Abstract Algebra

C. Caruvana, Ph.D.

Anno Accademico MMXXV

Contents

Preface & Prerequisite Comments **v**
Comment on Assumed Knowledge v

I Groups **1**

1 Brief History and a Motivating Example **3**

2 Brief Commentary on Mathematical Proofs **5**
2.1 Direct Proof 5
2.2 Proof by Contrapositive 6
2.3 Proof by Contradiction 6

3 Sets, Relations, and Functions **7**
3.1 Sets 7
3.2 Operations with Sets 8
3.3 Cartesian Products, Relations, and Functions 10
3.4 Equivalence Relations 13

4 Introducing Groups **15**
4.1 Getting to Know Groups 15
4.2 Subgroups 22
4.3 An Introduction to Homomorphisms 23
4.4 Cyclic Groups and Subgroups 26

5 Permutation Groups **33**
5.1 Permutations and Symmetric Groups 33
5.2 Dihedral Groups 42
5.3 Cayley's Theorem 43

6 Cosets and Lagrange's Theorem **45**
6.1 Cosets 45
6.2 Lagrange's Theorem 48

7 Creating New Groups From Old **51**
7.1 Direct Products . . . 51
7.2 Normal Subgroups and Quotient Groups . . . 56
7.2.1 Normal Subgroups . . . 56
7.2.2 Quotient Groups . . . 57
7.2.3 Isomorphism Theorems . . . 59

8 Classifying Finite Abelian Groups **67**

II Rings & Fields **75**

9 Introducing Rings **77**
9.1 Rings . . . 77
9.2 Ring Homomorphisms . . . 95
9.3 Ideals . . . 96
9.4 Prime and Maximal Ideals . . . 100

10 Polynomial Rings **103**
10.1 Polynomials Over a Ring . . . 103
10.2 The Polynomial Division Algorithm . . . 113
10.3 Irreducible Polynomials . . . 119

11 More on Integral Domains **125**
11.1 Factorization in Integral Domains . . . 125
11.2 Unique Factorization Domains . . . 129

Preface & Prerequisite Comments

These notes were prepared for the delivery of a course in Modern Algebra at Indiana University Kokomo for the 2025 academic year following Tom Judson's *Abstract Algebra: Theory and Applications*. For some select topics, additional inspiration was taken from Joseph Gallian's *Contemporary Abstract Algebra*. Any errors occurring within these notes are due to the present author.

Comment on Assumed Knowledge

Throughout these notes, we will assume the reader is familiar with the division algorithm for positive integers and fundamental properties of the determinant for square matrices. For reference, we list the particular facts that will be used throughout.

The Division Algorithm. For any integer n and any positive integer d, there are unique integers q and r with $0 \leq r < d$ such that $n = dq + r$.

Fact. For square matrices A and B, $\det(AB) = \det(A) \cdot \det(B)$.

Part I

Groups

Chapter 1

Brief History and a Motivating Example

The following historical commentary has been gathered from the Wikipedia articles on History of Algebra and Abstract Algebra.

The word *algebra* is derived from the Arabic *al-jabr* which appeared in the title of a treatise written in 830 by Al-Khwarizmi, a Persian mathematician. The treatise itself was about linear and quadratic equations. Note, however, that societies all around the world had independently developed their own studies of solving algebraic equations well before Al-Khwarizmi's treatise.

For the ancient Babylonians and the Greeks, algebraic concepts were largely geometric. In fact, Greek and Vedic Indian mathematicians used geometry to solve certain algebraic equations. Much later, Descartes (1596-1650) introduced modern notation and showed that problems of geometry can be expressed and solved in terms of algebra.

Diophantus and Brahmagupta, independently, moved away from the geometric perspective toward one focused on finding numbers that satisfied given equations.

Sharaf al-Din al-Tusi began the transition from static equation solving to a more functional approach, where functions are seen as dynamic entities representing motion.

It is not until the 19th and 20th centuries that *abstract algebra* is developed.

Abstract algebra, which we understand as the study of "algebraic structure," emerged through the identification of common themes in problems of number theory, geometry, analysis, and algebraic equation solving. In this semester, we will be focusing on a branch of abstract algebra known as *group theory*. Loosely speaking, a group is a collection of objects with a particular kind of operation. We will define them carefully after considering some motivating examples and setting up our mathematical formalism.

Example 1.0.1. Consider, for example, the possible symmetric positions of a line segment:

There are two possible actions on this structure: "do nothing" or reversing the poles. Note that reversing the poles twice has the same effect as "doing nothing."

Another way to model this example is to use "words." Let P represent the "positive" side and N represent the "negative" side. As presented, the word PN represents the initial

orientation. Reversing the poles yields the word NP. Thus, the possible states are thus PN and NP.

We can also abstract away the "actions" here of *doing nothing* and *reversing the poles*. Let 0 represent the action of "doing nothing" and 1 represent the action of "reversing the poles." In the following table, consider the rows as indicating which of the two actions to be applied first, and then the columns indicate which operation is done afterwards.

	0	1
0	0	1
1	1	0

As noted above and displayed in the second row, second column of the table, reversing the poles twice results in the same action as the "doing nothing" action. ⊣

Chapter 2

Brief Commentary on Mathematical Proofs

One of the primary activities in this course is proof-writing. Mathematical proofs are formal arguments that establish the logical necessity of a given statement. Generally speaking, a *statement* is any expression which has a truth value. For example, "2 is even" is a statement (and is true) and "$12 + 5 = 60$" is a statement (which is false). An expression like $x + 5$ is not a statement because there is not a coherent way to assign a truth value to it.

The general structure of discourse here consists of two things: there are the objects of discourse and there are statements referring to the objects.

Of primary importance are conditional statements of the form "if p, then q," where p and q are statements themselves. We will also deal with quantified statements:

- $\forall x\ P(x)$ is the statement "for every x, x satisfies P," and
- $\exists x\ P(x)$ is the statement "there exists an x such that x satisfies P,."

2.1 Direct Proof

A standard mathematical proof proving that the implication "if p, then q" is true starts by assuming p and deducing the logical necessity of the truth of q. We begin with a very basic example, introducing the conventional proof-writing format in the process.

Claim. Suppose x represents a real number. If $3x + 2 = 17$, then $x = 5$.

Proof. Suppose $3x + 2 = 17$. By subtracting 2 from both sides of the equation, we obtain that $3x = 15$. Then, dividing by 3 on both sides of the equation yields that $x = 5$. □

Formal proofs are written in a narrative fashion using the expected natural language, sometimes assisted by some mathematical symbols. Like anything else, proof-writing is a skill that improves with practice.

2.2 Proof by Contrapositive

Another valid proof technique for proving "if p, then q," is to prove what is known as the corresponding *contrapositive* statement: "if it's not true that q, then it's not true that p."

Let's see a basic example.

Claim. Suppose x represents a real number. If $x^3 - 1 < 0$, then $x < 1$.

Proof. We proceed by way of the contrapositive. So, suppose $x \geq 1$. It follows that $x^3 \geq 1$ and, then, that $x^3 - 1 \geq 0$. □

Sanity Check 2.2.1. If x is a real number and $x \geq 1$, why is it that $x^3 \geq 1$?

Exercise 2.2.2. An integer n is said to be *even* if there is an integer k such that $n = 2k$. In other words, n is even if 2 divides n with no remainder. Prove that, given an integer n, if n^2 is even, then n is even.

2.3 Proof by Contradiction

Proof by contradiction, also known as *reductio ad absurdum*, is an ancient argument style used often in Platonic dialogues. The basic format is this: First you pose the objection, "suppose what you claim to be true is false." You then attempt to deduct a logical impossibility. The standard introductory proof by contradiction (known certainly to the ancient Greeks) is that of the irrationality of $\sqrt{2}$.

Claim. The number $\sqrt{2}$ is not rational.

Proof. By way of contradiction, suppose that $\sqrt{2}$ is rational. By taking out any common factors, we can thus write $\sqrt{2} = \frac{p}{q}$ where p and q are integers with no common factors. Squaring both sides, we obtain that $2 = \frac{p^2}{q^2}$ and, thus, that $2q^2 = p^2$. It follows that p^2 is even and, thus, by Exercise 2.2.2, p is even. That is, $p = 2k$ for some integer k. It follows that

$$\begin{aligned} 2q^2 &= (2k)^2 \\ &= 4k^2. \end{aligned}$$

Dividing by 2 yields $q^2 = 2k^2$. Hence, q^2 is even which, again, by Exercise 2.2.2, asserts that q is even. We now see that p and q are both even, contradicting the assumption that p and q had no common factors. Therefore, $\sqrt{2}$ is irrational. □

Sanity Check 2.3.1. Why can we write ratios of integers in so-called reduced terms? (*Hint.* See the *Fundamental Theorem of Arithmetic*.)

Chapter 3

Sets, Relations, and Functions

The theory of sets provides us with a twofold benefit: they offer us a foundation on which to formalize our mathematics and they off us a relatively simple structure with which to continue practicing the basics of proof-writing.

3.1 Sets

A *set* is a well-defined collection of primitive objects. In the case of pure set theory, the *only* primitive objects are sets, themselves. In the language of set theory, we use the symbol $\in$ as a relation between an object (it can be a set) x and a set A in the following way: $x \in A$ is the statement that x is an element of the set A. For example, if we say that X is the set of all positive real numbers, then $1 \in X$ and $-5 \notin X$.

Guided by convention, we use $\mathbb{R}$ to refer to the set of all real numbers, $\mathbb{Z}$ to refer to the set of all integers, $\mathbb{Q}$ to refer to the set of all rational numbers, and $\mathbb{N}$ to refer to the set of all natural numbers. Not all mathematicians agree whether $\mathbb{N}$ contains 0 or not. To align with our chosen textbook, we will observe the convention that $\mathbb{N}$ consists of only the positive integers.

One can define (finite) or refer to sets using *set-roster* notation; e.g. $X = \{a, b, c\}$. Here, X is asserted to be the set containing as its elements a, b, and c.

The most common method to define or refer to sets is the *set-builder* notation;

$$X = \{x \in \mathbb{R} : x > 0\}.$$

Here, X is asserted to be the set of positive real numbers. The expression $\{x \in \mathbb{R} : x > 0\}$ can be read as:

> ➢ "The set of real numbers x such that x is positive."

Another set that will be important to us is the set $\mathbb{C}$ consisting of all complex numbers; that is, $\mathbb{C} = \{x + iy : x, y \in \mathbb{R}\}$ where i is chosen to be a solution to the equation $x^2 + 1 = 0$.

Note that our expression for $\mathbb{C}$ above is not of the form $\{x \in X : P(x)\}$ where X is a set and P is a *predicate* of x. We won't concern ourselves here with the details of this notation and will simply accept it as a valid incarnation of set-builder notation.

Definition 3.1.1. Two sets A and B are said to be *equal*, denoted $A = B$, if they consist of exactly the same elements.

Definition 3.1.2. For two sets A and B, we say that B is a *subset* of A, denoted by $B \subseteq A$, if every element of B is an element of A. If $B \subseteq A$ and $B \neq A$, then we say that B is a *proper subset* of A, which we may denote by $B \subsetneq A$.[1]

Pro Tip. Two sets A and B are equal if and only if $A \subseteq B$ and $B \subseteq A$. Hence, a common strategy for proving that two sets are equal is to prove that they are both subsets of the other.

Definition 3.1.3. The *empty set*, denoted by either $\emptyset$ or $\{\}$, is the unique set containing no elements.

3.2 Operations with Sets

We can form new sets from old with some basic operations.

Definition 3.2.1. Given two sets A and B, the *union* of A and B, denoted by $A \cup B$, is the set consisting of all x such that either $x \in A$ or $x \in B$.

Comment. In mathematics, the "or" is always the *inclusive* or. So, a phrase of the form "p or q" is only to be interpreted as meaning "p, or q, or both." We will address one way to deal with the exclusive or below.

Definition 3.2.2. Given two sets A and B, the *intersection* of A and B, denoted by $A \cap B$, is the set consisting of all x such that $x \in A$ and $x \in B$.

Exercise 3.2.1. Prove that, for sets A and B,

- $A \cup B = B \cup A$ and
- $A \cap B = B \cap A$.[2]

Exercise 3.2.2. Prove that, for sets A, B, and C,

- $A \cup (B \cup C) = (A \cup B) \cup C$ and
- $A \cap (B \cap C) = (A \cap B) \cap C$.[3]

Definition 3.2.3. Given two sets A and B, we define the *set difference* $A \setminus B$ to be

$$\{x \in A : x \notin B\}.$$

When a given context U (referred to as the *universal set*) is understood, we define the *complement* of a set $A \subseteq U$ to be $A' = U \setminus A$.

[1] Our chosen textbook uses the notation $\subset$ in place of $\subseteq$. As such, we must avoid using the visual relationships between $\leq$ and $\subseteq$, and $<$ and $\subset$ here.

[2] That is, the union and intersection operations are *commutative*.

[3] That is, the union and intersection operations are *associative*.

The *symmetric difference* between two sets exemplifies the concept of the exclusive or. The standard convention for the symmetric difference is as follows:

$$A \triangle B = (A \cup B) \setminus (A \cap B).$$

We will return to this later.

Definition 3.2.4. Two sets A and B are said to be *disjoint* if $A \cap B = \emptyset$.

The operations of union and intersection extend well beyond contexts where only two sets are involved.

Definition 3.2.5. Suppose $\mathcal{A}$ is a set of sets (that is, each $A \in \mathcal{A}$ is a set). We define

$$\bigcup \mathcal{A} = \bigcup_{A \in \mathcal{A}} A = \{x : \exists A \in \mathcal{A}\ (x \in A)\}.$$

Similarly, we define

$$\bigcap \mathcal{A} = \bigcap_{A \in \mathcal{A}} A = \{x : \forall A \in \mathcal{A}\ (x \in A)\}.$$

These will often show up over *countable* sets; e.g. if A_n is a set for each $n \in \mathbb{N}$, then

$$\bigcup_{n=1}^{\infty} A_n = \bigcup_{n \in \mathbb{N}} A_n = \{x : \exists n \in \mathbb{N}\ (x \in A_n)\}.$$

Proposition 3.2.6. For any set A,

- $A \cup A = A \cap A = A \cup \emptyset = A$ and
- $A \setminus A = A \cap \emptyset = \emptyset$.

Proposition 3.2.7 (Distributive Properties)**.** For sets A, B, and C,

- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and
- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

Theorem 3.2.8 (DeMorgan's Laws)**.** For any two subsets A and B of a given universal set U,

- $(A \cap B)' = A' \cup B'$ and
- $(A \cup B)' = A' \cap B'$.

Exercise 3.2.3. Prove Theorem 3.2.8.

3.3 Cartesian Products, Relations, and Functions

Cartesian products inherit their name from Descartes.

Definition 3.3.1. Given sets X and Y, the *Cartesian product* of X with Y is

$$X \times Y = \{(x, y) : x \in X, y \in Y\}.$$

For example, the usual coordinate plane is $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$.

Definition 3.3.2. For sets X and Y, a *relation* between X and Y is a subset of $X \times Y$. When $R \subseteq X \times Y$, we sometimes use the notation xRy to mean $(x, y) \in R$.

Definition 3.3.3. Given two sets X and Y, a *function* f with *domain* X and *codomain* Y is a relation between X and Y with the following properties:

- For every $x \in X$, there is some $y \in Y$ such that $(x, y) \in f$.[4]
- For every $x \in X$ and $y, z \in Y$, if $(x, y) \in f$ and $(x, z) \in f$, then $y = z$.[5]

We use the notation $f : X \to Y$ to denote that f is a function from X to Y. We will also use the notation $f(x)$ to denote the unique member of Y such that $(x, f(x)) \in f$. The *image* or *range* of a function $f : X \to Y$ is defined to be $f[X] = \{f(x) : x \in X\}$.[6]

Comment. We will say that two functions f and g are *equal*, denoted $f = g$, if they are equal *as sets*.

Example 3.3.4. The relation

$$R = \{(x, y) \in \mathbb{R}^2 : x^2 = y^4\}$$

is not a function. Consider the fact that both $(1, 1) \in R$ and $(1, -1) \in R$. ⊣

You may recall exercises from previous courses with the flavor of "rewrite *blah* as a function of *blech*" which involved algebraic manipulation of expressions. For example, in the equation $x = y^2$, x can be seen as a function of y, but y *cannot* be seen as a function of x. In these kinds examples, the issue of the domain tends to be ignored due to course focus, and they are typically intervals of real numbers, anyway.

We would like to point out here, though, that there are situations in which an equation can express y as a function of x, but cannot be rewritten using familiar algebraic tools in the form $y = f(x)$.

[4] In some contexts, *partial* functions are used and only have to satisfy the second property listed here (that is, they need not have full domain).

[5] This is formally stating that there is a unique element $y \in Y$ which satisfies $(x, y) \in f$. In less formal language, this is a version of the Vertical Line Test.

[6] Note the divergence in the notation here from the chosen textbook's. Either is acceptable, though it is sometimes useful to distinguish between a function being applied to particular elements of its domain and sets consisting of such applications.

Exercise 3.3.1. Show that the relation

$$R = \{(x, y) \in \mathbb{R}^2 : x = y + y^5\}$$

is a function. (*Hint.* You can use techniques from Calculus to show that $y = x + x^5$ is a bijection. Then, by the discussion below, it must have an inverse function, and that inverse function is exactly R.)

Before we get to function inverses, we will need the notion of *composition*.

Definition 3.3.5. Suppose $R \subseteq X \times Y$ and $S \subseteq Y \times Z$. The *composition* relation $S \circ R \subseteq X \times Z$ is defined to be

$$S \circ R = \{(x, z) \in X \times Z : \exists y \in Y \ ((x, y) \in R, (y, z) \in S)\}.$$

Exercise 3.3.2. Suppose $f : X \to Y$ and $g : Y \to Z$. Show that the composition $g \circ f$ is a function $X \to Z$.

Since relations serve as a broader context than functions, we start by defining the *inverse* of a relation, which always exists and is, itself, a relation.

Definition 3.3.6. Given $R \subseteq X \times Y$, we define the *inverse relation* of R to be

$$R^{-1} = \{(y, x) \in Y \times X : (x, y) \in R\}.$$

Under what conditions is R^{-1} guaranteed to be a function from Y to X? Examining Definition 3.3.3 in this context should naturally draw one to the following notions.

Definition 3.3.7. Suppose $f : X \to Y$ is a function. Then

- f is *injective* or *one-to-one* provided that, for $x_1, x_2 \in X$, if $f(x_1) = f(x_2)$, then $x_1 = x_2$.[7]
- f is *surjective* or *onto* if, for every $y \in Y$, there is some $x \in X$ such that $f(x) = y$.
- f is *bijective* if it is both injective and surjective.

Definition 3.3.8. For any space X, we define the *identity* map $\mathrm{id}_X : X \to X$ by the rule $\mathrm{id}_X(x) = x$. Note that id_X is a bijection.

Though the following does not align yet with Definition 3.3.6, we present it in this incarnation for the sake of familiarity.

Definition 3.3.9. Suppose $f : X \to Y$ and $g : Y \to X$. The function g is said to be the *inverse* function of f, denoted by f^{-1}, if $g \circ f = \mathrm{id}_X$ and $f \circ g = \mathrm{id}_Y$. In the case that a function $f : X \to Y$ has an inverse function, f is said to be *invertible*.

Theorem 3.3.10. A function is invertible if and only if it is bijective.

[7] This is a version of the Horizontal Line Test.

In other words, a function is invertible if and only if its inverse relation is, itself, a function.

Example 3.3.11. Consider $S : \mathbb{R} \to [0, \infty)$ defined by $S(x) = x^2$. Note that S is surjective but not injective. Hence, S is not invertible.

Note, however, that the standard square root function $\text{sqrt} : [0, \infty) \to [0, \infty)$, which chooses, for each $y \geq 0$, the *non-negative* solution to $x^2 = y$, is a function with the property that $S \circ \text{sqrt} = \text{id}_{[0,\infty)}$. On the other hand, for any $x \in \mathbb{R}$, $\text{sqrt} \circ S(x) = |x|$, the absolute value of x. ⊣

Exercise 3.3.3. Define $f : \mathbb{Z} \times \mathbb{N} \to \mathbb{Q}$ by the rule $f(k, n) = \dfrac{k}{n}$. Show that f is surjective but not injective.

Even when we restrict our attention to invertible functions, the composition operation fails to be commutative.

