 2$? It turns out that they're in there, too. In fact, *every* polynomial of degree n with real coefficients has n roots in $\mathbb{C}$! We need a lot more theory to discuss that, however, so we pass over it for the time being. In any case, we can now talk about a group that is both interesting and important.

Remark 2.72. You may wonder if we really *can* just make up some number i, and build a new set by adjoining it to $\mathbb{R}$. Isn't that just a little, oh, *imaginary?* Actually, no, it is quite concrete, and we can provide two very sound justifications.

First, mathematicians typically model the oscillation of a pendulum by a differential equations of the form $y'' + ay = 0$. As any book in the subject explains, we have good reason to solve such *differential* equations by resorting to *auxiliary polynomial* equations of the form $r^2 + a = 0$. The solutions to this equation are $r = \pm i\sqrt{a}$, so unless the oscillation of a pendulum is "imaginary", i is quite "real".

Second, we can construct from the real numbers a set that looks an awful lot like these purported complex numbers, using a very sensible approach, and we can even show that this set is isomorphic to the complex numbers in all the ways that we would like. That's a bit beyond us; you will learn more in Section 8.4.

The complex plane

We can diagram the real numbers along a line. In fact, it's quite easy to argue that what makes real numbers "real" is precisely the fact that they measure location or distance along a line. That's only one-dimensional, and you've seen before that we can do something similar on the plane or in space using $\mathbb{R}^2$ and $\mathbb{R}^3$.

What about the complex numbers? By definition, any complex number is the sum of its real and imaginary parts. We cannot simplify $a + bi$ any further using this representation, much as we cannot simplify the point $(a, b) \in \mathbb{R}^2$ any further. Since $\mathbb{R}^2$ forms a *vector space* over $\mathbb{R}$, does $\mathbb{C}$ also form a vector space over $\mathbb{R}$? In fact, it does! Here's a quick reminder of what makes a vector space:

- addition of vectors must satisfy closure and the associative, commutative, identity, and inverse properties;
- multiplication of vectors *by scalars* must have an identity scalar, must be associative on the scalars, and must satisfy the properties of distribution of scalars to vectors and vice-versa.

The properties for addition of vectors are precisely the properties of a group — and Theorem 2.71 tells us that $\mathbb{C}$ *is* a group under addition! All that remains is to show that $\mathbb{C}$ satisfies the required properties of multiplication. You will do that in Exercise 2.86.

Right now, we are more interested in the *geometric* implications of this relationship. We've already hinted that $\mathbb{C}$ and $\mathbb{R}^2$ have a similar structure. Let's start with the notion of *dimension.* Do you remember what that word means? Essentially, the dimension of a vectors space is the number of *basis vectors* needed to describe a vector space. Do $\mathbb{C}$ and $\mathbb{R}^2$ have the same dimension over $\mathbb{R}$? For that, we need to identify a *basis* of $\mathbb{C}$ over $\mathbb{R}$.

Theorem 2.73. $\mathbb{C}$ is a vector space over $\mathbb{R}$ with basis $\{1, i\}$.

Proof. We have already discussed why $\mathbb{C}$ is a vector space over $\mathbb{R}$; we still have to show that $\{0, i\}$ is a basis of $\mathbb{C}$. This is straightforward from the definition of $\mathbb{C}$, as any element can be written in terms of the basis elements as $a + bi = a \cdot 1 + b \cdot i$. □

We see from Theorem 2.73 that $\mathbb{C}$ and $\mathbb{R}^2$ do have the same dimension! After all, any point of $\mathbb{R}^2$ can be written as $(a, b) = a(1,0) + b(0,1)$, so a basis of $\mathbb{R}^2$ is $\{(1,0), (0,1)\}$.

This will hopefully prompt you to realize that $\mathbb{C}$ and $\mathbb{R}^2$ are identical as vector spaces. For our purposes, what matters that we can map any point of $\mathbb{C}$ to a unique point of $\mathbb{R}^2$, and vice-versa.

Theorem 2.74. There is a one-to-one, onto function from $\mathbb{C}$ to $\mathbb{R}^2$ that maps the basis vectors 1 to $(1,0)$ and i to $(0,1)$.

Proof. Let $\varphi : \mathbb{C} \to \mathbb{R}^2$ by $\varphi(a + bi) = (a, b)$. That is, we map a complex number to $\mathbb{R}^2$ by sending the real part to the first entry (the x-ordinate) and the imaginary part to the second entry (the y-ordinate). As desired, $\varphi(1) = (1,0)$ and $\varphi(i) = (0,1)$.