Example 3.3.12. Consider $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x + 2$ and $g : \mathbb{R} \to \mathbb{R}$ defined by $g(x) = x^3$. Note that

$$g \circ f(x) = (x + 2)^3$$

and that

$$f \circ g(x) = x^3 + 2.$$

Since $g \circ f(0) = 8$ and $f \circ g(0) = 2$, we see that $g \circ f \neq f \circ g$. ⊣

Exercise 3.3.4. Find a pair of functions $f, g : \mathbb{R} \to \mathbb{R}$ that aren't inverses of each other such that neither is the identity function $\text{id}_{\mathbb{R}}$ and the equation $g \circ f = f \circ g$ is satisfied.

Exercise 3.3.5. Suppose $f : W \to X$, $g : X \to Y$, and $h : Y \to Z$. Prove that composition is associative; that is, show that $(h \circ g) \circ f = h \circ (g \circ f)$.

Exercise 3.3.6. Suppose $f : X \to Y$ and $g : Y \to Z$. Prove each of the following.

(a) If f and g are both injective, then so is $g \circ f$.

(b) If f and g are both surjective, then so is $g \circ f$.

(c) If f and g are both bijections, then so is $g \circ f$.

Exercise 3.3.7. Consider the set $X = \{0, 1\}$ consisting of two elements. List all of the bijections of X. Can you see any similarity here with Example 1.0.1?

Another important concept related to functions is that of pre-images or fibers, though the standard notation overloads the notation for function inverses.

Definition 3.3.13. For a function $f : X \to Y$, the *pre-image* or *fiber* of a point $y \in Y$ is defined to be

$$f^{-1}(y) = \{x \in X : f(x) = y\}.$$

So a function is invertible if and only if every fiber consists of exactly one element of the domain.

Sanity Check 3.3.8. When f is invertible, what is the discrepancy between $f^{-1}(y)$ used as the *inverse function value at* y and $f^{-1}(y)$ used as the *fiber of* f *at* y?

3.4 Equivalence Relations

Definition 3.4.1. For a set X, a relation $\simeq \subseteq X \times X$ is said to be an *equivalence relation* if the following three properties hold:

- (Reflexivity) For every $x \in X$, $x \simeq x$.
- (Symmetry) For every $x, y \in X$, if $x \simeq y$, then $y \simeq x$.
- (Transitivity) For every $x, y, z \in X$, if both $x \simeq y$ and $y \simeq z$, then $x \simeq z$.

In this case, we will often use the phrasing "$\simeq$ is an equivalence relation *on* X."

The equality relation itself is an equivalence relation.

Exercise 3.4.1. Show that, if $R \subseteq X \times X$ is an equivalence relation, then

- $R^{-1} = R$ and
- $R \circ R = R$.

Exercise 3.4.2. Suppose $f : X \to Y$ and define $\simeq$ on X by the following rule: $x \simeq y$ provided that $f(x) = f(y)$. Show that $\simeq$ is an equivalence relation.

Exercise 3.4.3. Let $X = \mathbb{Z} \times \mathbb{N}$ and define $(k, m) \simeq (\ell, n)$ by

$$kn = \ell m.$$

Show that $\simeq$ is an equivalence relation.[8]

Definition 3.4.2. Suppose $\simeq$ is an equivalence relation on a set X. For $x \in X$, we define the $\simeq$-*equivalence class* (or simply called the *equivalence class* when there is no possibility for confusion) to be

$$[x]_{\simeq} = \{y \in X : x \simeq y\}.$$

When the equivalence relation $\simeq$ is understood, we may suppress the subscript and refer to the equivalence class of x as $[x]$.

Comment. Using the notion of equivalence classes, we can formally define the rational numbers to be the set of equivalence classes of the $\simeq$ relation defined in Exercise 3.4.3.

Exercise 3.4.4. Suppose $\simeq$ if an equivalence relation on a set X. Show that, for any $x, y \in X$, either $[x] = [y]$ or $[x] \cap [y] = \emptyset$.

Example 3.4.3. Let $n \geq 2$, $n \in \mathbb{Z}$, and define, for $p, q \in \mathbb{Z}$, $p \equiv q \pmod{n}$ if there exists $k \in \mathbb{Z}$ with $p - q = nk$. Then $\cdot \equiv \cdot \pmod{n}$ is an equivalence relation. ⊣

[8] In fact, it is the standard notion of equivalence for rational numbers.

Proof. For $p \in \mathbb{Z}$, note that $p - p = 0 = n \cdot 0$. Since $0 \in \mathbb{Z}$, $p \equiv p \pmod{n}$.

For symmetry, suppose $p \equiv q \pmod{n}$. By definition, that means there is some $k \in \mathbb{Z}$ for which $p - q = nk$. Note that $q - p = -nk = n \cdot (-k)$. Since $-k \in \mathbb{Z}$, we see that $q \equiv p \pmod{n}$.

Finally, for transitivity, suppose $p \equiv q \pmod{n}$ and $q \equiv r \pmod{n}$. Then let $k, \ell \in \mathbb{Z}$ be such that $p - q = nk$ and $q - r = n\ell$. Observe that

$$\begin{aligned} p - r &= p - q + q - r \\ &= nk + n\ell \\ &= n(k + \ell). \end{aligned}$$

Since $k + \ell \in \mathbb{Z}$, we see that $p \equiv r \pmod{n}$, finishing the proof. □

For each $n \geq 2$, $n \in \mathbb{Z}$, we define $\mathbb{Z}_n$ to be the set of equivalence classes for the equivalence relation $\cdot \equiv \cdot \pmod{n}$. We will also use the notation $k \,\%\, n$ to be the unique integer in the interval $[0, n)$ such that $k \equiv (k \,\%\, n) \pmod{n}$.

Exercise 3.4.5. Consider $\mathbb{Z}_2$.

(a) Show that,

- for any even $n \in \mathbb{Z}$, $[n] = [0]$ and
- for any odd $n \in \mathbb{Z}$, $[n] = [1]$.

Conclude that $\mathbb{Z}_2$ consists of exactly two elements.

(b) Consider addition in $\mathbb{Z}_2$ to be defined as $[x] + [y] = [x + y]$. Can you see any similarity here with Example 1.0.1?

Let $\mathbb{Z}^{\mathbb{Z}}$ denote the set of all functions $\mathbb{Z} \to \mathbb{Z}$.[9] For each $k \in \mathbb{Z}$, consider the function $g_k : \mathbb{Z} \to \mathbb{Z}$ defined by $g_k(n) = n + k$. Note that each g_k is a bijection and that $g_0 = \mathrm{id}_{\mathbb{Z}}$. We can also see each g_k as a shifting of $\mathbb{Z}$. For example, if we envision $\mathbb{Z}$ on a horizontal line ordered in increasing order, g_3 has the effect of shifting all points over by three units to the right, g_0 has the effect of doing nothing, and g_{-5} has the effect of shifting all points over by four units to the left.

Given the symbols above, we can define $\varphi : \mathbb{Z} \to \mathbb{Z}$ by $\varphi(k) = g_k$.[10] We will return to this idea later.

[9] There is a connection here with Cartesian products, but we won't get into those details now.

[10] Formally speaking, g is already a function from $\mathbb{Z}$ to $\mathbb{Z}^{\mathbb{Z}}$.

Chapter 4

Introducing Groups

4.1 Getting to Know Groups

Definition 4.1.1. Let G be a set and $\diamond : G \times G \to G$ be a function, also referred to as a *binary operation*. The set G with the binary operation $\diamond$, $(G, \diamond)$, is said to be a *group* if the following properties hold:

- (Associativity) For any $a, b, c \in G$, $a \diamond (b \diamond c) = (a \diamond b) \diamond c$.
- (Identity) There exists $e \in G$ such that, for every $g \in G$, $g \diamond e = e \diamond g = g$. Any such $e \in G$ is referred to as the[1] *identity*.
- (Inverses) For any $g \in G$, there is some $h \in G$ such that $g \diamond h = h \diamond g = e$, where e is the identity. As will be shown in Proposition 4.1.2, inverse elements are unique so we will use g^{-1} to denote the *inverse* of g.

Sanity Check 4.1.1. Suppose $(G, \diamond)$ is a group. Why is the identity $e \in G$ unique? (*Hint.* Suppose $e_1, e_2 \in G$ satisfy the identity condition for the group. Show that $e_1 = e_2$.)

Proposition 4.1.2. In any group, the inverse of a given element of the group is unique.

Proof. Let $(G, \diamond)$ be a group with identity e. Suppose $g \in G$ and $h_1, h_2 \in G$ are such that $g \diamond h_1 = h_1 \diamond g = e$ and $g \diamond h_2 = h_2 \diamond g = e$. Observe that

$$\begin{aligned} h_1 &= h_1 \diamond e \\ &= h_1 \diamond (g \diamond h_2) \\ &= (h_1 \diamond g) \diamond h_2 \\ &= e \diamond h_2 \\ &= h_2. \end{aligned}$$

Therefore, $h_1 = h_2$. □

[1] As will be shown in Sanity Check 4.1.1, there is only one element that satisfies the identity criterion.

A common practice in algebra is to use juxtaposition as the *multiplicative* operation of the group. In particular, we may say that $(G, \cdot)$ is a group and let $gh = g \cdot h$ for $g, h \in G$. We have chosen in this section to use $\diamond$ as the binary operation in many cases to help the reader dissociate from previous experience with common binary operations like $+$ and $\cdot$, though we will also employ the multiplicative convention where convenient. The motivation for using $\cdot$ over $+$ in the general setting is because of the association of $+$ with commutativity (Definition 4.1.4). If one recalls that $AB \neq BA$ in general for square matrices A and B, then one can hopefully appreciate the fact that $\cdot$ need not always commute, in general.

Notation. The $\cdot$ convention also inspires *exponent* notation. For a group $(G, \cdot)$, we will consider $g^0 = e$, the identity, and $g^1 = g$. Then, for $n \in \mathbb{N}$, assuming g^n has been defined, we set $g^{n+1} = g^n g$. We also define $g^{-n} = (g^n)^{-1}$.

Exercise 4.1.2. Let $(G, \cdot)$ be a group. Show that, for $g \in G$, $(g^{-1})^{-1} = g$.

Exercise 4.1.3. Let $(G, \cdot)$ be a group and define $\varphi : G \to G$ by $\varphi(g) = g^{-1}$. Show that φ is a bijection.

Exercise 4.1.4. Let $(G, \cdot)$ be a group and $g, h \in G$. Show that $(gh)^{-1} = h^{-1}g^{-1}$.

Exercise 4.1.5 (Left- and Right-cancellation Laws)**.** Let $(G, \cdot)$ be a group and $a, g, h \in G$. Show that

- $ag = ah \implies g = h$ and
- $ga = ha \implies g = h$.

Exercise 4.1.6. Let $(G, \cdot)$ be a group and $g, h \in G$. For $n, m \in \mathbb{Z}$, show that

- $g^n g^m = g^{n+m}$ and
- $(g^n)^m = g^{nm}$.

Exercise 4.1.7. Let $(G, \diamond)$ be a group and let $p \in G$. Define $\varphi : G \to G$ by $\varphi(g) = g \diamond p$. Show that φ is a bijection and determine the inverse function of φ.

Example 4.1.3. The integers $\mathbb{Z}$ with their usual binary operation of addition $+$ forms a group. ⊣

Definition 4.1.4. If a group $(G, \diamond)$ satisfies the property that $g \diamond h = h \diamond g$ for all $g, h \in G$, then G is said to be *Abelian* or *commutative*.

Note that the integer group $(\mathbb{Z}, +)$ is Abelian.

Exercise 4.1.8. Suppose $(G, \cdot)$ is an Abelian group. Show that, for any $g, h \in G$ and $n \in \mathbb{Z}$, $(gh)^n = g^n h^n$.

Example 4.1.5. Let $\mathrm{Sym}(\mathbb{R})$ consist of all bijections $\mathbb{R} \to \mathbb{R}$. With the operation of function composition $\circ$, $(\mathrm{Sym}(\mathbb{R}), \circ)$ is a group with identity $\mathrm{id}_{\mathbb{R}}$. By Example 3.3.12, this group is not commutative. ⊣

Example 4.1.6. Let $\mathbb{R}^+ = \{x \in \mathbb{R} : x > 0\}$. Then $(\mathbb{R}^+, \cdot)$, where $\cdot$ is the standard multiplication operation, is an Abelian group with identity 1. ⊣

Note that Examples 4.1.5 and 4.1.6 can justify the overloading of the $\cdot^{-1}$ operator seen in, for example, Calculus courses. In some cases, it is used to refer to the functional inverse, and in other cases, it is used to refer to the multiplicative inverse.

Exercise 4.1.9. Show that $(\mathbb{R}, \cdot)$, where $\cdot$ is multiplication, is not a group.

Exercise 4.1.10. Is $(\mathbb{N}, \cdot)$ a group? Why or why not?

Example 4.1.7. Expanding on Exercise 3.4.5, let $n \geq 2$, $n \in \mathbb{Z}$, and define $\oplus : \mathbb{Z}_n \times \mathbb{Z}_n \to \mathbb{Z}_n$ by the rule

$$[x] \oplus [y] = [x + y].$$

Then $(\mathbb{Z}_n, \oplus)$ is an Abelian group with identity $[0]$. As is common, we can identify each $[x]$ with the unique $0 \leq y < n$ for which $x \equiv y \pmod{n}$. In Figure 4.1, we provide the particular *Cayley table* for each of the groups $\mathbb{Z}_2$, $\mathbb{Z}_3$, and $\mathbb{Z}_5$, with the identifications specified above.

$\oplus$	0	1
0	0	1
1	1	0

$\oplus$	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

$\oplus$	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

Figure 4.1: Cayley tables for $(\mathbb{Z}_2, \oplus)$, $(\mathbb{Z}_3, \oplus)$, and $(\mathbb{Z}_5, \oplus)$, respectively.

⊣

Example 4.1.8. For $n \geq 2$, $n \in \mathbb{Z}$, define $* : \mathbb{Z}_n \times \mathbb{Z}_n \to \mathbb{Z}_n$ by the rule

$$[x] * [y] = [x \cdot y].$$

In Figure 4.2, we consider two Cayley tables for this operation in the contexts $n = 5, 6$.

Ignoring the zero rows and columns, we obtain the tables in Figure 4.3, highlighting zero entries in red. Note that $(\mathbb{Z}_5 \setminus \{0\}, *)$ forms a group but $(\mathbb{Z}_6 \setminus \{0\}, *)$ does not.

⊣

In fact, if we let $\mathbb{Z}_n^+ = \mathbb{Z}_n \setminus \{0\}$, then $(\mathbb{Z}_n^+, *)$ is a group if and only if n is prime. Recall that an integer $n \geq 2$ is said to be *composite* if there are integers $2 \leq a, b < n$ such that $ab = n$. If the integer $n \geq 2$ is not composite, it is *prime*.

Recall the standard notation of divisibility: if $n \in \mathbb{Z}$ and $d \in \mathbb{N}$, we write $d \mid n$ if there is an integer k such that $n = dk$.

A fact that we will use here without proof is

Theorem 4.1.9 (The Fundamental Theorem of Arithmetic)**.** Every integer greater than 1 has a unique prime factorization.

Using The Fundamental Theorem of Arithmentic, one can prove

Lemma 4.1.10 ("Euclid's Lemma")**.** If a prime p divides a product ab, then either $p \mid a$ or $p \mid b$.

Note that, by mathematical induction, Euclid's Lemma can be extended to the following assertion:

If a prime p divides a product $a_1 a_2 \cdots a_n$, then $p \mid a_j$ for some $j \in \{1, 2, \ldots, n\}$.

The following lemma can be seen as a rephrasing of Euclid's Lemma.

Lemma 4.1.11. If p is a prime number, then, for integers $0 \leq a, b < p$, if $ab \equiv 0 \pmod{p}$, then either $a = 0$ or $b = 0$.

Theorem 4.1.12. Let $p \geq 2$, $p \in \mathbb{Z}$, and consider $\mathbb{Z}_n^+$ with the operation $*$ defined in Example 4.1.8. Then $(\mathbb{Z}_p^+, *)$ is a group if and only if p is prime.

Proof. First note that, if p is composite, then we can write $p = ab$ for some $2 \leq a, b < p$. Note then that $ab \equiv 0 \pmod{p}$ and so $\mathbb{Z}_p^+$ is not closed under the operation $*$. Consequently, $(\mathbb{Z}_p^+, *)$ is not a group.

Now suppose p is prime. By Lemma 4.1.11, $\mathbb{Z}_p^+$ is closed under $*$. We also know that 1 satisfies the identity criterion and that multiplication is associative. So the only thing to show is that every element of $\mathbb{Z}_p^+$ has an inverse. So consider $2 \leq n < p$ and the set $\{n^k : k \in \mathbb{N}\}$. Note that $\{n^k : k \in \mathbb{N}\}$ is an infinite subset of natural numbers. However,

$*$	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

$*$	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	2	3	4	5
2	0	2	4	0	2	4
3	0	3	0	3	0	3
4	0	4	2	0	4	2
5	0	5	4	3	2	1

Figure 4.2: Cayley tables for $(\mathbb{Z}_5, *)$ and $(\mathbb{Z}_6, *)$, respectively.

$*$	1	2	3	4
1	1	2	3	4
2	2	4	1	3
3	3	1	4	2
4	4	3	2	1

$*$	1	2	3	4	5
1	1	2	3	4	5
2	2	4	0	2	4
3	3	0	3	0	3
4	4	2	0	4	2
5	5	4	3	2	1

Figure 4.3: Cayley tables for $(\mathbb{Z}_5 \setminus \{0\}, *)$ and $(\mathbb{Z}_6 \setminus \{0\}, *)$, respectively.

since $\mathbb{Z}_p^+$ is finite, $\{n^k \pmod p : k \in \mathbb{N}\}$ is also finite. Consequently, there must be some $j, k \in \mathbb{N}$, $j < k$, such that $n^j \equiv n^k \pmod p$. It follows that

$$p \mid n^j - n^k = n^j(1 - n^{k-j}).$$

Since $p \nmid n$, it must be the case that $p \mid 1 - n^{k-j}$, which is equivalent to $n^{k-j} \equiv 1 \pmod p$. Since $2 \leq n < p$, we must have that $k - j > 1$. Hence, $n \cdot n^{k-j-1} \equiv 1 \pmod p$. Therefore, n has a multiplicative inverse. □

We are now equipped to prove

Theorem 4.1.13 (Wilson's Theorem)**.** An integer $p > 1$ is prime if and only if

$$(p-1)! \equiv -1 \pmod p$$

where $x!$ is the factorial of x. Equivalently, the integer $p > 1$ is prime if and only if

$$\frac{(p-1)! + 1}{p}$$

is an integer.

Proof. [2] First, suppose p is composite. We proceed here by cases.

First, suppose $p > 4$. Then we write $p = ab$ where $2 \leq a, b < p$. Since $p > 4$, we can assume without loss of generality that $a \geq 3$. Note now that $2 \leq a - 1$ and $1 \leq b - 1$, and, thus,

$$2 \leq (a-1)(b-1) = ab - a - b + 1 \implies a + b \leq ab - 1 = p - 1.$$

It follows that

$$(p-1)! = 1 \cdot 2 \cdots a(a+1) \cdots (a+b) \cdots (ab-1).$$

In particular, there is an integer m for which

$$(p-1)! = a(a+1) \cdots (a+b)m.$$

Note that

$$b \mid (a+1)(a+2) \cdots (a+b)$$

[2] This proof is adapted from comments appearing in math.se/307 and math.se/164852.

since the right-hand expression is the product of b consecutive integers. Hence, there is an integer k for which

$$(a+1)(a+2)\cdots(a+b) = bk.$$

We can thus rewrite our equation above as

$$(p-1)! = abkm = pkm.$$

It follows that $(p-1)! \equiv 0 \pmod{p}$.

In the case that $p = 4$, note that $3! = 6 \equiv 2 \pmod{4}$. Since $2 \not\equiv -1 \pmod{4}$, we have finished this direction of the proof.

Finally, suppose p is prime. In this portion of the proof, we wish to show that the product $(p-1)!$ consists of two types of elemental products, one of which produces a factor of -1 and the other producing a factor of 1. So consider

$$P = \{(n,m) \in \mathbb{Z}_p^+ \times \mathbb{Z}_p^+ : n \leq m \wedge nm \equiv 1 \pmod{p}\}.$$

Since, by Theorem 4.1.12, $(\mathbb{Z}_p^+, *)$ is a group, we have that, for every $x \in \mathbb{Z}_p^+$ there is some $y \in \mathbb{Z}_p^+$ such that either $(x,y) \in P$ or $(y,x) \in P$.

We now consider

$$H = \{n \in \mathbb{Z}_p^+ : \exists m \in \mathbb{Z}_p^+ \ (n,m) \in P\}.$$

By the uniqueness of inverses, we know that $n \mapsto n^{-1}$, $H \to \mathbb{Z}_p^+$, is an injective function. Also, by the property mentioned above, note that, for any $x \in \mathbb{Z}_p^+ \setminus H$, there is some $y \in H$ such that $x = y^{-1}$.

We now split H into two disjoint subsets:

$$H_{id} = \{n \in H : n = n^{-1}\} \text{ and } H_\star = \{n \in H : n \neq n^{-1}\}.$$

We start by showing that $H_{id} = \{[1], [-1]\}$. So let $n \in H_{id}$ and note that $n = n^{-1}$ is equivalent to $n^2 \equiv 1 \pmod{p}$. It follows that

$$p \mid n^2 - 1 = (n+1)(n-1) \implies p \mid n+1 \vee p \mid n-1.$$

Hence, if $n^2 \equiv 1 \pmod{p}$, then either $n \equiv 1 \pmod{p}$ or $n \equiv -1 \pmod{p}$. This establishes that $H_{id} \subseteq \{[1], [-1]\}$. For equality, simply note that $p-1 \equiv -1 \pmod{p}$ and that $(-1)^2 = 1$.

Finally, enumerate $H_\star = \{a_1, \ldots, a_k\}$ and note that

$$(p-1)! = (p-1)a_1a_1^{-1}a_2a_2^{-1}\cdots a_ka_k^{-1} \equiv -1 \pmod{p}.$$

This finishes the proof. □

Example 4.1.14. Let $\mathbb{F}$ be either $\mathbb{R}$ or $\mathbb{C}$. Then define

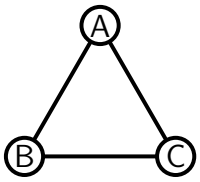
$$\mathrm{GL}(2,\mathbb{F}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : ad - bc \neq 0 \right\}$$

and

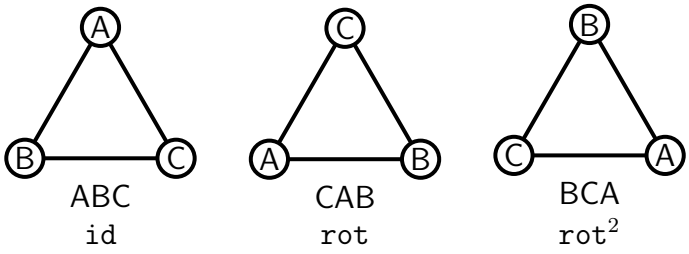
$$\mathrm{SL}(2,\mathbb{F}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : ad - bc = 1 \right\}.$$

With the operation of matrix multiplication, both $\mathrm{GL}(2,\mathbb{F})$ and $\mathrm{SL}(2,\mathbb{F})$ are non-Abelian groups. $\mathrm{GL}(2,\mathbb{F})$ is referred to as the *general linear group* and $\mathrm{SL}(2,\mathbb{F})$ is referred to as the *special linear group*. ⊣

Example 4.1.15. Consider the symmetries of an equilateral triangle with labeled vertices and code each state of the triangle with a word where the first letter corresponds to the top vertex, the second letter corresponds to the bottom left vertex, and the third letter corresponds to the bottom right vertex:

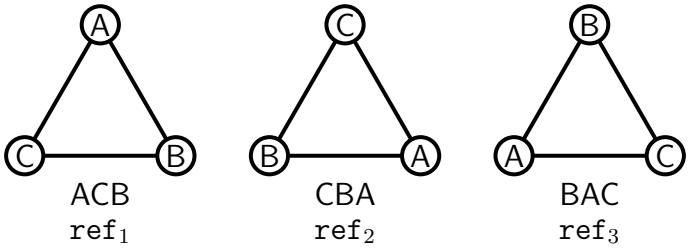


Let `id` represent the null action and let `rot` denote a counterclockwise rotation. We use juxtaposition to indicate actions done in succession, so the for orbit of `rot` looks like this:



Note here[3] that $\texttt{rot}^{-1} = \texttt{rot}^2$ and that $\texttt{rot}^3 = \texttt{id}$.

We also identify three possible reflections:



It is clear that $\texttt{ref}_j^2 = \texttt{id}$ for $j = 1, 2, 3$.

Now, for any two action `a` and `b` listed above, let $\texttt{a} \cdot \texttt{b}$ represent the action obtained by first applying `a` and then applying `b`. It can then be verified that

$$\texttt{rot} \cdot \texttt{ref}_1 = \texttt{ref}_2 \qquad \texttt{rot} \cdot \texttt{ref}_2 = \texttt{ref}_3 \qquad \texttt{rot} \cdot \texttt{ref}_3 = \texttt{ref}_1$$

$$\texttt{rot}^2 \cdot \texttt{ref}_1 = \texttt{ref}_3 \qquad \texttt{rot}^2 \cdot \texttt{ref}_2 = \texttt{ref}_1 \qquad \texttt{rot}^2 \cdot \texttt{ref}_3 = \texttt{ref}_2$$

It can be verified directly as above, or by applying Exercise 4.1.4 to the equations above, that

$$\texttt{ref}_1 \cdot \texttt{rot} = \texttt{ref}_3 \qquad \texttt{ref}_2 \cdot \texttt{rot} = \texttt{ref}_1 \qquad \texttt{ref}_3 \cdot \texttt{rot} = \texttt{ref}_2$$

$$\texttt{ref}_1 \cdot \texttt{rot}^2 = \texttt{ref}_2 \qquad \texttt{ref}_2 \cdot \texttt{rot}^2 = \texttt{ref}_3 \qquad \texttt{ref}_3 \cdot \texttt{rot}^2 = \texttt{ref}_1$$

[3] One clockwise rotation undoes one counter-clockwise rotation.

Let[4]

$$S_3 = \{\mathtt{id}, \mathtt{rot}, \mathtt{rot}^{-1}, \mathtt{ref}_1, \mathtt{ref}_2, \mathtt{ref}_3\}.$$

Then $(S_3, \cdot)$ is a non-Abelian group and its corresponding Cayley table is:

$\cdot$	$\mathtt{id}$	$\mathtt{rot}$	$\mathtt{rot}^{-1}$	$\mathtt{ref}_1$	$\mathtt{ref}_2$	$\mathtt{ref}_3$
$\mathtt{id}$	$\mathtt{id}$	$\mathtt{rot}$	$\mathtt{rot}^{-1}$	$\mathtt{ref}_1$	$\mathtt{ref}_2$	$\mathtt{ref}_3$
$\mathtt{rot}$	$\mathtt{rot}$	$\mathtt{rot}^{-1}$	$\mathtt{id}$	$\mathtt{ref}_2$	$\mathtt{ref}_3$	$\mathtt{ref}_1$
$\mathtt{rot}^{-1}$	$\mathtt{rot}^{-1}$	$\mathtt{id}$	$\mathtt{rot}$	$\mathtt{ref}_3$	$\mathtt{ref}_1$	$\mathtt{ref}_2$
$\mathtt{ref}_1$	$\mathtt{ref}_1$	$\mathtt{ref}_3$	$\mathtt{ref}_2$	$\mathtt{id}$	$\mathtt{rot}^{-1}$	$\mathtt{rot}$
$\mathtt{ref}_2$	$\mathtt{ref}_2$	$\mathtt{ref}_1$	$\mathtt{ref}_3$	$\mathtt{rot}$	$\mathtt{id}$	$\mathtt{rot}^{-1}$
$\mathtt{ref}_3$	$\mathtt{ref}_3$	$\mathtt{ref}_2$	$\mathtt{ref}_1$	$\mathtt{rot}^{-1}$	$\mathtt{rot}$	$\mathtt{id}$

⊣

Comment. Note that every element of S_3 in Example 4.1.15 codes a bijection of a given set of three elements. In fact, $(S_3, \cdot)$ is exactly the group of bijections on the set $\{A, B, C\}$ under function composition.

Another thing to note here is that a copy of $(\mathbb{Z}_3, \oplus)$ lives inside of S_3; indeed,

$$(\{\mathtt{id}, \mathtt{rot}, \mathtt{rot}^{-1}\}, \cdot)$$

can be verified to be a copy of $(\mathbb{Z}_3, \oplus)$.

Exercise 4.1.11. Let X be a non-empty set and, for $A, B \subseteq X$, define the *symmetric difference* between A and B to be

$$A \triangle B = (A \cup B) \setminus (A \cap B).$$

Let $\wp(X)$ be the set of all subsets of X. Show that $(\wp(X), \triangle)$ is an Abelian group.

4.2 Subgroups

Definition 4.2.1. For a group $(G, \cdot)$, we say that $H \subseteq G$ is a *subgroup* of G if $(H, \cdot)$ is a group. The singleton set consisting of the identity element is a subgroup of any given group, and is referred to as the *trivial subgroup*. If H is a proper subset of G and is a subgroup of G, we say that H is a *proper subgroup* of G.

Example 4.2.2. The integer group $(\mathbb{Z}, +)$ is a (proper) subgroup of $(\mathbb{R}, +)$. ⊣

Since the associativity of the operation is inherited, to determine whether a subset of a group is a subgroup, it suffices to check that the subset contains the group identity and is *closed* under the group operations.

Proposition 4.2.3. Suppose $(G, \cdot)$ is a group and $H \subseteq G$. Then H is a subgroup of G if and only if the following three conditions are satisfied:

[4] More on the choice of notation later.

- $e \in H$.
- For $g, h \in H$, $gh \in H$.
- For $h \in H$, $h^{-1} \in H$.

In fact, there is a single condition to check that verifies whether a subset is a subgroup.

Proposition 4.2.4. Let $(G, \cdot)$ be a group and $H \subseteq G$. Then H is a subgroup of G if and only if $H \neq \emptyset$ and, for $g, h \in H$, $gh^{-1} \in H$.

Proof. First, note that, if H is a subgroup, $e \in H$ so $H \neq \emptyset$. Then, for $g, h \in H$, $h^{-1} \in H$ and so $gh^{-1} \in H$.

For the reverse direction, we verify the conditions of Proposition 4.2.3. Since $H \neq \emptyset$ there is some $x \in H$. By the hypothesis, $xx^{-1} = e \in H$.

Now, for any $y \in H$, since $e \in H$, $ey^{-1} = y^{-1} \in H$.

Finally, consider $g, h \in H$. From Exercise 4.1.2, we know that $h = (h^{-1})^{-1}$. From the argument above, we also know that $h^{-1} \in H$. Hence, by the hypothesis, we have that

$$gh = g(h^{-1})^{-1} \in H,$$

finishing the proof. □

Exercise 4.2.1. Let $\mathbb{F}$ be $\mathbb{R}$ or $\mathbb{C}$. Show that, as defined in Example 4.1.14, $\mathrm{SL}(2, \mathbb{F})$ is a subgroup of $\mathrm{GL}(2, \mathbb{F})$. (*Hint.* There is an important property about determinants that can be helpful here.)

Example 4.2.5. The subgroups of $(\mathbb{Z}_4, \oplus)$ are $\{0\}$, $\{0, 2\}$, and $\mathbb{Z}_4$. It is straightforward to verify that each three of these are subgroups of $\mathbb{Z}_4$. To see that they are the only ones, suppose you have a subgroup H. It will suffice to show that, if $1 \in H$, then $H = \mathbb{Z}_4$. Indeed, note that $-1 \equiv 3 \pmod 4$ so $3 \in H$. Moreover, $1 + 1 = 2 \in H$. ⊣

Exercise 4.2.2. Let $(G, \cdot)$ be a group and $\mathcal{H}$ be a set of subgroups of G. Show that $\bigcap \mathcal{H}$ is a subgroup of G.

Exercise 4.2.3. Provide a counterexample to the following statement: Let $(G, \cdot)$ be a group and suppose that H_1 and H_2 are subgroups of G. Then $H_1 \cup H_2$ is a subgroup of G.

4.3 An Introduction to Homomorphisms

Definition 4.3.1. Suppose $(G, \diamond)$ and $(H, \bullet)$ are groups. A function $\varphi : G \to H$ is said to be a *homomorphism* (of groups) if, for all $g, h \in G$,

$$\varphi(g \diamond h) = \varphi(g) \bullet \varphi(h).$$

Example 4.3.2. Consider the groups $(\mathbb{R}, +)$ and $(\mathbb{R}^+, \cdot)$, where $\mathbb{R}^+ = \{x \in \mathbb{R} : x > 0\}$. Define $\varphi : \mathbb{R} \to \mathbb{R}^+$ by the rule $\varphi(x) = 2^x$. Then φ is a homomorphism. Indeed, check that

$$\begin{aligned} \varphi(x + y) &= 2^{x+y} \\ &= 2^x \cdot 2^y \\ &= \varphi(x) \cdot \varphi(y). \end{aligned}$$

⊣

Exercise 4.3.1. Suppose $\varphi : G \to H$ is a homomorphism from the group $(G, \diamond)$ to the group $(H, \bullet)$. Let e_G be the identity element of G and e_H be the identity element of H. Prove that $\varphi(e_G) = e_H$. (*Hint.* First establish that $\varphi(g) \bullet \varphi(e_G) = \varphi(g)$ for any $g \in G$.)

Exercise 4.3.2. Suppose $\varphi : G \to H$ is a homomorphism from the group $(G, \diamond)$ to the group $(H, \bullet)$. Show that, for any $g \in G$, $\varphi(g^{-1}) = \varphi(g)^{-1}$.

Exercise 4.3.3. Suppose $\varphi : G \to H$ is a homomorphism between multiplicative groups G and H. Show that, for any positive integer n, $\varphi(g^n) = \varphi(g)^n$.

Definition 4.3.3. Suppose $\varphi : G \to H$ is a homomorphism from the group $(G, \diamond)$ to the group $(H, \bullet)$. Let e_H be the identity element of H. We define the *kernel* of φ to be

$$\ker(\varphi) = \varphi^{-1}(e_H) = \{g \in G : \varphi(g) = e_h\}.$$

Exercise 4.3.4. Suppose $\varphi : G \to H$ is a homomorphism from the group $(G, \diamond)$ to the group $(H, \bullet)$. Prove that $\ker(\varphi)$ is a subgroup of G.

Proposition 4.3.4. Suppose $\varphi : G \to H$ is a homomorphism from the group $(G, \diamond)$ to the group $(H, \bullet)$. Suppose E is a subgroup of G. Then

$$\varphi[E] = \{\varphi(g) : g \in E\}$$

is a subgroup of H.

Proof. To prove the proposition, we will apply Proposition 4.2.4.

Since E is a subgroup of G, $e_G \in E$. Then, by Exercise 4.3.1, $\varphi(e_G) = e_H \in \varphi[E]$. Hence, $\varphi[E] \neq \emptyset$.

To finish the proof, suppose $h_1, h_2 \in \varphi[E]$. By definition, there are $g_1, g_2 \in E$ such that $\varphi(g_1) = h_1$ and $\varphi(g_2) = h_2$. Since E is a subgroup of G and $g_1, g_2 \in E$, we see that $g_1 \diamond g_2^{-1} \in E$. By the corresponding application of Exercise 4.3.2, we see that

$$\begin{aligned} \varphi(g_1 \diamond g_2^{-1}) &= \varphi(g_1) \bullet \varphi(g_2^{-1}) \\ &= \varphi(g_1) \bullet \varphi(g_2)^{-1} \\ &= h_1 \bullet h_2^{-1}. \end{aligned}$$

Again, as $g_1 \diamond g_2^{-1} \in E$, we see that

$$h_1 \bullet h_2^{-1} = \varphi(g_1 \diamond g_2^{-1}) \in \varphi[E].$$

Conclusively, Proposition 4.2.4 applies, and $\varphi[E]$ is a subgroup of H. □

Exercise 4.3.5. Suppose $\varphi : G \to H$ is a homomorphism from the group $(G, \diamond)$ to the group $(H, \bullet)$. Suppose E is a subgroup of H. Show that

$$\varphi^{-1}(E) = \{g \in G : \varphi(g) \in E\}$$

is a subgroup of G.

Exercise 4.3.6. Suppose $(G, \diamond)$, $(H, \bullet)$, and $(K, *)$ are groups and that $\varphi : G \to H$ and $\psi : H \to K$ are homomorphisms. Show that $\psi \circ \varphi$ is a homomorphism.

Definition 4.3.5. Suppose $(G, \diamond)$ and $(H, \bullet)$ are groups. A bijection $\varphi : G \to H$ is said to be an *isomorphism* if φ is a homomorphism and its inverse φ^{-1} is also a homomorphism. In such a case, we say that $(G, \diamond)$ and $(H, \bullet)$ are *isomorphic*, which is denoted by $G \cong H$.

In fact, the algebraic structure necessitates that any bijective homomorphism be an isomorphism.

Exercise 4.3.7. Suppose $\varphi : G \to H$ is a bijective homomorphism from the group $(G, \diamond)$ to the group $(H, \bullet)$. Show that φ^{-1} is necessarily a homomorphism.

Whether or not a homomorphism is injective depends completely on its kernel.

Exercise 4.3.8. Suppose $\varphi : G \to H$ is a homomorphism from the group $(G, \diamond)$ to the group $(H, \bullet)$. Show that φ is injective if and only if its kernel is the trivial subgroup of G.

Exercise 4.3.9. Let $R = \{\mathtt{id}, \mathtt{rot}, \mathtt{rot}^{-1}\}$, as in the context of Example 4.1.15.

(a) Show that R is a subgroup of S_3.

(b) Show that $(R, \cdot)$ and $(\mathbb{Z}_3, \oplus)$ are isomorphic.

Example 4.3.6. Let $G = \mathbb{Z}_2 \times \mathbb{Z}_2$ and define $+$ by $(a, b) + (c, d) = (a \oplus c, b \oplus d)$.[5] Note that

$$G = \{(0, 0), (0, 1), (1, 0), (1, 1)\},$$

so G consists of four distinct elements. We argue here that G and $\mathbb{Z}_4$ are not isomorphic.

Note that every element of G is its own inverse. However, in $\mathbb{Z}_4$, the inverse of 1 is 3. So G and $\mathbb{Z}_4$ are not isomorphic. ⊣

Exercise 4.3.10. For the group G defined in Example 4.3.6, show that G has exactly 5 subgroups.

Homomorphisms also preserve algebraic properties, like that of being Abelian.

Proposition 4.3.7. Suppose $(G, \diamond)$ and $(H, \bullet)$ are groups and that $\varphi : G \to H$ is a surjective homomorphism. If G is Abelian, then H is Abelian.

Proof. Let $h_1, h_2 \in H$ be arbitrary. Since φ is surjective, there are $g_1, g_2 \in G$ for which $\varphi(g_1) = h_1$ and $\varphi(g_2) = h_2$. Note then that

$$h_1 \bullet h_2 = \varphi(g_1) \bullet \varphi(g_2) = \varphi(g_1 \diamond g_2) = \varphi(g_2 \diamond g_1) = \varphi(g_2) \bullet \varphi(g_1) = h_2 \bullet h_1.$$

Therefore, H is Abelian. □

Since the trivial mapping $\varphi : G \to \{1\}$ defined by $\varphi(x) = 1$ is a homomorphism from any group to the trivial group, the reverse direction to Proposition 4.3.7 does not necessarily hold.

[5] This, as we will elaborate on later, is known as a *direct product*.

4.4 Cyclic Groups and Subgroups

The groups $(\mathbb{Z}, +)$ and $(\mathbb{Z}_n, \oplus)$, where $n \in \mathbb{N}$, are examples of what we call *cyclic groups* since they can be seen as *generated* by a single non-identity element. In these cases, the element 1 can be seen as the generating element.

Definition 4.4.1. For a group $(G, \cdot)$ and $a \in G$, let $\langle a \rangle = \{a^k : k \in \mathbb{Z}\}$. We will refer to $\langle a \rangle$ as a *cyclic subgroup* of G, and we will say that a is the *generator* of $\langle a \rangle$.

If $G = \langle a \rangle$ for some $a \in G$, we say that G is a *cyclic group*.

Theorem 4.4.2. Let $(G, \cdot)$ be a group and $a \in G$. Then $\langle a \rangle$ is a subgroup of G and, furthermore, $\langle a \rangle$ is the smallest (with respect to subset inclusion) subgroup of G which contains a.

Based on Exercise 4.2.2, letting $\mathcal{H}_a$ be the set of all subgroups of a group $(G, \cdot)$ containing $a \in G$, we can see that

$$\langle a \rangle = \bigcap \mathcal{H}_a.$$

Exercise 4.4.1. Let $(G, \cdot)$ be a group with no proper nontrivial subgroups. Show that G is necessarily cyclic.