Is this a bijection? We see that φ is one-to-one by the fact that if $\varphi(a + bi) = \varphi(c + di)$, then $(a, b) = (c, d)$; equality of points in $\mathbb{R}^2$ implies that $a = c$ and $b = d$; equality of complex numbers implies that $a + bi = c + di$. We see that φ is onto by the fact that for any $(a, b) \in \mathbb{R}^2$, $\varphi(a + bi) = (a, b)$. □

Since $\mathbb{R}^2$ has a nice, geometric representation as the x-y plane, we can represent complex numbers in the same way. That motivates our definition of the **complex plane**, which is nothing more than a visualization of $\mathbb{C}$ in $\mathbb{R}^2$.

Take a look at Figure 2.8. We have labeled the x-axis as $\mathbb{R}$ and the y-axis as $i\mathbb{R}$. We call the former the **real axis** and the latter the **imaginary axis** of the complex plane. This agrees with our mapping above, which sent the real part of a complex number to the x-ordinate, and the imaginary part to the y-ordinate. Thus, the complex number $2 + 3i$ corresponds to the point $(2, 3)$, while the complex number $-2i$ corresponds to the point $(0, -2)$.

We could say a great deal about the complex plane, but that would distract us from our main goal, which is to proceed further in group theory. Even so, we should not neglect one important and beautiful point.

Roots of unity

Any root of the polynomial $f(x) = x^n - 1$ is called a **root of unity**. These are very important in the study of polynomial roots. At least some of them satisfy a very nice form.

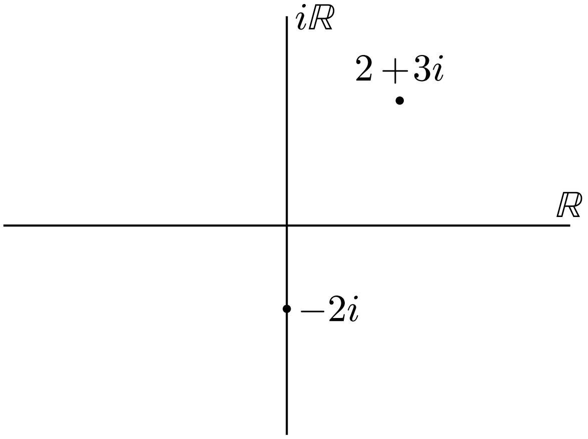


Figure 2.8. Two elements of $\mathbb{C}$, visualized as points on the complex plane

Theorem 2.75. Let $n \in \mathbb{N}^+$. The complex number

$$\omega = \cos\left(\frac{2\pi}{n}\right) + i\sin\left(\frac{2\pi}{n}\right)$$

is a root of $f(x) = x^n - 1$.

To prove Theorem 2.75, we need a different property of ω.

Lemma 2.76. If ω is defined as in Theorem 2.75, then

$$\omega^m = \cos\left(\frac{2\pi m}{n}\right) + i\sin\left(\frac{2\pi m}{n}\right)$$

for every $m \in \mathbb{N}^+$.

Proof. We proceed by induction on m. For the *inductive base,* the definition of ω shows that ω^1 has the desired form. For the *inductive hypothesis,* assume that ω^m has the desired form; in the *inductive step,* we need to show that

$$\omega^{m+1} = \cos\left(\frac{2\pi(m+1)}{n}\right) + i\sin\left(\frac{2\pi(m+1)}{n}\right).$$

To see why this is true, use the trigonometric sum identities $\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$

and $\sin(\alpha+\beta) = \sin\alpha\cos\beta + \sin\beta\cos\alpha$ to rewrite ω^{m+1}, like so:

$$\begin{aligned}
\omega^{m+1} &= \omega^m \cdot \omega \\
&\underset{\substack{\text{ind.}\\\text{hyp.}}}{=} \left[\cos\left(\frac{2\pi m}{n}\right) + i\sin\left(\frac{2\pi}{n}\right)\right] \\
&\quad \cdot \left[\cos\left(\frac{2\pi}{n}\right) + i\sin\left(\frac{2\pi}{n}\right)\right] \\
&= \cos\left(\frac{2\pi m}{n}\right)\cos\left(\frac{2\pi}{n}\right) + i\sin\left(\frac{2\pi}{n}\right)\cos\left(\frac{2\pi m}{n}\right) \\
&\quad + i\sin\left(\frac{2\pi m}{n}\right)\cos\left(\frac{2\pi}{n}\right) - \sin\left(\frac{2\pi m}{n}\right)\sin\left(\frac{2\pi}{n}\right) \\
&= \left[\cos\left(\frac{2\pi m}{n}\right)\cos\left(\frac{2\pi}{n}\right) - \sin\left(\frac{2\pi m}{n}\right)\sin\left(\frac{2\pi}{n}\right)\right] \\
&\quad + i\left[\sin\left(\frac{2\pi}{n}\right)\cos\left(\frac{2\pi m}{n}\right)\right. \\
&\qquad \left. + \sin\left(\frac{2\pi m}{n}\right)\cos\left(\frac{2\pi}{n}\right)\right] \\
&= \cos\left(\frac{2\pi(m+1)}{n}\right) + i\sin\left(\frac{2\pi(m+1)}{n}\right).
\end{aligned}$$

□

Once we have Lemma 2.76, proving Theorem 2.75 is spectacularly easy.