We will record the following basic observation here for use later.

Lemma 4.4.3. In a group $(G, \cdot)$, given $a, b \in G$, if there is some $k \in \mathbb{Z}$ for which $b = a^k$, then $\langle b \rangle \subseteq \langle a \rangle$.

Proof. Suppose that $a, b \in G$ and $k \in \mathbb{Z}$ are as in the hypothesis. We show that $\langle b \rangle \subseteq \langle a \rangle$. So let $x \in \langle b \rangle$, which means that $x = b^m$ for some $m \in \mathbb{Z}$. Note then that

$$x = b^m = (a^k)^m = a^{km} \in \langle a \rangle.$$

Since $x \in \langle b \rangle$ was arbitrary, we see that $\langle b \rangle \subseteq \langle a \rangle$. □

Exercise 4.4.2. Let $(G, \cdot)$ be a group and $a \in G$. Show that the mapping $k \mapsto a^k$, $\mathbb{Z} \to G$, is a homomorphism from the group $(\mathbb{Z}, +)$ to $(G, \cdot)$.

Conclude that every cyclic subgroup of a given group is a homomorphic image of $(\mathbb{Z}, +)$.

It follows immediately that homomorphisms preserve the algebraic property of being cyclic.

Proposition 4.4.4. Suppose $(G, \diamond)$ and $(H, \bullet)$ are groups and that $\varphi : G \to H$ is a surjective homomorphism. If G is cyclic, then H is cyclic.

Proof. By Exercise 4.4.2, there exists a surjective homomorphism $\psi : \mathbb{Z} \to G$. Then, by Exercise 4.3.6, $\varphi \circ \psi : \mathbb{Z} \to H$ is a surjective homomorphism. It follows that H is cyclic with generator $\varphi \circ \psi(1)$. □

Since $(\mathbb{Z}, +)$ is Abelian, we can use Proposition 4.3.7 to conclude that:

Theorem 4.4.5. Every cyclic group is Abelian.

As the contrapositive of Theorem 4.4.5, if a group is non-Abelian, then it is not cyclic. For example, $\mathrm{GL}(2,\mathbb{R})$ is not cyclic since it is non-Abelian.

Theorem 4.4.6. Every non-trivial subgroup of $(\mathbb{Z},+)$ is of the form $n\mathbb{Z} = \{nk : k \in \mathbb{Z}\}$ for some $n \in \mathbb{N}$.

Proof. It is straight-forward to verify that $n\mathbb{Z}$, $n \in \mathbb{N}$, is a subgroup of $(\mathbb{Z},+)$. So we need only show these are the only subgroups. Let H be a non-trivial subgroup of $(\mathbb{Z},+)$. Since H is non-trivial, it must contain some $k \neq 0$. Moreover, as H is a subgroup, $-k \in H$. Since either k or $-k$ is positive, H contains at least one positive element. Let m be the smallest positive integer in H. Evidently, $m\mathbb{Z} \subseteq H$.

To finish the proof, we show that $H \subseteq m\mathbb{Z}$. So suppose $h \in H$. Since $h \in \mathbb{Z}$ and m is a positive integer, we can write $h = mq + r$ where $0 \leq r < m$. Note that $mq \in m\mathbb{Z} \subseteq H$, and so $h - mq = r \in H$. Since $0 \leq r < m$ and m was set to be the smallest positive element of H, $r = 0$. That is, $h = mq \in m\mathbb{Z}$. □

Corollary 4.4.7. Every subgroup of a cyclic group is cyclic.

Some cyclic groups, like $(\mathbb{Z},+)$ itself, are infinite. Others, like $(\mathbb{Z}_n,\oplus)$, $n \in \mathbb{N}$, are finite. In $\mathbb{Z}_6$, note that the "powers" of 2 form the sequence $(2,4,0,2,4,0,\ldots)$. Similarly, note that the "powers" of 5 form the sequence $(5,4,3,2,1,0,5,4,3,2,1,0,\ldots)$. In both (cycling) sequences, we reach the identity. In the case of 2, we reach the identity as the third iterate. In the case of 5, we reach the identity as the sixth iterate.

Definition 4.4.8. Let $(G,\cdot)$ be a group. For $a \in G$, we define the *order* of a, denoted by $\mathrm{ord}(a)$, to be the least positive integer n such that $a^n = e_G$, if any such integer exists. Otherwise, we say that the order of a is infinite.

In the example involving $\mathbb{Z}_6$ above, we see that the element 2 has order 3 and that the element 5 has order 6.

Lemma 4.4.9. Let $(G,\cdot)$ be a group and suppose $a \in G$ is such that $a \neq e_G$. If $a^n = e_G$ for a positive integer n, then the order of of a divides n.

Proof. Let $d = \mathrm{ord}(a)$ and suppose that $a^n = e_G$. By the definition of the order, $d \leq n$ and, by the division algorithm, we can write $n = dq + r$ where q and r are integers with $0 \leq r < d$. Note then that

$$e_G = a^n = a^{dq+r} = (a^d)^q a^r = a^r.$$

Since $r < d$ and $a^r = e_G$, it must be the case that $r = 0$. Hence, $n = dq$, which establishes that d divides n. □

The word *order* is also overloaded with the following definition, but the dependence on context should avoid potential confusion.

Definition 4.4.10. Let $(G,\cdot)$ be a group. If G is finite, we use the *order* of G to mean the cardinality of G. When G is infinite, we say that the *order* of G is infinite.

Remark. Note here that the order of an element $g \in G$ corresponds to the order of the corresponding cyclic subgroup $\langle g \rangle$.

As some quick examples, the trivial group consisting of the identity element has order 1, the group $(\mathbb{Z}_7, \oplus)$ has order 7, and the group $(\mathbb{Z}, +)$ has infinite order.

Exercise 4.4.3. Suppose $(G, \cdot)$ is a cyclic group of order n. Show that G is isomorphic to $(\mathbb{Z}_n, \oplus)$.

In the context of $\mathbb{Z}_n$, we can capture the order of any element in terms of its gcd with n. To accomplish this, we will use the following fact without proof.

Theorem 4.4.11 (Bézout's Identity)**.** Let a and b be integers. Then there exist integers x and y such that

$$ax + by = \gcd(a, b).$$

More generally, given integers $a_1, a_2, \ldots, a_n$, there are integers $x_1, x_2, \ldots, x_n$ such that

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = \gcd(a_1, a_2, \ldots, a_n).$$

Toward our goal of characterizing the order of elements in $\mathbb{Z}_n$, we start with a basic observation regarding divisors of n.

Lemma 4.4.12. If $a \in \mathbb{Z}_n$ is such that a divides n, then $\text{ord}(a) = \dfrac{n}{a}$.

Proof. If a divides n, then there is an integer k for which $ak = n$. Since $0 < a < n$ and $n > 0$, we see that k must be positive. Note that, for any ℓ with $0 < \ell < k$, $a\ell < ak = n$. So k is indeed the least positive integer for which $ak \equiv 0 \pmod{n}$. Since $k = \dfrac{n}{a}$, we see that $\text{ord}(a) = \dfrac{n}{a}$. □

Lemma 4.4.13. For any $a \in \mathbb{Z}_n$, where $n > 1$ and $a > 0$, $\langle a \rangle = \langle \gcd(a, n) \rangle$.

Proof. First, note that $a = \gcd(a, n) \cdot k$ for some $k \in \mathbb{Z}$ since $\gcd(a, n)$ divides a. Hence, by Lemma 4.4.3, $\langle a \rangle \subseteq \langle \gcd(a, n) \rangle$.

To finish the proof, we show that $\langle \gcd(a, n) \rangle \subseteq \langle a \rangle$. By Bézout's Identity, there are integers x and y such that $ax + ny = \gcd(a, n)$. In particular, $ax \equiv \gcd(a, n) \pmod{n}$. So, again, Lemma 4.4.3 applies to guarantee that $\langle \gcd(a, n) \rangle \subseteq \langle a \rangle$.

Therefore, $\langle a \rangle = \langle \gcd(a, n) \rangle$. □

Proposition 4.4.14. Given $a \in \mathbb{Z}_n$, where $n > 1$ and $a > 0$,

$$\text{ord}(a) = \frac{n}{\gcd(a, n)}.$$

Proof. By Lemma 4.4.13, $\langle a \rangle = \langle \gcd(a, n) \rangle$, In particular, this means that

$$\text{ord}(a) = \text{ord}(\gcd(a, n)).$$

Since $\gcd(a, n)$ divides n, Lemma 4.4.12 affirms that

$$\text{ord}(a) = \text{ord}(\gcd(a, n)) = \frac{n}{\gcd(a, n)}.$$

□

In fact, for any finite cyclic group G of order n, G has a subgroup of order d for any divisor d of n.

Proposition 4.4.15. Let $(G, \cdot)$ be a cyclic group of order n and let $d > 1$ be a divisor of n. Then G has a subgroup of order d.

Proof. Let n and d be as in the hypothesis. Then $n = dk$ for some positive integer k. Note that G is isomorphic to $\mathbb{Z}_n$, so we restrict our attention to $\mathbb{Z}_n$. Note that $k \in \mathbb{Z}_n$ and that the order of k is d. Therefore, $\langle k \rangle$ is a subgroup of order d. □

Exercise 4.4.4. Suppose $(G, \cdot)$ is a group which has a unique nontrivial proper subgroup. Show that G is a cyclic group of order p^2, where p is a prime number.

For elements in a group that commute and for which their corresponding cyclic subgroups only have the identity element in common, we can compute the order of their product based on their respective orders.

Proposition 4.4.16. Suppose $(G, \cdot)$ is a group and that $g, h \in G$ are elements of finite order such that $gh = hg$ and that $\langle g \rangle \cap \langle h \rangle = \{e_G\}$. Then

$$\text{ord}(gh) = \text{lcm}(\text{ord}(g), \text{ord}(h)).$$

More generally, if $g_1, g_2, \ldots, g_k \in G$ commute, are each of finite order, and

$$\langle g_j \rangle \cap H_j = \{e_G\}$$

where

$$H_j = \{g_1^{\alpha_1} g_2^{\alpha_2} \cdots g_k^{\alpha_k} : \alpha_1, \alpha_2, \ldots, \alpha_k \in \mathbb{Z}, \alpha_j = 0\},$$

for every positive integer $j \leq k$,

$$\text{ord}(g_1 g_2 \cdots g_k) = \text{lcm}(\text{ord}(g_1), \text{ord}(g_2), \ldots, \text{ord}(g_k)).$$

Proof. Let $p = \text{ord}(g)$, $q = \text{ord}(h)$, $m = \text{lcm}(p, q)$, and $r = \text{ord}(gh)$. Since m is a multiple of both p and q, then we can write $m = pa$ and $m = qb$ for integers a and b. Note then that

$$(gh)^m = g^m h^m = (g^p)^a (h^q)^b = e_G.$$

So $r \leq m$.

To show that $m \leq r$, we first by claiming that $r \geq \max\{p, q\}$. To see this, consider the fact that

$$g^r h^r = (gh)^r = e_G \implies g^r = h^{-r}.$$

It follows that $g^r \in \langle g \rangle \cap \langle h \rangle$, and so $g^r = e_G$. In a similar fashion, $h^r \in \langle g \rangle \cap \langle h \rangle = \{e_G\}$. So $p \leq r$ and $q \leq r$. Hence, we can write $r = ps_1 + t_1$ and $r = qs_2 + t_2$ where s_1, t_1, s_2, and t_2 are integers with $0 \leq t_1 < p$ and $0 \leq t_2 < q$. Then note that

$$e_G = (gh)^r = g^r h^r = (g^p)^{s_1} g^{t_1} (h^q)^{s_2} h^{t_2} = g^{t_1} h^{t_2}.$$

Since $t_1 < r$ and $t_2 < q$, it must be the case that $t_1 = t_2 = 0$. In particular, r is a multiple of both p and q. Therefore, $m \leq r$.

Now, for the more general case, suppose the proposition holds for all $k \leq n$ for $n \geq 2$ and suppose $g_1, g_2, \ldots, g_k, g_{k+1} \in G$ satisfy the stated hypotheses. Note that, in particular, $g_1, g_2, \ldots, g_k$ satisfy the stated hypotheses, so we can use the inductive hypothesis to conclude that

$$\text{ord}(g_1 g_2 \cdots g_k) = \text{lcm}(\text{ord}(g_1), \text{ord}(g_2), \ldots, \text{ord}(g_k)).$$

Now, since all of the g_j commute with each other, then $h := g_1 g_2 \cdots g_k$ and g_{k+1} commute with each other. Moreover,

$$\langle g_{k+1} \rangle \cap \langle h \rangle \subseteq \langle g_{k+1} \rangle \cap H_{k+1} = \{e_G\}.$$

Hence, the former assertion of the proposition applies to guarantee that

$$\text{ord}(hg_{k+1}) = \text{lcm}(\text{ord}(h), \text{ord}(g_{k+1})) = \text{lcm}(\text{ord}(g_1), \text{ord}(g_2), \ldots, \text{ord}(g_{k+1})),$$

finishing the proof. □

The condition that $\langle g \rangle \cap \langle h \rangle = \{e_G\}$ in Proposition 4.4.16 is a necessary one since

$$\text{ord}(gg^{-1}) = 1,$$

regardless of what properties g may have.

For the more general statement, consider $\{01, 10, 11\}$, as the three non-identity elements of the group in Example 4.3.6 written in string form. Note that these elements all commute with each other and that they generate cyclic subgroups that have pairwise trivial intersection. However,

$$\text{ord}(01 + 10 + 11) = \text{ord}(00) = 1 \neq \text{lcm}(2, 2, 2).$$

The condition that the elements g and h commute in Proposition 4.4.16 is also a necessary one. In fact, for elements that don't commute, it's possible to have two elements of finite order with a product that has infinite order.

Exercise 4.4.5. Let

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

and

$$B = \begin{bmatrix} 0 & -1 \\ 1 & -1 \end{bmatrix}.$$

Prove that, as elements of $(\text{GL}(2, \mathbb{R}), \cdot)$, A and B both have finite orders, but that AB has infinite order.

Example 4.4.17 (Roots of Unity)**.** Recall that the complex numbers $\mathbb{C}$ can be expressed in the form $x + iy$ where $x, y \in \mathbb{R}$ and i is chosen to be a solution to the equation $x^2 + 1 = 0$. After this choice has been made, there are two roots to the equation $x^2 + 1 = 0$ over $\mathbb{C}$: i and $-i$. Let $\mathbb{C}^* = \mathbb{C} \setminus \{0\}$. Then $(\mathbb{C}^*, \cdot)$, where $\cdot$ is the usual multiplication of complex numbers, forms an Abelian group. We define the *modulus* of a complex number $x + iy$ to be

$$|x + iy| = \sqrt{x^2 + y^2}.$$

It can then be shown that

$$\mathbb{S}^1 = \{z \in \mathbb{C} : |z| = 1\}$$

is a subgroup of $(\mathbb{C}^*, \cdot)$. $\mathbb{S}^1$ is known as the *circle group*.

Let $n \in \mathbb{N}$ and consider

$$H_n = \{z \in \mathbb{C} : z^n = 1\}.$$

The elements of H_n are referred to as the n^{th} *roots of unity*. As you will be asked to show in Exercise 4.4.6, H_n is a cyclic subgroup of $\mathbb{S}^1$ of order n; indeed, $(H_n, \cdot)$ is isomorphic to $(\mathbb{Z}_n, \oplus)$. ⊣

Exercise 4.4.6. Show that, for $n \in \mathbb{N}$, $(H_n, \cdot)$ is cyclic. (*Hint.* Use DeMoivre's Theorem.)

Chapter 5

Permutation Groups

Permutation groups form another important category of groups and arise naturally from the study of geometric symmetries. Permutation groups also have applications to the theory of finding solutions to polynomial equations.

5.1 Permutations and Symmetric Groups

Definition 5.1.1. Given a set X, a *permutation* of X is a bijection $p : X \to X$.

Definition 5.1.2. For a nonempty set X, we define the *symmetric group* on X to be the set S_X of all permutations of X endowed with the binary operation of composition $\circ$.

If the set X is finite, with $n \in \mathbb{N}$ elements (which, without loss of generality, can be taken to be $\{1, 2, \ldots, n\}$), then we let S_n denote S_X.

Definition 5.1.3. A group is called a *permutation group* if it is a subgroup of $(S_n, \circ)$ for some $n \in \mathbb{N}$.

Exercise 5.1.1. Show that $(S_n, \circ)$ is of order $n!$.

Recall Example 4.1.15 and observe that $(S_3, \circ)$ in our updated notation corresponds exactly to the group of symmetries of an equilateral triangle. Note also that $(S_3, \circ)$ is not Abelian. Hence:

Comment. In general, permutation groups are not Abelian.

However, full symmetric groups tend to be strictly larger than groups of geometric symmetries. For example, a square has 8 symmetries (the identity, a counterclockwise rotation by 90° which generates three non-identity rotations, and four reflections), but $(S_4, \circ)$ is a group of order $4! = 24$ by Exercise 5.1.1. By labeling the four corners of the square, we can see that the symmetries of the square naturally forms a subgroup of $(S_4, \circ)$, so the symmetries of the square is a non-trivial example of a permutation group.

When working with symmetric groups, it is convenient to employ *cycle notation*.

Definition 5.1.4. For a permutation p of a set X, we define the *support* of p to be

$$\text{supp}(p) = \{x \in X : p(x) \neq x\}.$$

That is, the support of p is the set of points of X that are moved to points other than themselves by p.

Definition 5.1.5. A *cycle* of length $k \geq 2$ is a permutation σ of $n \geq k$ symbols such that there exists a collection $\{a_1, a_2, \ldots, a_k\}$ of distinct symbols such that $\sigma(a_j) = a_{j+1}$, for $1 \leq j < k$, $\sigma(a_k) = a_1$, and $\text{supp}(\sigma) = \{a_1, a_2, \ldots, a_k\}$. In such a case, we use the *cycle notation*

$$(a_1\ a_2\ a_3\ \cdots\ a_k)$$

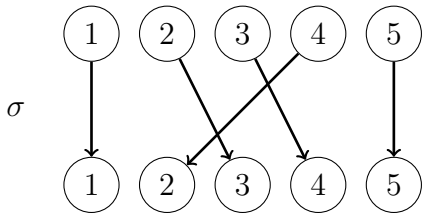
for σ. For convenience, we will use () to denote the identity permutation, which we will call the *trivial* cycle or permutation.

Comment. Since we will generally be phrasing symmetric groups as the group of permutations of $\{1, 2, \ldots, n\}$, for $n \in \mathbb{N}$, given any cycle σ of S_n, we can choose a_1 to be the minimal element of the support of σ. This uniquely determines a cycle notation for σ.

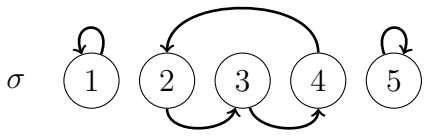
Example 5.1.6. Consider $X = \{1, 2, 3, 4, 5\}$. By (2 3 4), we mean the permutation σ of X where

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

Another representation of σ:



We may also describe σ using a directed graph:



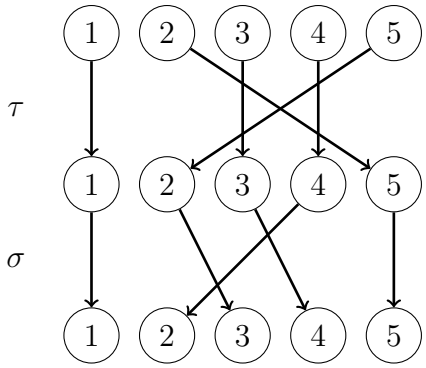
In this case, note that $\text{supp}(\sigma) = \{2, 3, 4\}$. ⊣

Remark. When using cycle notation, we will use juxtaposition to represent the operation of function composition.

Example 5.1.7. Let $X = \{1, 2, 3, 4, 5\}$, $\sigma = (2\ 3\ 4)$, and $\tau = (2\ 5)$. Then we use the notation $\sigma\tau = (2\ 3\ 4)(2\ 5)$ to be $\sigma \circ \tau$, which is the mapping

$$\sigma\tau(1) = 1 \qquad \sigma\tau(2) = 5 \qquad \sigma\tau(3) = 4 \qquad \sigma\tau(4) = 2 \qquad \sigma\tau(5) = 3$$

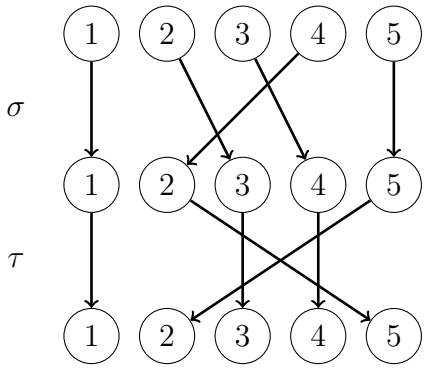
Note that this aligns with the following graphical representation (note that τ is applied *before* σ because of how we notate function composition):



We can express $\sigma\tau$ using cycle notation:

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

Note that switching the order generates a different permutation:



That is,

$$\tau\sigma(1) = 1 \qquad \tau\sigma(2) = 3 \qquad \tau\sigma(3) = 4 \qquad \tau\sigma(4) = 5 \qquad \tau\sigma(5) = 2$$

We can also express $\tau\sigma$ using cycle notation:

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

In this case, we have that $\sigma\tau \neq \tau\sigma$. ⊣

So, in general, cycles do not commute with each other. However, there are cases in which cycles *do* commute.

Definition 5.1.8. Two cycles σ and τ are said to be *disjoint* if $\text{supp}(\sigma) \cap \text{supp}(\tau) = \emptyset$.

Exercise 5.1.2. Show that, if σ and τ are disjoint cycles, then $\sigma\tau = \tau\sigma$.

Despite the scenario in Example 5.1.7 where $\sigma\tau$ and $\tau\sigma$ could be rewritten as single cycles, not all permutations can be expressed as a single cycle. Indeed, for $X = \{1, 2, 3, 4, 5\}$, the permutation (1 2)(4 5) cannot be simplified into a single cycle. We can, however, express permutations as products of disjoint cycles.

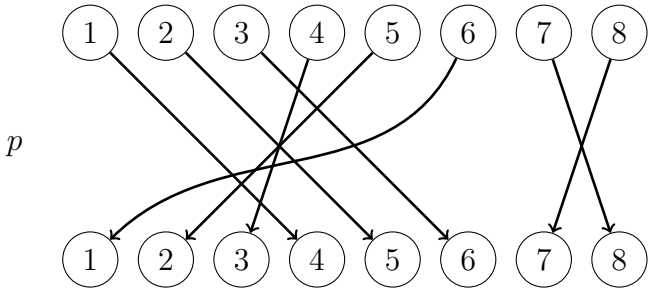
Theorem 5.1.9. Every permutation in S_n can be written as the product of disjoint cycles.

Proof. Let $X = \{1, 2, \ldots, n\}$ and $\sigma \in S_n$. Then define $A_1 = \{\sigma^m(1) : m \in \mathbb{Z}\}$. Note that $1 \in A_1$, so $A_1 \neq \emptyset$. Also note that A_1 is the support of a cycle containing 1.

Now, for $k \in \mathbb{N}$, suppose we have defined $\{A_j : j \leq k\}$. If $X = \bigcup\{A_j : j \leq k\}$, then each A_j corresponds to a cycle, and these cycles are pair-wise disjoint, so we are done. Otherwise, we can let x be the minimal element of $X \setminus \bigcup\{A_j : j \leq k\}$. Then we define $A_{k+1} = \{\sigma^m(x) : m \in \mathbb{Z}\}$. Note that $x \in A_{k+1}$, so $A_{k+1} \neq \emptyset$.

Since X is finite and each A_k to be defined by the process above is non-empty and pair-wise disjoint, the process must terminate at some finite stage. It follows that σ can be rewritten as a product of disjoint cycles. □

To see how the proof above functions in context, let $X = \{1, 2, 3, 4, 5, 6, 7, 8\}$ and consider the permutation p described by



Note that the *orbit* of 1 under p generates the cycle

$$(1\ 4\ 3\ 6).$$

Then $A_1 = \{1, 3, 4, 6\}$. Then $X \setminus A_1 = \{2, 5, 7, 8\}$. The smallest value of this set is 2 and the orbit of 2 under σ generates the cycle

$$(2\ 5).$$

Then $A_2 = \{2, 5\}$ and $X \setminus (A_1 \cup A_2) = \{7, 8\}$. The smallest of these values is 7, and 7 generates the final cycle

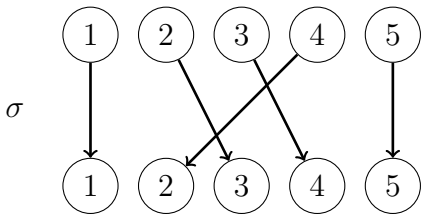
$$(7\ 8).$$

Therefore,

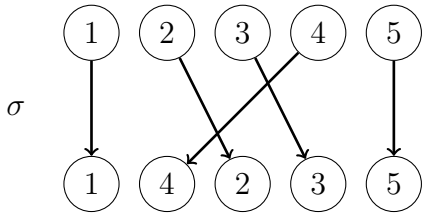
$$p = (1\ 4\ 3\ 6)(2\ 5)(7\ 8). \tag{5.1}$$

Another useful way to represent permutations is as matrices.

Example 5.1.10. Refer back to the σ and τ of Example 5.1.7. To faithfully capture the "action" of the permutation, we want to find a matrix for σ which describes the rearrangement of the symbols, as in the following diagram:



If we keep the same "action," but shuffle the locations, then we obtain



Rewriting these as column matrices, we are looking for a matrix which accomplishes the following "action":

$$\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} \longrightarrow \begin{bmatrix} 1 \\ 4 \\ 2 \\ 3 \\ 5 \end{bmatrix}$$

Adding subscripts to the right-hand column matrix indicating their position, we can recover the original bijection:

$$\begin{matrix} 1_1 \\ 4_2 \\ 2_3 \\ 3_4 \\ 5_5 \end{matrix}$$

The value of the entry is mapped to its subscript under the original bijection.

Now, thinking of using row swapping operations, we can capture σ with

$$\hat{\sigma} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and, in a similar way, τ can be captured with

$$\hat{\tau} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Another way to read these matrices is to note that, for example, in $\hat{\sigma}$, the second column has the 1 in the third row. In the original bijection, 2 is mapped to 3. Similarly, the fourth column has the 1 in the second row, and the original bijection sent 4 to 2.

Now, we can verify that

$$\hat{\sigma}\begin{bmatrix}1\\2\\3\\4\\5\end{bmatrix} = \begin{bmatrix}1&0&0&0&0\\0&0&0&1&0\\0&1&0&0&0\\0&0&1&0&0\\0&0&0&0&1\end{bmatrix}\begin{bmatrix}1\\2\\3\\4\\5\end{bmatrix} = \begin{bmatrix}1\\4\\2\\3\\5\end{bmatrix}$$

and that

$$\hat{\tau}\begin{bmatrix}1\\2\\3\\4\\5\end{bmatrix} = \begin{bmatrix}1&0&0&0&0\\0&0&0&0&1\\0&0&1&0&0\\0&0&0&1&0\\0&1&0&0&0\end{bmatrix}\begin{bmatrix}1\\2\\3\\4\\5\end{bmatrix} = \begin{bmatrix}1\\5\\3\\4\\2\end{bmatrix}.$$

Note that

$$\hat{\sigma}\hat{\tau}\begin{bmatrix}1\\2\\3\\4\\5\end{bmatrix} = \begin{bmatrix}1\\4\\5\\3\\2\end{bmatrix} \quad \text{and} \quad \hat{\tau}\hat{\sigma}\begin{bmatrix}1\\2\\3\\4\\5\end{bmatrix} = \begin{bmatrix}1\\5\\2\\3\\4\end{bmatrix}.$$

Note that, interpreting the resulting column matrix, particularly in the case $\hat{\sigma}\hat{\tau}$, we have a map which does the following:

$$1 \mapsto 1, \quad 4 \mapsto 2, \quad 5 \mapsto 3, \quad 3 \mapsto 4, \quad 2 \mapsto 5$$

Observe that this corresponds exactly to $\sigma\tau$ from Example 5.1.7. ⊣

Fact 5.1.11. For each $p \in S_n$, define

$$\hat{p} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{bmatrix}$$

where

$$a_{j,k} = \begin{cases} 1, & j = p(k) \\ 0, & \text{otherwise} \end{cases}$$

Then the mapping $p \mapsto \hat{p}$, $S_n \to \mathrm{GL}(n, \mathbb{R})$, is an injective homomorphism of groups. Moreover, $\det(\hat{p}) = \pm 1$, which can be verified by considering cofactor expansions and the fact that each row/column of $\hat{p}$ has exactly one occurrence of 1. At each recursive step in the cofactor expansion, one will be able to select a row/column with exactly one entry of 1 contributing a 1 or -1 to the determinant. At "the bottom" of the recursion, one is left computing either

$$\det\begin{bmatrix}1&0\\0&1\end{bmatrix} = 1$$

or

$$\det \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = -1.$$

Hence, the final determinant is either 1 or -1.

Definition 5.1.12. Given a permutation $p \in S_n$, we define the *sign* of p to be

$$\operatorname{sgn}(p) = \det(\hat{p}).$$

By Fact 5.1.11, $\operatorname{sgn}(p) = \pm 1$.

Exercise 5.1.3. Consider $G = \{-1, 1\}$. Show that the map $\operatorname{sgn} : S_n \to G$ is a homomorphism from $(S_n, \circ) \to (G, \cdot)$.

Definition 5.1.13. For S_n, we define the *alternating group* A_n to consist of all $p \in S_n$ with $\operatorname{sgn}(p) = 1$. Note that, by Exercise 4.3.4, A_n is a subgroup of S_n.

There is another way to motivate the alternating group, and that's via *transpositions*.

Definition 5.1.14. Any cycle of the form[1] $(a\ b)$ is called a *transposition*.

Comment. Note that any transposition is its own inverse. Hence, every transposition is of order 2.

In the context of $(S_3, \circ)$, using the notation for the group elements in Example 4.1.15,

$$\begin{array}{lllllllll} \mathtt{id} & = & () & \mathtt{rot} & = & (\mathsf{A\ C\ B}) & \mathtt{rot}^2 & = & (\mathsf{A\ B\ C}) \\ \mathtt{ref}_1 & = & (\mathsf{B\ C}) & \mathtt{ref}_2 & = & (\mathsf{A\ C}) & \mathtt{ref}_3 & = & (\mathsf{A\ B}) \end{array}$$

Note from the corresponding Cayley table, that

$$(\mathsf{A\ C\ B}) = (\mathsf{A\ C})(\mathsf{B\ C}) \quad \text{and} \quad (\mathsf{A\ B\ C}) = (\mathsf{A\ B})(\mathsf{B\ C}).$$

So then the subgroup $\{(), (\mathsf{A\ C})(\mathsf{B\ C}), (\mathsf{A\ B})(\mathsf{B\ C})\}$ of $(S_3, \circ)$ or order 3 consists of elements that can be written as the product of an even number of transpositions. In fact, this subgroup is exactly A_3.

Exercise 5.1.4. Show that the subgroup $\{\mathtt{id}, \mathtt{rot}, \mathtt{rot}^{-1}\}$ of S_3 is A_3 by computing the sign of every element of S_3.

Proposition 5.1.15. Every non-trivial cycle can be written as a product of transpositions.

Proof. Consider the cycle $(a_1\ a_2\ \cdots\ a_k)$, where $k \geq 2$, and observe that

$$(a_1\ a_2\ \cdots\ a_k) = (a_1\ a_2)(a_2\ a_3)\cdots(a_{k-1}\ a_k).$$

Indeed, for a_j, where $1 \leq j < k$, the first transposition affecting a_j will be $(a_j\ a_{j+1})$, taking a_j to a_{j+1}. Then a_{j+1} appears in no further transpositions. In the case we have a_k, a_k first gets moved to a_{k-1}, which then gets moved to a_{k-2} in the transposition $(a_{k-2}\ a_{k-1})$. This continues in this way until we reach $(a_1\ a_2)$, finally taking a_2 to a_1. So, in the end, a_k is mapped to a_1, finishing the proof. □

[1] Note that $a \neq b$ by convention.

We note however that the decomposition of cycles into a product of transpositions is not unique. Indeed, note that

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

Nevertheless, we have the following result governing the parity of products of transpositions.

Theorem 5.1.16. If a cycle σ can be written as the product of an even number of transpositions, then any decomposition of σ into a product of transpositions must consist of an even number of transpositions. Consequently, if a cycle σ can be written as the product of an odd number of transpositions, then any decomposition of σ into a product of transpositions must consist of an odd number of transpositions.

Proof. Note that every transposition τ has the property that $\text{sgn}(\tau) = -1$. Since the sgn function is a homomorphism, any product of an even number of transpositions has a positive sign, and any product of an odd number of transpositions has an odd sign. □

Remark. Using the cycle decomposition into transpositions as in the proof of Proposition 5.1.15, we see that a $(k+1)$-cycle, where $k \in \mathbb{N}$, can be written as a product of k-many transpositions. Hence, any $(2k+1)$-cycle is even and any $(2k)$-cycle is odd.

In fact, as long as $n \geq 3$, the alternating group is generated by the 3-cycles.

Theorem 5.1.17. For $n \geq 3$, any even permutation can be written as the product of 3-cycles. In other words, the alternating group A_n is generated by the 3-cycles.

Proof. It suffices to show that a product of transpositions can be written as a product of 3-cycles. So let $(a\ b)$ and $(c\ d)$ be transpositions (hence, $a \neq b$ and $c \neq d$). We proceed by cases.

If $\{a, b\} = \{c, d\}$, then $(a\ b) = (c\ d)$. It follows that

$$(a\ b)(c\ d) = () = (1\ 2\ 3)(1\ 3\ 2).$$

Now suppose that $|\{a, b\} \cap \{c, d\}| = 1$. Without loss of generality, suppose $b = d$. Note then that

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

For the final case, suppose that $(a\ b)$ and $(c\ d)$ are disjoint transpositions. Then verify that

$$(a\ b)(c\ d) = (a\ d\ c)(a\ b\ c).$$

Finally, since any even permutation can be written as a product of an even number of transpositions, and each product of a pair of transpositions can be written as a product of 3-cycles, then every even permutation can be written as a prodct of 3-cycles. □

Proposition 5.1.18. In any finite permutation group, the number of even permutations is equal to the number of odd permutations. In other words, exactly half of all the permutations in a finite permutation group are even.

Proof. Consider the permutation group S_n, where $n > 1$. For convenience, let $B = S_n \setminus A_n$, the set of odd permutations. Fix a transposition $\tau \in S_n$ and define a function $f : B \to A_n$ by the rule $f(p) = \tau p$. Note that the map f is defined since τp is an even permutation whenever p is an odd permutation. To finish the proof, we will show that f is a bijection.

First, suppose $p, q \in B$ are such that

$$\tau p = f(p) = f(q) = \tau q.$$

We can then left-multiply by $\tau^{-1} = \tau$ to yield that $p = q$. That is, f is injective.

To see that f is a surjection, let $p \in A_n$ and consider $\tau p \in B$. Since $\tau^{-1} = \tau$, we see that $f(\tau p) = \tau\tau p = p$. Hence, f is a surjection. □

We show here that we can compute the order of any permutation using only information about its "cycle type."

Lemma 5.1.19. The order of any n-cycle is n.

Proof. Suppose $\sigma = (a_0\ a_1\ a_2\ \dots\ a_{n-1})$. Note that, for $k \geq 1$, $\sigma^k(a_0) = a_{k\,\%\,n}$. So, for any k, $1 \leq k < n$, $\sigma^k \neq ()$. However, $\sigma^n = ()$ since, for any j, $0 \leq j < n$, $\sigma^n(a_j) = a_{(j+n)\,\%\,n} = a_j$. Therefore, $\operatorname{ord}(\sigma) = n$. □

Since all permutation can be expressed as a product of disjoint cycles, a common way to refer to permutations is in terms of their *cycle type*. For example,

$$(1\ 2)(3\ 4)(5\ 6\ 7)$$

is a cycle of type $(2, 2, 3)$.

Proposition 5.1.20. The order of a permutation of cycle type $(n_1, n_2, \dots, n_k)$ is

$$\operatorname{lcm}(n_1, n_2, \dots, n_k).$$

Proof. Let σ_j be the n_j-cycle, for $1 \leq j \leq k$, and let

$$H_j = \{\sigma_1^{\alpha_1}\sigma_2^{\alpha_2}\cdots\sigma_k^{\alpha_k} : \alpha_1, \alpha_2, \dots, \alpha_k \in \mathbb{Z}, \alpha_j = 0\}.$$

To verify that

$$\langle\sigma_j\rangle \cap H_j = \{()\},$$

note that, for any integer a, $\operatorname{supp}(\sigma_j^a) \subseteq \operatorname{supp}(\sigma_j)$. Also, note that, for integers $\alpha_1, \alpha_2, \dots, \alpha_k$ where $\alpha_j = 0$,

$$\begin{aligned}&\operatorname{supp}\left(\sigma_1^{\alpha_1}\sigma_2^{\alpha_2}\cdots\sigma_k^{\alpha_k}\right)\\ &\subseteq \operatorname{supp}(\alpha_1) \cup \operatorname{supp}(\sigma_2) \cup \cdots \cup \operatorname{supp}(\sigma_{j-1}) \cup \operatorname{supp}(\sigma_{j+1}) \cup \cdots \cup \operatorname{supp}(\sigma_k),\end{aligned}$$

since the cycles are disjoint. That is,

$$\operatorname{supp}(\sigma_j) \cap \operatorname{supp}\left(\sigma_1^{\alpha_1}\sigma_2^{\alpha_2}\cdots\sigma_k^{\alpha_k}\right) = \emptyset.$$

Consequently,

$$\langle\sigma_j\rangle \cap H_j = \{()\}.$$

By Lemma 5.1.19, we know that each σ_j has order n_j. By Exercise 5.1.2, we have that disjoint cycles commute. Therefore, by Proposition 4.4.16, the proposition obtains. □

We can also use combinatorial techniques to determine the total number of k-cycles in a symmetric group.

Proposition 5.1.21. Let $n > 1$ and k be an integer $2 \leq k \leq n$. Then there are

$$\frac{P(n,k)}{k} = \frac{n!}{k \cdot (n-k)!}$$

distinct k-cycles in S_n.

Proof. First, note that there are

$$P(n,k) = \frac{n!}{(n-k)!}$$

tuples of k distinct elements from $X := \{1, 2, \ldots, n\}$. Given a particular tuple of distinct elements from X, note that k of the corresponding cycles are equivalent; indeed, note that

$$(a_1\ a_2\ \cdots\ a_k) = (a_2\ a_3\ \cdots\ a_k\ a_1) = \cdots = (a_k\ a_1\ \cdots\ a_{k-1}).$$

Therefore, the total number of distinct k-cycles is

$$\frac{P(n,k)}{k},$$

finishing the proof. □

5.2 Dihedral Groups

The *dihedral groups* are special subgroups of permutation groups corresponding to rigid motions of regular polygons.

Definition 5.2.1. Consider a regular (all angles are equal and all side lengths are equal) polygon with n vertices. The group of rigid motions of the given polygon, called a *dihedral group*, is denoted by D_n, and can be viewed as a subgroup of S_n.[2]

Theorem 5.2.2. The order of D_n, for $n \geq 3$, is $2n$.

Proof. Any rigid motion of a polygon with n vertices can be fully described by where it sends a single pair of adjacent vertices. Then, for a given fixed pair of adjacent vertices (a, b), the vertex a can be moved to one of n vertices, and b can then be sent to one of the two vertices which are adjacent to the vertex a was moved to. Hence, there are at most $2n$ rigid motions.

To see that this upper-bound is attained, note that there are n distinct rotations. To address the number of reflections, we consider two cases.

When n is odd, there is a single reflection that fixes a given vertex. Hence, there are n distinct reflections. Since no rotation fixes any vertices, we see that the total number of rigid motions is at least $n + n = 2n$.

[2]Note that group theorists typically write D_{2n} to refer to D_n, which is partially motivated by Theorem 5.2.2.

When n is even, any reflection that fixes one vertex fixes another vertex opposite to the vertex under consideration. So there are $n/2$ reflections that fix vertices. In this context, there are also $n/2$ reflections that pass through two opposing sides of the given polygon. Hence, there are $\frac{n}{2} + \frac{n}{2} = n$ reflections. Like above, the total number of rigid motions is at least $n + n = 2n$.

Conclusively, the order of D_n is $2n$. □

When working in the context of a dihedral group D_n, we can label the vertices of the corresponding polygon with the integers 1 through n in a counterclockwise fashion. We can then let r denote the counterclockwise rotation that takes vertex 1 to vertex 2. With this choice, r has order n within D_n; that is, $r^n = 1$.