Proof of Theorem 2.75. Substitution and the lemma give us

$$\begin{aligned}
\omega^n - 1 &= \left[\cos\left(\frac{2\pi n}{n}\right) + i\sin\left(\frac{2\pi n}{n}\right)\right] - 1 \\
&= \cos 2\pi + i\sin 2\pi - 1 \\
&= (1 + i\cdot 0) - 1 = 0,
\end{aligned}$$

so ω is indeed a root of $x^n - 1$. □

As promised, $\langle\omega\rangle$ gives us a nice group.

Theorem 2.77. The nth roots of unity are $\Omega_n = \{1, \omega, \omega^2, \ldots, \omega^{n-1}\}$, where ω is defined as in Theorem 2.75. They form a cyclic group of order n under multiplication.

The theorem does not claim merely that Ω_n is a list of *some* nth roots of unity; it claims that Ω_n is a list of *all* nth roots of unity. Our proof is going to cheat a little bit, because we don't quite have the machinery to prove that Ω_n is an exhaustive list of the roots of unity. We will eventually, however, and you should be able to follow the general idea now. The idea is called *unique factorization.* Basically, let f be a polynomial of degree n. Suppose that we have n roots of f; call them $\alpha_1, \alpha_2, \ldots, \alpha_n$. The parts you have to take on faith (for now) are twofold. First, $x - \alpha_i$ is a factor of f for each α_i. Each linear factor adds one to the degree of a polynomial, and

f has degree n, so the number of linear factors cannot be more than n. Second, and this is not quite so clear, there is only one way to factor f into linear polynomials

(You can see this in the example above with $x^4 - 1$, but Theorem 7.41 on page 204 will have the details. You should have seen that theorem in your precalculus studies, and since it doesn't depend on anything in this section, the reasoning is not circular.)

If you're okay with that, then you're okay with everything else.

Proof. For $m \in \mathbb{N}^+$, we use the associative property of multiplication in $\mathbb{C}$ and the commutative property of multiplication in $\mathbb{N}^+$:

$$(\omega^m)^n - 1 = \omega^{mn} - 1 = \omega^{nm} - 1 = (\omega^n)^m - 1 = 1^m - 1 = 0.$$

Hence ω^m is a root of unity for any $m \in \mathbb{N}^+$. If $\omega^m = \omega^\ell$, then

$$\cos\left(\frac{2\pi m}{n}\right) = \cos\left(\frac{2\pi \ell}{n}\right) \quad \text{and} \quad \sin\left(\frac{2\pi m}{m}\right) = \sin\left(\frac{2\pi \ell}{n}\right),$$

and we know from trigonometry that this is possible only if

$$\begin{aligned} \frac{2\pi m}{n} &= \frac{2\pi \ell}{n} + 2\pi k \\ \frac{2\pi}{n}(m-\ell) &= 2\pi k \\ m - \ell &= kn. \end{aligned}$$

That is, $m - \ell$ is a multiple of n. Since Ω_n lists only those powers from 0 to $n-1$, the powers must be distinct, so Ω_n contains n distinct roots of unity. (See also Exercise 2.85.) As there can be at most n distinct roots, Ω_n is a complete list of nth roots of unity.

Now we show that Ω_n is a cyclic group.

(closure) Let $x, y \in \Omega_n$; you will show in Exercise 2.82 that $xy \in \Omega_n$.

(associativity) The complex numbers are associative under multiplication; since $\Omega_n \subseteq \mathbb{C}$, the elements of Ω_n are also associative under multiplication.

(identity) The multiplicative identity in $\mathbb{C}$ is 1. This is certainly an element of Ω_n, since $1^n = 1$ for all $n \in \mathbb{N}^+$.

(inverses) Let $x \in \Omega_n$; you will show in Exercise 2.83 that $x^{-1} \in \Omega_n$.

(cyclic) Theorem 2.75 tells us that $\omega \in \Omega_n$; the remaining elements are powers of ω. Hence $\Omega_n = \langle \omega \rangle$.

□

Combined with the explanation we gave earlier of the complex plane, Theorem 2.77 gives us a wonderful symmetry for the roots of unity.

Example 2.78. We'll consider the case where $n = 7$. According to the theorem, the 7th roots of unity are $\Omega_7 = \{1, \omega, \omega^2, \ldots, \omega^6\}$ where

$$\omega = \cos\left(\frac{2\pi}{7}\right) + i \sin\left(\frac{2\pi}{7}\right).$$

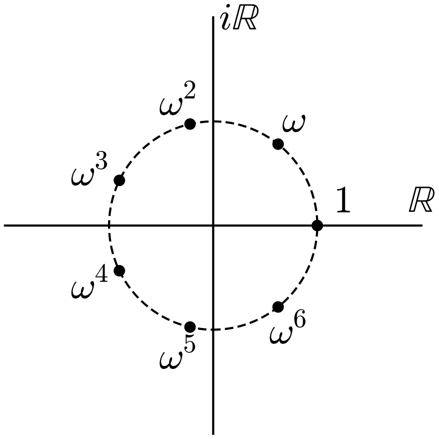


Figure 2.9. The seventh roots of unity, on the complex plane

According to Lemma 2.76,

$$\omega^m = \cos\left(\frac{2\pi m}{7}\right) + i\sin\left(\frac{2\pi m}{7}\right),$$

where $m = 0, 1, \ldots, 6$. By substitution, the angles we are looking at are

$$0, \frac{2\pi}{7}, \frac{4\pi}{7}, \frac{6\pi}{7}, \frac{8\pi}{7}, \frac{10\pi}{7}, \frac{12\pi}{7}.$$

Recall that in the complex plane, any complex number $a + bi$ corresponds to the point (a, b) on $\mathbb{R}^2$. The Pythagorean identity $\cos^2\alpha + \sin^2\alpha = 1$ tells us that the coordinates of the roots of unity lie on the unit circle. Since the angles are at equal intervals, they divide the unit circle into seven equal arcs! See Figure 2.9.

Although we used $n = 7$ in this example, we used no special properties of that number in the argument. That tells us that this property is true for any n: the nth roots of unity divide the unit circle of the complex plane into n equal arcs!

Here's an interesting question: is ω is the only generator of Ω_n? In fact, no. A natural follow-up: are *all* the elements of Ω_n generators of the group? Likewise, no. Well, which ones are? We are not yet ready to give a precise criterion that signals which elements generate Ω_n, but they do have a special name.

Definition 2.79. We call any generator of Ω_n a **primitive nth root of unity**.

Exercises.

Unless stated otherwise, $n \in \mathbb{N}^+$ and ω is a primitive n-th root of unity.

Exercise 2.80. Show that $\mathbb{C}$ is a group under addition.

Exercise 2.81.

(a) Find all the primitive square roots of unity, all the primitive cube roots of unity, and all the primitive quartic (fourth) roots of unity.

(b) Sketch *all* the square roots of unity on a complex plane. (Not just the primitive ones, but all.) Repeat for the cube and quartic roots of unity, each on a separate plane.

(c) Are any cube roots of unity *not* primitive? what about quartic roots of unity?

Exercise 2.82.

(a) Suppose that a and b are both positive powers of ω. Adapt Lemma 2.76 to show that ab is also a power of ω.

(b) Explain why this shows that Ω_n is closed under multiplication.

Exercise 2.83.

(a) Let ω be a 14th root of unity; let $\alpha = \omega^5$, and $\beta = \omega^{14-5} = \omega^9$. Show that $\alpha\beta = 1$.

(b) More generally, let ω be a primitive n-th root of unity, Let $\alpha = \omega^a$, where $a \in \mathbb{N}$ and $a < n$. Show that $\beta = \omega^{n-a}$ satisfies $\alpha\beta = 1$.

(c) Explain why this shows that every element of Ω_n has an inverse.

Exercise 2.84. Suppose β is a root of $x^n - b$.

(a) Show that $\omega\beta$ is also a root of $x^n - b$, where ω is *any* nth root of unity.

(b) Use (a) and the idea of unique factorization that we described right before the proof of Theorem 2.77 to explain how we can use β and Ω_n to list all n roots of $x^n - b$.

Exercise 2.85.

(a) For each $\omega \in \Omega_6$, find $x, y \in \mathbb{R}$ such that $\omega = x + yi$. Plot all the points (x, y) on a graph.

(b) Do you notice any pattern to the points? If not, repeat part (a) for Ω_7, Ω_8, etc., until you see the pattern.

Exercise 2.86.

(a) Show that $\mathbb{C}$ satisfies the requirements of a vector space for scalar multiplication.

(b) Show that $\mathbb{C}$ and $\mathbb{R}^2$ are isomorphic as monoids under addition.