Now, let s denote the reflection that fixes vertex 1. We argue here that any other reflection can be obtained as a product $r^k s$, for some $0 \leq k < n$. Consider an arbitrary reflection g and note that g moves vertex 1 to some other vertex $1 \leq k \leq n$. Then observe that $g = r^{k-1}s$.

Remark. Using the notation above, we have just finished arguing that

$$D_n = \{r^k : 0 \leq k < n\} \cup \{r^k s : 0 \leq k < n\}$$

where $\{r^k : 0 \leq k < n\}$ is the set of all rotations and $\{r^k s : 0 \leq k < n\}$ is the set of all reflections.

Exercise 5.2.1. In the context of D_n, show that $sr^k s = r^{-k}$, where k is an integer. (*Hint.* Note that $r^k s$ is a reflection and that every reflection has order 2.)

5.3 Cayley's Theorem

Exercise 5.3.1. Let $(G, \cdot)$ be a group and, for $g \in G$, define $T_g : G \to G$ by the rule $T_g(x) = gx$. Show that T_g is a permutation of G.

Exercise 5.3.2. Suppose $(G, \cdot)$ is a group and let S_G denote the symmetric group on G. Define $\varphi : G \to S_G$ by the rule $\varphi(g) = T_g$, where T_g is as defined in Exercise 5.3.1. Show that φ is an injective homomorphism.

As a direct consequence of Exercise 5.3.2, we have:

Theorem 5.3.1 (Cayley's Theorem)**.** Every group is (isomorphic to) a permutation group. In particular, if $(G, \cdot)$ is a group of order $n \in \mathbb{N}$, it is isomorphic to a subgroup of S_n.

Remark. Since Fact 5.1.11 affirms that S_n, where $n \in \mathbb{N}$, is isomorphic to a subgroup of $\mathrm{GL}(2, \mathbb{R})$, any group of order n is isomorphic to a subgroup of $\mathrm{GL}(2, \mathbb{R})$.

Chapter 6

Cosets and Lagrange's Theorem

6.1 Cosets

Definition 6.1.1. Let $(G, \cdot)$ be a group and H be a subgroup of G. We let $gH = \{gh : h \in H\}$ and refer to gH as a *left-coset*; likewise, we say that $Hg = \{hg : h \in H\}$ is a *right-coset*.

Example 6.1.2. Consider the symmetric group S_3 and the subgroup $H = \{(), (1\ 2)\}$. We will list both the left-cosets of H and the right-cosets of H. First, note that

$$S_3 = \{(), (1\ 2), (1\ 3), (2\ 3), (1\ 2\ 3), (1\ 3\ 2)\}.$$

We can list out the elements gh where $g \in S_3$ and $h \in H$ in tabular form:

	()	(1 2)
()	()	(1 2)
(1 2)	(1 2)	()
(1 3)	(1 3)	(1 2 3)
(2 3)	(2 3)	(1 3 2)
(1 2 3)	(1 2 3)	(1 3)
(1 3 2)	(1 3 2)	(2 3)

So the distinct left-cosets of H in this context are

$$\begin{cases} H = (1\ 2)H = \{(), (1\ 2)\} \\ (1\ 3)H = (1\ 2\ 3)H = \{(1\ 3), (1\ 2\ 3)\} \\ (2\ 3)H = (1\ 3\ 2)H = \{(2\ 3), (1\ 3\ 2)\} \end{cases}$$

We can also list out the elements hg where $g \in S_3$ and $h \in H$ in tabular form:

	()	(1 2)	(1 3)	(2 3)	(1 2 3)	(1 3 2)
()	()	(1 2)	(1 3)	(2 3)	(1 2 3)	(1 3 2)
(1 2)	(1 2)	()	(1 3 2)	(1 2 3)	(2 3)	(1 3)

So the distinct right-cosets of H in this context are

$$\begin{cases} H = H(1\ 2) = \{(), (1\ 2)\} \\ H(1\ 3) = H(1\ 3\ 2) = \{(1\ 3), (1\ 3\ 2)\} \\ H(2\ 3) = H(1\ 2\ 3) = \{(2\ 3), (1\ 2\ 3)\} \end{cases}$$

Note in both cases that there are three distinct cosets. ⊣

Lemma 6.1.3. Let $(G, \cdot)$ be a group and H be a subgroup of G. Suppose $g_1, g_2 \in G$. Then the following are equivalent:

(i) $g_1 H = g_2 H$

(ii) $Hg_1^{-1} = Hg_2^{-1}$

(iii) $g_2 \in g_1 H$

(iv) $g_2^{-1} \in Hg_1^{-1}$

(v) $g_1^{-1} g_2 \in H$

(vi) $g_2^{-1} g_1 \in H$

Proof. (i) $\Longrightarrow$ (ii): Suppose that $g_1 H = g_2 H$.

We will start by showing that $Hg_1^{-1} \subseteq Hg_2^{-1}$. So consider $hg_1^{-1} \in Hg_1^{-1}$, for $h \in H$, and note that

$$(hg_1^{-1})^{-1} = g_1 h^{-1} \in g_1 H = g_2 H.$$

Then there is some $k \in H$ for which $(hg_1^{-1})^{-1} = g_2 k$. Observe that

$$hg_1^{-1} = \left((hg_1^{-1})^{-1}\right)^{-1} = (g_2 k)^{-1} = k^{-1} g_2^{-1} \in Hg_2^{-1}.$$

Hence, $Hg_1^{-1} \subseteq Hg_2^{-1}$.

We now show that $Hg_2^{-1} \subseteq Hg_1^{-1}$. So consider hg_2^{-1}, where $h \in H$. Like above, $\left(hg_2^{-1}\right)^{-1} = g_2 h^{-1} \in g_2 H = g_1 H$. Then there is some $k \in H$ for which $\left(hg_2^{-1}\right)^{-1} = g_1 k$. As before, observe that

$$hg_2^{-1} = \left(\left(hg_2^{-1}\right)^{-1}\right)^{-1} = (g_1 k\ )^{-1} = k^{-1} g_1^{-1} \in Hg_1^{-1}.$$

So $Hg_2^{-1} \subseteq Hg_1^{-1}$.

Thus, $Hg_1^{-1} = Hg_2^{-1}$.

(ii) $\Longrightarrow$ (iii): Suppose $Hg_1^{-1} = Hg_2^{-1}$. Then there are $h, k \in H$ such that $hg_1^{-1} = kg_2^{-1}$. Note that

$$\begin{aligned} hg_1^{-1} = kg_2^{-1} &\Longrightarrow hg_1^{-1} g_2 = k \\ &\Longrightarrow g_1^{-1} g_2 = h^{-1} k \\ &\Longrightarrow g_2 = g_1 h^{-1} k \in g_1 H. \end{aligned}$$

(iii) $\implies$ (iv): Suppose $g_2 \in g_1H$. Then there is some $h \in H$ for which $g_2 = g_1h$. It follows that

$$g_2^{-1} = (g_1h)^{-1} = h^{-1}g_1^{-1} \in Hg_1^{-1}.$$

(iv) $\implies$ (v): Suppose that $g_2^{-1} \in Hg_1^{-1}$. Then there is some $h \in H$ for which $g_2^{-1} = hg_1^{-1}$. Note then that $g_2 = g_1h^{-1}$ and, furthermore, that $g_1^{-1}g_2 = h^{-1} \in H$.

(v) $\implies$ (vi): Suppose that $g_1^{-1}g_2 \in H$. Then

$$g_2^{-1}g_1 = \left(g_1^{-1}g_2\right)^{-1} \in H.$$

(vi) $\implies$ (i): Suppose $g_2^{-1}g_1 \in H$.

We start by showing that $g_1H \subseteq g_2H$. So let $h \in H$ be arbitrary and note that

$$g_1h = g_2g_2^{-1}g_1h = g_2\left(g_2^{-1}g_1h\right) \in g_2H$$

since $g_2^{-1}g_1 \in H$.

We finish by showing that $g_2H \subseteq g_1H$. So consider g_2h for $h \in H$. Note that $g_1^{-1}g_2 = \left(g_2^{-1}g_1\right)^{-1} \in H$ since $g_2^{-1}g_1 \in H$. It follows that

$$g_2h = g_1g_1^{-1}g_2h = g_1\left(g_1^{-1}g_2h\right) \in g_1H$$

since $g_1^{-1}g_2h \in H$.

Therefore, $g_1H = g_2H$. □

Exercise 6.1.1. Let $(G, \cdot)$ be a group and H be a subgroup of G. Define a relation $\simeq$ on G by the rule $g \simeq h$ provided that $g^{-1}h \in H$. Show that $\simeq$ is an equivalence relation on G and that the equivalence classes of $\simeq$ are precisely the left-cosets of H.

Exercise 6.1.2. Let $(G, \cdot)$ be a group and H be a subgroup of G. Show that the left-cosets form a partition of G.

Exercise 6.1.3. Let $(G, \cdot)$ be a group and H be a subgroup of G. Fix $g \in G$ and define $t_\ell : H \to gH$ and $t_r : H \to Hg$ by $t_\ell(x) = gx$ and $t_r(x) = xg$. Show that both t_ℓ and t_r are bijections and conclude that each left-coset of H and each right-coset of H has the same cardinality as H.

Proposition 6.1.4. Let $(G, \cdot)$ be a group and H be a subgroup of G. Then the number of distinct left-cosets of H is the same as the number of distinct right-cosets of H.

Proof. Let L_H be the set of left-cosets of H and R_H be the set of right-cosets of H. We define $\gamma : \mathsf{L}_H \to \mathsf{R}_H$ by $\gamma(gH) = Hg^{-1}$, where $g \in G$. Note that γ is seen to be both well-defined and injective by Lemma 6.1.3. So, to finish the proof, we need only show that γ is surjective. Note that an arbitrary right-coset of H must be Hg for some $g \in G$. Then observe that

$$\gamma(g^{-1}H) = H(g^{-1})^{-1} = Hg,$$

finishing the proof. □

6.2 Lagrange's Theorem

Now that we know that all cosets are of the same size and that the number of left-cosets is equal to the number of right-cosets, we define the index of a subgroup within a group.

Definition 6.2.1. For a group $(G, \cdot)$, we say that the *index* of a subgroup H of G, denoted by $[G : H]$, is the number of distinct left-cosets of H.

Theorem 6.2.2 (Lagrange's Theorem)**.** Suppose $(G, \cdot)$ is a finite group and that H is a subgroup of G. Then

$$[G : H] = \frac{|G|}{|H|}.$$

Proof. We show that $|G| = [G : H] \cdot |H|$. Suppose $[G : H] = j \in \mathbb{N}$ and let $R := \{g_1, g_2, \ldots, g_j\}$ be such that $\{g_1H, g_2H, \ldots, g_jH\}$ is the complete set of distinct left-cosets. Define $p : R \times H \to G$ by the rule $p(g, h) = gh$. We will show that p is a bijection.

First, suppose $(g_k, h_1), (g_\ell, h_2) \in R \times H$ are such that

$$p(g_k, h_1) = g_k h_1 = g_\ell h_2 = p(g_\ell, h_2).$$

Note then that

$$g_\ell^{-1} g_k = h_2 h_1^{-1} \in H.$$

By Lemma 6.1.3, $g_\ell H = g_k H$. Hence, by our definition of R, $g_\ell = g_k$. It follows that $h_1 = h_2$, so p is injective.

To see that p is surjective, let $g \in G$. Since the left-cosets of H partition G, there exists some $g_k \in R$ such that $g \in g_k H$. Then there is some $h \in H$ for which $g = g_k h$. Note then that $p(g_k, h) = g$. So p is a surjection.

Since p is a bijection,

$$|G| = |R \times H| = [G : H] \cdot |H|,$$

finishing the proof. □

Corollary 6.2.3. In a finite group G, the order of any element divides $|G|$. Consequently, if $n = |G|$ and $g \in G$, $g^n = e_G$.

Proof. Let $n = |G|$, $g \in G$, and $r = \text{ord}(g)$. Note that $r = |\langle g \rangle|$ and that $\langle g \rangle$ is a subgroup of G. By Lagrange's Theorem, r divides n. Hence, there is some integer k for which $rk = n$. It follows that $g^n = g^{rk} = (g^r)^k = e_G$. □

As a particular application, we obtain Fermat's Little Theorem.

Corollary 6.2.4 (Fermat's Little Theorem)**.** If p is a prime and $a \in \mathbb{Z}$, then $a^p \equiv a \pmod{p}$.

Proof. By Theorem 4.1.12, we know that $\mathbb{Z}_p^+$ is a group under multiplication. Note that the order of $\mathbb{Z}_p^+$ is $p - 1$. If $a \equiv 0 \pmod{p}$, then it is clear that $a^p \equiv a \pmod{p}$. So suppose $a \not\equiv 0 \pmod{p}$. Then there is some $x \in \mathbb{Z}_p^+$ for which $a \equiv x \pmod{p}$. By Corollary 6.2.3, $x^{p-1} \equiv 1 \pmod{p}$. Hence, $x^p \equiv x \pmod{p}$ which yields that $a^p \equiv a \pmod{p}$. □

Exercise 6.2.1. Suppose $(G, \cdot)$ is a group and H is a subgroup of G with the property that $|H| \geq \frac{|G|}{2}$. Show that, for every $g \in G$ and every $h \in H$, $ghg^{-1} \in H$.

Exercise 6.2.2. Let $(G, \cdot)$ be a group and H be a subgroup of G. Suppose that, for every $g \in G$ and every $h \in H$, $ghg^{-1} \in H$. Show that, for every $g \in G$, $gH = Hg$; that is, that the left- and right-cosets are identical.

Recall Proposition 4.4.15 which can be seen as a converse of Lagrange's Theorem in the context of cyclic groups. The converse to Lagrange's Theorem in the more general context does not hold, however.

Example 6.2.5. The alternating group A_4 is of order 12 and has no subgroup of order 6. ⊣

Proof. First note that the order of A_4 is $\frac{4!}{2} = 12$. Now suppose $H \subseteq A_4$ is a subgroup with at least 6 elements. By Exercise 6.2.1, we know that $php^{-1} \in H$ for any $p \in A_4$ and $h \in H$. By Proposition 5.1.21, there are

$$\frac{P(4,3)}{3} = 8$$

distinct 3-cycles in S_3. Note that every 3-cycle is an element of A_4. So A_4 has 4 elements which are not 3-cycles. It follows that H must contain at least one 3-cycle, say $(a\ b\ c)$. Let $d \in \{1,2,3,4\} \setminus \{a,b,c\}$ and note that

$$(a\ b\ d)(a\ b\ c)(a\ b\ d)^{-1} = (a\ b\ d)(a\ b\ c)(a\ d\ b) = (b\ d\ c) \in H.$$

Since H is a subgroup, we also have that

$$(a\ b\ c)(b\ d\ c) = (a\ b\ d) \in H.$$

Accounting for inverses, note that we have so far that

$$\{(), (a\ b\ c), (a\ c\ b), (b\ d\ c), (b\ c\ d), (a\ b\ d), (a\ d\ b)\} \subseteq H.$$

Hence, $|H| \geq 7$. Therefore, A_4 has no subgroup of order 6. □

Chapter 7

Creating New Groups From Old

In this section, we will discuss two common ways to create new mathematical objects from existing examples: via Cartesian products and via "quotients." We will also transition to using generic group operations that may be easier to write by hand now that we've had time to get accustomed to the general context of group theory.

7.1 Direct Products

Definition 7.1.1. Suppose $(G, *)$ and $(H, \circ)$ are groups. We define the binary operation $\cdot$ on the Cartesian product $G \times H$ by the rule

$$(g_1, h_1) \cdot (g_2, h_2) = (g_1 * g_2, h_1 \circ h_2).$$

When G and H refer to groups, the notation $G \times H$ will be understood going forward as meaning the Cartesian product $G \times H$ paired with this binary operation, and we will refer to the pair $(G \times H, \cdot)$ as the *direct product* of $(G, *)$ and $(H, \circ)$.

Exercise 7.1.1. Given groups $(G, *)$ and $(H, \circ)$, show that the binary operation on $G \times H$ defined in Definition 7.1.1 satisfies the group axioms.

Exercise 7.1.2. Suppose $(G, *)$ and $(H, \circ)$ are nontrivial groups. Let $\varphi : G \to G \times H$ and $\psi : H \to G \times H$ be defined by $\varphi(g) = (g, e_H)$ and $\psi(h) = (e_G, h)$. Show that both φ and ψ are injective homomorphisms. Conclude that, in this way, $G \times H$ contains at least two distinct nontrivial proper subgroups: one isomorphic to G and the other isomorphic to H.

Exercise 7.1.3. For two groups G and H, define $\pi_1 : G \times H \to G$ and $\pi_2 : G \times H \to H$ by the rules $\pi_1(g, h) = g$ and $\pi_2(g, h) = h$. Show that both π_1 and π_2 are surjective homomorphisms.

When creating new objects from old, it is common to ask which properties are preserved by the new construction.

Proposition 7.1.2. Suppose $(G, *)$ and $(H, \circ)$ are groups. Then the direct product $G \times H$ is Abelian if and only if both G and H are Abelian.

Proof. First, suppose G and H are both Abelian. Let $(g_1, h_1), (g_2, h_2) \in G \times H$ and note that

$$\begin{aligned}(g_1, h_1) \cdot (g_2, h_2) &= (g_1 * g_2, h_1 \circ h_2) \\ &= (g_2 * g_1, h_2 \circ h_1) \\ &= (g_2, h_2) \cdot (g_1, h_1).\end{aligned}$$

Conclusively, $G \times H$ is Abelian.

The other direction follows from Exercise 7.1.3 and Proposition 4.3.7, which states that the homomorphic image of an Abelian group is Abelian. □

The case for the property of being cyclic does not work out similarly.

Example 7.1.3. The direct product of two cyclic groups may fail to be cyclic. Indeed, consider $(\mathbb{Z}_2, \oplus)$ and the direct product with itself, $\mathbb{Z}_2 \times \mathbb{Z}_2$. A straightforward computation shows that no element of $\mathbb{Z}_2 \times \mathbb{Z}_2$ generates the entire group; indeed, each non-identity element has order 2. ⊣

Theorem 7.1.4. For positive integers a and b, $\mathbb{Z}_{ab} \cong \mathbb{Z}_a \times \mathbb{Z}_b$ if and only if $\gcd(a, b) = 1$.

Proof. First, assume that $\gcd(a, b) = 1$. Note that 1 has order a in $\mathbb{Z}_a$ and 1 has order b in $\mathbb{Z}_b$. By Proposition 4.4.16, the order of $(1, 1)$ in $\mathbb{Z}_a \times \mathbb{Z}_b$ is

$$\operatorname{lcm}(a, b) = \frac{ab}{\gcd(a, b)} = ab.$$

Since the cardinality of $\mathbb{Z}_a \times \mathbb{Z}_b$ is ab, we see that $\mathbb{Z}_a \times \mathbb{Z}_b$ is cyclic. Hence, $\mathbb{Z}_a \times \mathbb{Z}_b$ is isomorphic to $\mathbb{Z}_{ab}$.

Now assume that $\gcd(a, b) > 1$ and consider $(x, y) \in \mathbb{Z}_a \times \mathbb{Z}_b$, arbitrary. We first note that $\operatorname{ord}(x)$ is a divisor of a and that $\operatorname{ord}(y)$ is a divisor of b. In particular, we can let k_1 and k_2 be the integers for which $\operatorname{ord}(x) \cdot k_1 = a$ and $\operatorname{ord}(y) \cdot k_2 = b$. We can also let ℓ_1 and ℓ_2 be the integers for which

$$\operatorname{lcm}(a, b) = a\ell_1 = \operatorname{ord}(x)k_1\ell_1$$

and

$$\operatorname{lcm}(a, b) = b\ell_2 = \operatorname{ord}(y)k_2\ell_2.$$

It follows that

$$\operatorname{lcm}(\operatorname{ord}(x), \operatorname{ord}(y)) \leq \operatorname{lcm}(a, b),$$

and thus that

$$\begin{aligned}\operatorname{ord}((x, y)) &= \operatorname{ord}((x, e_H) \cdot (e_G, y)) \\ &= \operatorname{lcm}(\operatorname{ord}(x), \operatorname{ord}(y)) \\ &\leq \operatorname{lcm}(a, b) \\ &= \frac{ab}{\gcd(a, b)} \\ &< ab.\end{aligned}$$

Hence, no element of $\mathbb{Z}_a \times \mathbb{Z}_b$ generates $\mathbb{Z}_a \times \mathbb{Z}_b$, so $\mathbb{Z}_a \times \mathbb{Z}_b$ is not cyclic. It is therefore not isomorphic to $\mathbb{Z}_{ab}$ since $\mathbb{Z}_{ab}$ is cyclic. □

Corollary 7.1.5. Let G be a cyclic group of order n and let

$$n = p_1^{a_1} p_2^{a_2} \cdots p_k^{a_k}$$

be the prime factorization of n, where $\{p_1, p_2, \ldots, p_k\}$ is a set of distinct primes and each a_j is a positive integer. Then

$$G \cong \mathbb{Z}_{p_1^{a_1}} \times \mathbb{Z}_{p_2^{a_2}} \times \cdots \times \mathbb{Z}_{p_k^{a_k}}.$$

The following factorization theorem will be of use to us later.

Theorem 7.1.6. Suppose H and K are subgroups of a group G with the properties that $H \cap K = \{e_G\}$, $G = \{hk : h \in H, k \in K\}$, and $hk = kh$ for all $h \in H$ and $k \in K$. Then $G \cong H \times K$. Moreover, the natural mapping $(h, k) \mapsto hk$, $H \times K \to G$, is a witnessing isomorphism.

Proof. Define $\varphi : H \times K \to G$ by $\varphi(h, k) = hk$. Note that φ is a surjection by the hypothesis.

To see that φ is a homomorphism, let $(h_1, k_1), (h_2, k_2) \in H \times K$ be arbitrary. Then note that

$$\varphi((h_1, k_1) \cdot (h_2, k_2)) = \varphi(h_1 h_2, k_1 k_2) = h_1 h_2 k_1 k_2 = h_1 k_1 h_2 k_2 = \varphi(h_1, k_1) \cdot \varphi(h_2, k_2).$$

The last thing to show is that φ is injective. So suppose $(h_1, k_1), (h_2, k_2) \in H \times K$ are such that $\varphi(h_1, k_1) = \varphi(h_2, k_2)$. Note that this asserts that $h_1 k_1 = h_2 k_2$, which can be rewritten as $h_2^{-1} h_1 = k_2 k_1^{-1}$. Since $h_2^{-1} h_1 \in H$, $h_2^{-1} h_1 = k_2 k_1^{-1} \in K$, and $H \cap K = \{e_G\}$, $h_2^{-1} h_1 = k_2 k_1^{-1} = e_G$. It follows that $h_1 = h_2$ and that $k_2 = k_1$. Therefore, $(h_1, k_1) = (h_2, k_2)$, and the proof is complete. □

In the context of Abelian groups (though it is also true in a slightly more general context) we can extend this result in a natural way to a finite number of factors. Before stating the extension, let's overload some notation. For a group $(G, \cdot)$ and $A \subseteq G$, let $\langle A \rangle = \bigcap \mathcal{H}_A$, where $\mathcal{H}_A$ is the set of all subgroups of G that contain A. In other words, $\langle A \rangle$ is the smallest subgroup of G which contains every element of A.

Corollary 7.1.7. Suppose G is an Abelian group with subgroups $H_1, H_2, \ldots, H_n$ with the properties that $H_j \cap \tilde{H}_k = \{e_G\}$ for $j \neq k$, where

$$\tilde{H}_k = \left\langle \bigcup \{H_\ell : 1 \leq \ell \leq n, \ell \neq j\} \right\rangle,$$

and

$$G = \{h_1 h_2 \cdots h_n : h_1 \in H_1, h_2 \in H_2, \ldots, h_n \in H_n\}.$$

Then

$$G \cong H_1 \times H_2 \times \cdots \times H_n.$$

Before proving the corollary, we comment that the condition on the H_j in the hypothesis is, in general, necessary. Indeed, consider $G = \mathbb{Z}_2 \times \mathbb{Z}_2$ where the elements are written as strings and let $H_1 = \{00, 01\}$, $H_2 = \{00, 10\}$, and $H_3 = \{00, 11\}$. Note that each H_j is a subgroup of G and that $H_j \cap H_k = \{00\}$ for $j \neq k$. However, $G \ncong H_1 \times H_2 \times H_3$ since

$H_1 \times H_2 \times H_3$ has order 8. The primary issue in this example is that the elements of $H_2 \cup H_3$ generate the entire group. Note that $10 + 11 = 01$ and so

$$\langle H_2 \cup H_3 \rangle = H_2 + H_3 = \{00, 10, 11, 01\} = G.$$

So these H_j do not satisfy the hypotheses of Corollary 7.1.7.

Proof of Corollary 7.1.7. We induct on the number of factors n. Note that, when $n = 2$, the result becomes a direct application of Theorem 7.1.6. So suppose the result holds for every $n \leq k$ where $k \geq 2$ and suppose $H_1, H_2, \ldots, H_k, H_{k+1}$ are as in the hypotheses.

Let

$$K = \langle H_1 \cup H_2 \cup \cdots \cup H_k \rangle$$

and notice that $K \cap H_{k+1} = \{e_G\}$ by the hypothesis. Since every element of G can be written as $h_1 h_2 \cdots h_k h_{k+1}$ where $h_j \in H_j$ by the hypothesis and $h_1 h_2 \cdots h_k \in K$,

$$G = \{y h_{k+1} : y \in K, h_{k+1} \in H_{k+1}\},$$

and so Theorem 7.1.6 applies to assert that

$$G \cong K \times H_{k+1}.$$

So, to finish the proof, we need only show that

$$K \cong H_1 \times H_2 \times \cdots \times H_k.$$

By the hypotheses in the statement of the corollary and the inductive hypothesis, it suffices to show that

$$K = \{h_1 h_2 \cdots h_k : h_1 \in H_1, h_2 \in H_2, \ldots, h_k \in H_k\}.$$

Indeed, note that, for any $1 \leq j \leq k$,

$$H_j \cap \left\langle \bigcup \{H_\ell : 1 \leq \ell \leq k, \ell \neq j\} \right\rangle \subseteq H_j \cap \left\langle \bigcup \{H_\ell : 1 \leq \ell \leq k+1, \ell \neq j\} \right\rangle = \{e_G\},$$

so the subgroups $H_1, H_2, \ldots, H_k$ satisfy the requisite condition.

Now, let

$$K' = \{h_1 h_2 \cdots h_k : h_1 \in H_1, h_2 \in H_2, \ldots, h_k \in H_k\}$$

and note that, since G is Abelian, K' is a subgroup of G. Note also that $K \subseteq K'$ since

$$H_1 \cup H_2 \cup \cdots \cup H_k \subseteq K'$$

and K' is a subgroup of G. On the other hand, if $h_1 h_2 \cdots h_k \in K'$, note that $h_j \in K$ for each $j = 1, 2, \ldots, k$, and so $h_1 h_2 \cdots h_k \in K$. That is, $K' \subseteq K$.

Finally, as

$$K = \{h_1 h_2 \cdots h_k : h_1 \in H_1, h_2 \in H_2, \ldots, h_k \in H_k\}$$

and the requisite conditions hold for $H_1, H_2, \ldots, H_k$, the inductive hypothesis applies to guarantee that

$$K \cong H_1 \times H_2 \times \cdots \times H_k.$$

Therefore,

$$G \cong K \times H_{k+1} \cong H_1 \times H_2 \times \cdots \times H_k \times H_{k+1},$$

finishing the proof. □

To elaborate a bit more on the condition on the subgroups in the statement of Corollary 7.1.7, consider $G = \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$. Considering the way the group is defined, we can verify that the subgroups defined by $H_1 = \{000, 001\}$, $H_2 = \{000, 010\}$, and $H_3 = \{000, 100\}$ meet the criteria in the corollary. As an example, note that $\langle H_1 \cup H_2 \rangle = \{000, 001, 010, 011\}$. Then $\langle H_1 \cup H_2 \rangle \cap H_3 = \{000\}$. It may help here to think about $\langle H_1 \cup H_2 \rangle$ as the *span* of $\{001, 010\}$. In this way, 011 is a *linear combination* of 001 and 010.

By way of contrast, suppose we randomly reach in and grab three distinct elements of G of order 2, say 111, 100, and 011. Let $K_1 = \{000, 111\}$, $K_2 = \{000, 100\}$, and $K_3 = \{000, 011\}$. Note that these are all subgroups of G and that $K_j \cap K_\ell = \{000\}$ for $j \neq \ell$. However, $111 \in K_1 \cap \langle K_2 \cup K_3 \rangle$, so this choice $\{111, 100, 011\}$ of elements, unlike $\{001, 010, 100\}$, doesn't form a *basis* for G. In a sense, $\{001, 010, 100\}$ is *linear independent* whereas $\{111, 100, 011\}$ is not.

Ultimately, without having some intimate knowledge of a given group G like we have in the particular example $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$, finding the right subgroups that satisfy the condition given in Corollary 7.1.7 can be challenging, if not impossible.

We end the present discussion on direct products by noting that the direct product construction is not the only way to assign a group structure to a Cartesian product of groups. We include the following example as just a hint of what is possible with Cartesian products, in general.

Example 7.1.8. Consider $(\mathbb{R}^*, \cdot)$ and $(\mathbb{R}, +)$. Define $*$ on the Cartesian product $\mathbb{R}^* \times \mathbb{R}$ by the rule

$$(a, b) * (c, d) = (ac, ad + b).$$

Then $(\mathbb{R}^* \times \mathbb{R}, *)$ is a non-Abelian group, which we prove below. Note then that the direct product $\mathbb{R}^* \times \mathbb{R}$ cannot be isomorphic to $(\mathbb{R}^* \times \mathbb{R}, *)$ since the direct product $\mathbb{R}^* \times \mathbb{R}$ is Abelian. ⊣

Proof. We prove that $(\mathbb{R}^* \times \mathbb{R}, *)$ is a non-Abelian group.

To address associativity, consider $(u, v), (w, x), (y, z) \in \mathbb{R}^* \times \mathbb{R}$. Then note that

$$\begin{aligned}((u, v) * (w, x)) * (y, z) &= (uw, ux + v) * (y, z) \\ &= (uwy, uwz + ux + v)\end{aligned}$$

and that

$$\begin{aligned}(u, v) * ((w, x) * (y, z)) &= (u, v) * (wy, wz + x) \\ &= (uwy, u(wz + x) + v) \\ &= (uwy, uwz + ux + v).\end{aligned}$$

Hence, $*$ is associative.

Now we check that $(1, 0)$ serves as the identity for $*$. Note that

$$(1, 0) * (x, y) = (x, 1 \cdot y + 0) = (x, y)$$

and that

$$(x, y) * (1, 0) = (x, x \cdot 0 + y) = (x, y).$$

For inverses, consider $(x, y) \in \mathbb{R}^* \times \mathbb{R}$ and consider $\left(\frac{1}{x}, -\frac{y}{x}\right)$. Note that $\frac{1}{x} \in \mathbb{R}^*$ since $x \neq 0$ and so $\left(\frac{1}{x}, -\frac{y}{x}\right) \in \mathbb{R}^* \times \mathbb{R}$. Then, observe that

$$(x, y) * \left(\frac{1}{x}, -\frac{y}{x}\right) = (1, -y + y) = (1, 0)$$

and that

$$\left(\frac{1}{x}, -\frac{y}{x}\right) * (x, y) = \left(1, \frac{y}{x} - \frac{y}{x}\right) = (1, 0).$$

Conclusively, $(\mathbb{R}^* \times \mathbb{R}, *)$ is a group. The last thing to show is that it is not Abelian. So consider $(2, 0)$ and $(1, 2)$. Note that

$$(2, 0) * (1, 2) = (2, 4)$$

and that

$$(1, 2) * (2, 0) = (2, 2).$$

Thus, $(2, 0) * (1, 2) \neq (1, 2) * (2, 0)$. □

7.2 Normal Subgroups and Quotient Groups

7.2.1 Normal Subgroups

Definition 7.2.1. A subgroup H of a group $(G, \cdot)$ is said to be *normal* if, for every $g \in G$ and every $h \in H$, $ghg^{-1} \in H$.

Proposition 7.2.2. If $(G, \cdot)$ is an Abelian group, then every subgroup of G is normal.

Exercise 7.2.1. Suppose $(G, *)$ and $(H, \circ)$ are groups. Show that the isomorphic copy of G in $G \times H$ obtained via $g \mapsto (g, e_H)$, $G \to G \times H$, and that the isomorphic copy of H in $G \times H$ obtained via $h \mapsto (e_G, h)$, $H \to G \times H$, are both normal subgroups of $G \times H$.

Proposition 7.2.3. Let $(G, *)$ and $(H, \circ)$ be groups and suppose that $\varphi : G \to H$ is a homomorphism. Then $\ker(\varphi)$ is a normal subgroup of G.

Proof. By Exercise 4.3.4, we have that $\ker(\varphi)$ is a subgroup of G, so we need only show it's a normal subgroup. So let $g \in G$ be arbitrary and $h \in \ker(\varphi)$. Note that

$$\varphi(g * h * g^{-1}) = \varphi(g) \circ \varphi(h) \circ \varphi(g)^{-1} = \varphi(g) \circ e_H \circ \varphi(g)^{-1} = e_H.$$

That is, $g * h * g^{-1} \in \ker(\varphi)$, and the proof is complete. □

Example 7.2.4. By Exercise 5.1.3 and Proposition 7.2.3, we see that A_n is a normal subgroup of S_n for each $n \in \mathbb{N}$. ⊣

Example 7.2.5. Consider the symmetric group S_3 and the subgroup $H = \{(), (1\ 2)\}$. Then H is not a normal subgroup of S_3. Indeed, note that

$$\begin{aligned}(1\ 3)(1\ 2)(1\ 3)^{-1} &= (1\ 3)(1\ 2)(1\ 3) \\ &= (2\ 3) \notin H.\end{aligned}$$

⊣

Note that, if K is a subgroup of H, and H is a subgroup of G, then K is a subgroup of G. Introducing the modifier of normality breaks this transitivity, in general, however.

Example 7.2.6. There is a group G in which we have a normal subgroup H of G and a normal subgroup K of H for which K is not a normal subgroup of G. Note here that

- H being a normal subgroup of G means that $\forall g \in G \, \forall h \in H \ (ghg^{-1} \in H)$ and that
- K being a normal subgroup of H means that $\forall h \in H \, \forall k \in K \ (hkh^{-1} \in K)$.

For K to be normal in G, we would need that, for every $g \in G$ and every $k \in K$, $gkg^{-1} \in K$.

Consider D_4 and, using the notation of §5.2, we write

$$D_4 = \{1, r, r^2, r^3, s, rs, r^2s, r^3s\}.$$

Let $H = \{1, r^2, s, r^2s\}$ and $K = \{1, s\}$. To show that H is a subgroup of D_4, we refer to its Cayley table, where some computations are aided by the use of Exercise 5.2.1:

	1	r^2	s	r^2s
1	1	r^2	s	r^2s
r^2	r^2	1	r^2s	s
s	s	r^2s	1	r^2
r^2s	r^2s	s	r^2	1

Since $[G : H] = \frac{|G|}{|H|} = 2$, we see that H is a normal subgroup of G by Exercise 6.2.1.

Note also that K is clearly a subgroup of H and that $[H : K] = 2$. Again, by Exercise 6.2.1, K is a normal subgroup of H.

However, K is not a normal subgroup of D_4. Indeed, note that

$$rsr^{-1} = r^2s \notin K.$$

⊣

Exercise 7.2.2. For a group G, let $g \in G$ and define $\varphi : G \to G$ by the rule $\varphi(x) = gxg^{-1}$. Show that φ is an isomorphism.

Exercise 7.2.3. Suppose G is a group and that H is a unique subgroup of G with order k, for some positive integer k. Show that H is a normal subgroup of G.

7.2.2 Quotient Groups

When dealing with a normal subgroup, we can turn the set of cosets into a group in its own right.

Definition 7.2.7. Suppose N is a normal subgroup of a group $(G, \cdot)$. Then we can define a binary operation on the cosets of N in the following way:

$$(aN) \cdot (bN) = abN.$$

As will be shown below, this turns the set of cosets into a group, which we will denote by G/N, and refer to as a *quotient group* or a *factor group*.

Proposition 7.2.8. The set of cosets of a normal subgroup N of a group $(G, \cdot)$ with the binary operation defined in Definition 7.2.7 forms a group.

Proof. We first show that the binary operation is well-defined. So suppose we have $a, b, x, y \in G$ such that $aN = xN$ and $bN = yN$. We will show that $abN = xyN$. To accomplish this, we will show that $xy \in abN$. First, note that, since $aN = xN$, $x \in aN$. So there are $n \in N$ for which $x = an$. Note then that $a^{-1}x = n \in N$. By normality, $y^{-1}a^{-1}xy = y^{-1}ny \in N$. So let $m = y^{-1}ny$ and consider the fact that

$$y^{-1}a^{-1}xy = m \implies a^{-1}xy = ym \in yN = bN.$$

Since $ym = bN$, there is some $\ell \in N$ for which $ym = b\ell$. So now we have that $a^{-1}xy = b\ell$. Finally, this yields that $xy = ab\ell \in abN$.

For associativity, note that $(aN \cdot bN) \cdot cN = abN \cdot cN = abcN$ and that $aN \cdot (bN \cdot cN) = aN \cdot bcN = abcN$.

For identity, note that $N = e_G N$ where e_G is the identity of G serves as the identity.

Lastly, check that $g^{-1}N$ serves as the inverse of gN, for any $g \in G$. □

For the group $(\mathbb{Z}, +)$ and a subgroup $n\mathbb{Z}$, where $n > 1$, since $\mathbb{Z}$ is Abelian, $n\mathbb{Z}$ is a normal subgroup. Moreover, a straightforward argument verifies that the quotient group $\mathbb{Z}/n\mathbb{Z}$ and $\mathbb{Z}_n$ are isomorphic.

Example 7.2.9. The normality condition in Proposition 7.2.8 is necessary. Consider the symmetric group S_3 and the subgroup $H = \{(), (1\ 2)\}$. The set of left-cosets and right-cosets are not equal, as demonstrated in Example 6.1.2. Even if we consider only the set of left-cosets, the operation defined in Definition 7.2.7 is not valid. Consider for example, the cosets $H = (1\ 2)H$ and $(1\ 3)H$. In one representation, $H \cdot (1\ 3)H = (1\ 3)H$. In another representation,

$$(1\ 2)H \cdot (1\ 3)H = (1\ 2)(1\ 3)H = (1\ 3\ 2)H \neq (1\ 3)H.$$

⊣

Definition 7.2.10. For a group $(G, \cdot)$ and a normal subgroup N of G, we define the *canonical homomorphism* $\varphi : G \to G/N$ by $\varphi(g) = gN$.

Sanity Check 7.2.4. Verify that φ defined in Definition 7.2.10 is a homomorphism of groups.

Either by direct arguments or by appealing to the canonical homomorphism, it can be shown that the properties of being Abelian or cyclic are preserved by the formation of a quotient group.

Remark. Note that, by Proposition 4.3.7, if $(G, \cdot)$ is an Abelian group and H is a (necessarily normal) subgroup of G, then G/H is an Abelian group.

Similarly, if $(G, \cdot)$ is a cyclic group and H is a (necessarily normal) subgroup of G, then G/H is cyclic by Proposition 4.4.4.

However, quotient groups may gain properties that neither the parent group nor the generating normal subgroup have.

Example 7.2.11. Consider the symmetric group S_4 and its alternating subgroup A_4. Note that S_4/A_4 is isomorphic to $(\mathbb{Z}_2, \oplus)$, and hence cyclic (and consequently Abelian). However, neither S_4 nor A_4 are Abelian. Indeed, note that $(1\ 2\ 3), (1\ 2\ 4) \in A_4$. Then note that

$$(1\ 2\ 3)(1\ 2\ 4) = (1\ 3)(2\ 4)$$

and that

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

Hence,

$$(1\ 2\ 3)(1\ 2\ 4) \neq (1\ 2\ 4)(1\ 2\ 3),$$

establishing that neither A_4 nor S_4 are Abelian. ⊣

It may be tempting to think of the forming of factor groups as a sort of division and that, for a group G and a normal subgroup N, $G \cong (G/N) \times N$. However, this is generally not the case.

Example 7.2.12. Recall again $D_4 = \{1, r, r^2, r^3, s, rs, r^2s, r^3s\}$ and $R = \{1, r, r^2, r^3\}$. We know that, by Exercise 5.2.1, R is a normal subgroup of D_4. We also know that

$$[D_4 : R] = \frac{|D_4|}{|R|} = \frac{8}{4} = 2.$$

It follows that D_4/R is a group of order 2, and hence $D_4/R \cong \mathbb{Z}_2$. Note that R and D_4/R are Abelian, so $(D_4/R) \times R$ is an Abelian group. Hence, since D_4 is not Abelian, $D_4 \ncong (D_4/R) \times R$. ⊣

7.2.3 Isomorphism Theorems

Theorem 7.2.13 (The First Isomorphism Theorem)**.** Suppose $\varphi : G \to H$ is a surjective homomorphism and that $K = \ker(\varphi)$. Let $\psi : G \to G/K$ be the canonical homomorphism. Then there exist a unique isomorphism $\vartheta : G/K \to H$ such that $\varphi = \vartheta \circ \psi$.

The content of The First Isomorphism Theorem can be summarized in the *commuting diagram* Figure 7.1.

Proof. Let $\varphi : G \to H$ and $\psi : G \to G/K$ be as in the hypothesis. Define $\vartheta : G/K \to H$ by the rule

$$\vartheta(gK) = \varphi(g).$$

We first verify that ϑ is well-defined. So consider $gK, hK \in G/K$ such that $gK = hK$. By Lemma 6.1.3, $h^{-1}g \in K = \ker(\varphi)$. That is,

$$e_H = \varphi(h^{-1}g) = \varphi(h)^{-1}\varphi(g).$$

It follows that

$$\vartheta(hK) = \varphi(h) = \varphi(g) = \vartheta(gK).$$

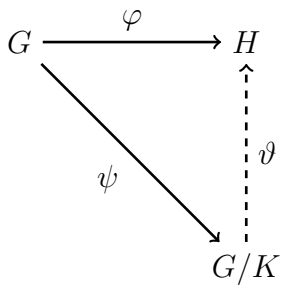


Figure 7.1: Commuting Diagram for The First Isomorphism Theorem

Hence, ϑ is well-defined.

To see that ϑ is a homomorphism, let $gK, hK \in G/K$ and note that

$$\begin{aligned}\vartheta(gK)\vartheta(hK) &= \varphi(g)\varphi(h)\\ &= \varphi(gh)\\ &= \vartheta(ghK)\\ &= \vartheta((gK)(hK)).\end{aligned}$$

Hence, ϑ is a homomorphism.

We know show that ϑ is injective. Suppose $gK, hK \in G/K$ are such that $\vartheta(gK) = \vartheta(hK)$. It follows that

$$\begin{aligned}\varphi(g) = \varphi(h) &\implies e_H = \varphi(g)^{-1}\varphi(h) = \varphi(g^{-1}h)\\ &\implies g^{-1}h \in \ker(\varphi) = K.\end{aligned}$$

By Lemma 6.1.3, $gK = hK$. Hence, ϑ is injective.

To see that ϑ is surjective, let $h \in H$. Since φ was assumed to be surjective, there is some $g \in G$ for which $\varphi(g) = h$. Note then that $gK \in G/K$ and $\vartheta(gK) = \varphi(g) = h$. Thus, ϑ is surjective.

Observe that, for any $g \in G$,

$$\vartheta \circ \psi(g) = \vartheta(gK) = \varphi(g).$$

That is, $\varphi = \vartheta \circ \psi$.

The final thing to prove is that ϑ is the unique isomorphism with the specified properties. So supposed that $\lambda : G/K \to H$ is an isomorphism such that $\varphi = \lambda \circ \psi$. Let $gK \in G/K$ be arbitrary and note that

$$\lambda(gK) = \lambda(\psi(g)) = \lambda \circ \psi(g) = \varphi(g) = \vartheta(gK).$$

Since $\lambda(gK) = \vartheta(gK)$ for every $gK \in G/K$, we see that $\lambda = \vartheta$, establishing the uniqueness of ϑ. □

We provide some examples here elaborating on how The First Isomorphism Theorem can be used to capture certain quotient groups as familiar groups.

Example 7.2.14. Consider the group $(\mathbb{R}, +)$ and the subgroup $\mathbb{Z}$ of $\mathbb{R}$. Since $\mathbb{R}$ is Abelian, $\mathbb{Z}$ is a normal subgroup of $\mathbb{R}$. Now consider the map $\varphi : \mathbb{R} \to \mathbb{S}^1$ defined by

$$\varphi(t) = \cos(2\pi t) + i \sin(2\pi t)$$

where $\mathbb{S}^1$ is the circle group defined in Example 4.4.17. It can be shown that φ is a surjective homomorphism. Note also that

$$\ker(\varphi) = \mathbb{Z}.$$

Then, by The First Isomorphism Theorem, $\mathbb{R}/\mathbb{Z} \cong \mathbb{S}^1$. ⊣

Sanity Check 7.2.5. Verify that the φ defined in Example 7.2.14 is a surjective homomorphism with $\ker(\varphi) = \mathbb{Z}$.

In the next example, we will demonstrate one method for finding ideas for homomorphisms yielding a given normal subgroup as the kernel (other than the canonical homomorphism).

Example 7.2.15. Let

$$G = \left\{ \begin{bmatrix} x & y \\ 0 & 1 \end{bmatrix} : x, y \in \mathbb{C}, x \neq 0 \right\}.$$

Note that the identity matrix is an element of G, that

$$\begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x & y \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} ax & ay + b \\ 0 & 1 \end{bmatrix} \in G,$$

and that

$$\begin{bmatrix} x & y \\ 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{x} & -\frac{y}{x} \\ 0 & 1 \end{bmatrix} \in G.$$

Hence, G is a subgroup of $\mathrm{GL}(2, \mathbb{C})$.

Let

$$H = \left\{ \begin{bmatrix} 1 & x \\ 0 & 1 \end{bmatrix} : x \in \mathbb{C} \right\}.$$

A routine computation shows that H is a normal subgroup of G.

To begin to understand the quotient group, we look into the left-coset relation. That is, what properties must two matrices have to guarantee their membership in the same left-coset? Again, recall Lemma 6.1.3 and note that

$$\begin{aligned} \begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x & y \\ 0 & 1 \end{bmatrix}^{-1} &= \begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{x} & -\frac{y}{x} \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{a}{x} & \frac{bx - ay}{x} \\ 0 & 1 \end{bmatrix}. \end{aligned}$$

Then, observe that, for

$$\begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x & y \\ 0 & 1 \end{bmatrix}^{-1} \in H,$$

we would need $x = a$. From here, we can conclude that

$$\begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix} H = \begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix} H.$$

Note that we chose to use the zero in the top-right entry as the canonical representative since it can be verified that

$$D := \left\{ \begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix} : a \in \mathbb{C}^* \right\}$$

is a group.

Finally, define $\varphi : G \to D$ by the rule

$$\varphi \begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix}.$$

As you will be asked to check, φ is a surjective homomorphism and $\ker(\varphi) = H$. Therefore, by The First Isomorphism Theorem, $G/H \cong D$. ⊣

Exercise 7.2.6. Refer to Example 7.2.15.

(a) Show that H is a normal subgroup of G.

(b) Show that D is a subgroup of $\mathrm{GL}(2, \mathbb{C})$.

(c) Verify that φ is a surjective homomorphism with $\ker(\varphi) = H$.

Remark. In Example 7.2.15, note that

$$\begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix} H = \begin{bmatrix} a & 1 \\ 0 & 1 \end{bmatrix} H,$$

as well. In such a way,

$$\begin{bmatrix} a & 1 \\ 0 & 1 \end{bmatrix}$$

can serve as a canonical representative for the coset

$$\begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix} H,$$

but the set

$$\left\{ \begin{bmatrix} a & 1 \\ 0 & 1 \end{bmatrix} : a \in \mathbb{C}^* \right\}$$

of coset representatives is not a subgroup of G, so we can't naturally generate a homomorphism using these choices.

Exercise 7.2.7. Let $(G, \cdot)$ and $(H, \cdot)$ be groups. Let $\hat{G} = \{(g, e_H) : g \in G\}$ and $\hat{H} = \{(e_G, h) : h \in H\}$. Verify that both $\hat{G}$ and $\hat{H}$ are normal subgroups of $G \times H$ and show that $(G \times H)/\hat{G} \cong H$ and that $(G \times H)/\hat{H} \cong G$.

Definition 7.2.16. Let $(G, *)$ be a group and $A, B \subseteq G$. Then we define

$$A * B = \{a * b : a \in A, b \in B\}.$$

Exercise 7.2.8. Suppose that H and K are subgroups of a group $(G, \cdot)$. Show that

$$H \cup K \subseteq HK.$$

Theorem 7.2.17 (The Second Isomorphism Theorem)**.** Suppose that H is a (not necessarily normal) subgroup of a group $(G, *)$ and that N is a normal subgroup of G. Then $H * N$ is a subgroup of G, $H \cap N$ is a normal subgroup of H, and

$$H/(H \cap N) \cong (H * N)/N.$$

Proof. For notational convenience, we treat the group G as a multiplicative group. We start by showing that HN is a subgroup of G. First note that $e_G \in H \cap N$ so $e_G = e_G \cdot e_G \in HN$.

Now consider $h_1n_1, h_2n_2 \in HN$ where $h_1, h_2 \in H$ and $n_1, n_2 \in N$. Note that

$$h_1n_1h_2n_2 = h_1(h_2h_2^{-1})n_1h_2n_2 = h_1h_2(h_2^{-1}n_1h_2)n_2.$$

Since N is a normal subgroup of G, $h_2^{-1}n_2h_2 \in N$, and so we have that $h_1h_2 \in H$ and $(h_2^{-1}n_1h_2)n_2 \in N$. That is, $h_1n_1h_2n_2 \in HN$.

For the fact that the inverse of every element of HN is a member of HN, let $hn \in HN$ where $h \in H$ and $n \in N$. Note that

$$(hn)^{-1} = n^{-1}h^{-1} = h^{-1}hn^{-1}h^{-1} = h^{-1}(hn^{-1}h^{-1}).$$

Again, since N is a normal subgroup of G, $hn^{-1}h^{-1} \in N$ and so $(hn)^{-1} \in HN$.

We now show that $H \cap N$ is a normal subgroup of H. By Exercise 4.2.2, we know that $H \cap N$ is a subgroup of G, and since $H \cap N \subseteq H$, we see that $H \cap N$ is a subgroup of H. So we need only show it is a normal subgroup of H. So let $h \in H$ and $n \in H \cap N$ be arbitrary. Since N is a normal subgroup of G, we know that $hnh^{-1} \in N$. Note also that $hnh^{-1} \in H$ since $h, n \in H$. Hence, $hnh^{-1} \in H \cap N$, establishing that $H \cap N$ is a normal subgroup of H.

Finally, define $\varphi : H \to (HN)/N$ by the rule $\varphi(h) = hN$. Note that each coset of $(HN)/N$ is of the form hnN where $h \in H$ and $n \in N$. Also, note that $hnN = hN$. So we see that φ is a surjection. Note also that φ is a homomorphism since

$$\varphi(h_1)\varphi(h_2) = (h_1N)(h_2N) = h_1h_2N = \varphi(h_1h_2).$$

Hence, φ is a surjective homomorphism.

Note that $h \in \ker(\varphi)$ if and only if $\varphi(h) = N$. Moreover, $\varphi(h) = hN = N$ holds if and only if $h \in N$ by Lemma 6.1.3. So $\ker(\varphi) = H \cap N$, and so

$$H/(H \cap N) \cong (HN)/N$$

by The First Isomorphism Theorem. □

The normality of N in The Second Isomorphism Theorem is necessary to guarantee that $H * N$ is a subgroup of G.

Example 7.2.18. Consider

$$D_4 = \{1, r, r^2, r^3, s, rs, r^2s, r^3s\}$$

and let $H = \{1, s\}$ and $K = \{1, rs\}$. Note that both H and K are subgroups of G, but neither one is a normal subgroup. Note also that

$$HK = \{1, s, rs, r^3\}$$

since $srs = r^3$. Then note that $(r^3)^{-1} = r \notin HK$. That is, HK is not a subgroup of D_4. $\dashv$

Under the assumption that N is a normal subgroup of a group G and H is a subgroup of G, HN is actually the smallest subgroup of G containing both H and N.

Proposition 7.2.19. Suppose $(G, *)$ is a group, N is a normal subgroup of G, and H is a (not necessarily normal) subgroup of G. Then $H*N$ is the smallest subgroup of G containing $H \cup N$.

Proof. By Exercise 7.2.8, we know that $H \cup N \subseteq H * N$. By The Second Isomorphism Theorem, $H * N$ is a subgroup of G. To see that it is the smallest, suppose K is a subgroup of G such that $H \cup N \subseteq K$. We show that $H * N \subseteq K$. So let $h * n \in H * N$ be arbitrary, where $h \in H$ and $n \in N$. Note that $h \in K$ and $n \in K$. Since K is a subgroup of G, $h * n \in K$. Since $h * n$ was arbitrary, we see that $H * N \subseteq K$. $\square$

With the established context, given a normal subgroup N of a group G and a subgroup H of G, we can see $H \cap N$ as the largest subgroup contained in both H and N (kind of like a greatest common divisor), and HN as the smallest subgroup containing both H and N (kind of like a least common multiple). In this way, The Second Isomorphism Theorem can be loosely seen as a very abstract generalization of the familiar identity

$$ab = \gcd(a, b) \cdot \operatorname{lcm}(a, b),$$

for positive integers a and b.

Example 7.2.20. Consider the group $\mathbb{Z}_{60}$ and note that $\mathbb{Z}_{60}$ is Abelian, so all subgroups of $\mathbb{Z}_{60}$ are normal. Observe that

$$\langle 10 \rangle = \{0, 10, 20, 30, 40, 50\}$$

and that

$$\langle 6 \rangle = \{0, 6, 12, 18, 24, 30, 36, 42, 48, 54\}.$$

Then

$$\langle 10 \rangle \cap \langle 6 \rangle = \{0, 30\} = \langle 30 \rangle.$$

To find $\langle 10 \rangle + \langle 6 \rangle$, we use a table:

	0	6	12	18	24	30	36	42	48	54
0	0	6	12	18	24	30	36	42	48	54
10	10	16	22	28	34	40	46	52	58	4
20	20	26	32	38	44	50	56	2	8	14
30	30	36	42	48	54	0	6	12	18	24
40	40	46	52	58	4	10	16	22	28	34
50	50	56	2	8	14	20	26	32	38	44

It follows that $\langle 10\rangle + \langle 6\rangle = \langle 2\rangle$.

By The Second Isomorphism Theorem, we see that

$$\langle 10\rangle/\langle 30\rangle \cong \langle 2\rangle/\langle 6\rangle \text{ and } \langle 6\rangle/\langle 30\rangle \cong \langle 2\rangle/\langle 10\rangle.$$

If we consider orders, $|\langle 10\rangle| = 6$, $|\langle 6\rangle| = 10$, $|\langle 30\rangle| = 2 = \gcd(6, 10)$, and $|\langle 2\rangle| = 30 = \operatorname{lcm}(6, 10)$. Then isolating one of the congruences above and substituting in the corresponding equation involving the orders, we have that

$$\frac{|\langle 10\rangle|}{|\langle 30\rangle|} = \frac{|\langle 2\rangle|}{|\langle 6\rangle|}$$
$$\frac{6}{\gcd(6, 10)} = \frac{\operatorname{lcm}(6, 10)}{10},$$

which provides the equation

$$6 \cdot 10 = \gcd(6, 10) \cdot \operatorname{lcm}(6, 10) = 2 \cdot 30.$$

⊣

We end our discussion of isomorphism theorems with one more application of The Second Isomorphism Theorem.

Definition 7.2.21. A group G is said to be *metabelian* if there exists an Abelian normal subgroup N for which G/N is also Abelian.

Exercise 7.2.9. Refer to Example 7.2.15 and show that G is a metabelian group which is not Abelian using H as the desired Abelian normal subgroup.

Proposition 7.2.22. Every subgroup of a metabelian group is metabelian.

Proof. Let G be a metabelian group and let H be a subgroup of G. Since G is metabelian, we can let N be an Abelian normal subgroup of G with the property that G/N is Abelian. Note then that $H \cap N$ is an Abelian normal subgroup of H.

Since HN is a subgroup of G, $(HN)/N$ is a subgroup of G/N. Hence, $(HN)/N$ is Abelian. By The Second Isomorphism Theorem, $H/(H \cap N) \cong (HN)/N$, and so $H/(H \cap N)$ is Abelian.

Therefore, H is metabelian. □

Chapter 8

Classifying Finite Abelian Groups

Recall that the direct product of Abelian groups is an Abelian group. So any group of the form

$$G = \mathbb{Z}_{p_1^{a_1}} \times \mathbb{Z}_{p_2^{a_2}} \times \cdots \times \mathbb{Z}_{p_n^{a_n}},$$

where each p_j is a prime number and a_j is a positive integer, is a finite Abelian group. Also note that the group G is cyclic if and only if the primes p_j are all distinct by the proof of Theorem 7.1.4. In particular, part of the proof of Theorem 7.1.4 shows that, if a and b are positive integers with a common factor, then $\mathbb{Z}_a \times \mathbb{Z}_b$ is not cyclic. The goal of this section is to prove that every finite Abelian group is isomorphic to a group of the form

$$\mathbb{Z}_{p_1^{a_1}} \times \mathbb{Z}_{p_2^{a_2}} \times \cdots \times \mathbb{Z}_{p_n^{a_n}},$$

where each p_j is a prime number and a_j is a positive integer.

Lemma 8.0.1. Suppose G is a finite Abelian group and that p is a prime which divides the order of G. Then G has an element of order p.

Proof. We proceed by induction on the order of G.

If $|G| = 1$, there is nothing to show since no prime p divides 1. Hence, the statement of the lemma holds trivially in this case.

Now suppose $k \geq 1$ and that, for any group H with $|H| \leq k$ and any prime p which divides k, H has an element of order p. Suppose G is an Abelian group of order $k+1$ and suppose p is a prime which divides $k+1$. We will proceed by cases.

First, suppose G has no proper nontrivial subgroups. From here it follows that G must be a cyclic group of finite order, in which case $p = k+1$ and hence any generator of G has order p.

Now, suppose G has a proper nontrivial subgroup H. Note that $1 < |H| < |G| = k+1$, which means that $|H| \leq k$.

If p divides $|H|$, then the inductive hypothesis applies to guarantee the existence of an element $h \in H$ with order p. Note that the order of h in G is also p, and so G has an element of order p.

Otherwise, p does not divide $|H|$. Since G is Abelian, H is a normal subgroup of G, and so

$$|G| = |H| \cdot [G : H] = |H| \cdot |G/H|$$

by Lagrange's Theorem. Since p is prime and p divides $|G|$ but doesn't divide $|H|$, p must divide $|G/H|$. Since $|H| > 1$, $|G/H| < |G| = k+1$, which establishes that $|G/H| \leq k$. Hence, the inductive hypothesis applies to guarantee the existence of some $gH \in G/H$ with order p. Since $p \geq 2$, by the definition of order, we see that $gH \neq H$, and so $g \notin H$. On the other hand, $H = (gH)^p = g^pH$, which implies that $g^p \in H$. Let $m = |H|$ and note that, since p is prime and p does not divide m, $\gcd(m,p) = 1$. By Bézout's Identity, we can let x and y be integers such that $px + my = 1$. Since $g^p \in H$ and $|H| = m$, Corollary 6.2.3 guarantees that $(g^m)^p = (g^p)^m = e_G$.

Our final claim is that g^m has order p. By Lemma 4.4.9, the order of g^m must divide p since $(g^m)^p = e_G$. Hence, the order of g^m is either 1 or p, so we show that the order of g^m is not 1. Since the only element of order 1 in any group is the identity, we need only show that $g^m \neq e_G$. Note that

$$\begin{aligned} g &= g^{px+my} \\ &= g^{px}g^{my} \\ &= (g^p)^x(g^m)^y \end{aligned}$$

which implies that

$$(g^m)^{-y}g = (g^p)^x \in H.$$

Since $g \notin H$, it cannot be the case that $g^m = 1$. Therefore, g^m has order p. □

Definition 8.0.2. For a prime number p, we say that a group G is a *p-group* if every element of G has as its order some power of p.

For example, both $\mathbb{Z}_9$ and $\mathbb{Z}_3 \times \mathbb{Z}_3$ are 3-groups.

Remark. Note that any nontrivial p-group has an element of order p, in particular. Indeed, suppose G is a p-group and let g be a non-identity element. Then $\text{ord}(g) = p^m$ for some positive integer m. Note then that

$$e_G = g^{p^m} = g^{p \cdot p^{m-1}} = \left(g^{p^{m-1}}\right)^p.$$

By the way the order is defined, we see that $g^{p^{m-1}}$ is an element of order p.

Exercise 8.0.1. Show that every subgroup of a p-group is itself a p-group.

Lemma 8.0.3. The homomorphic image of a p-group is itself a p-group.

Proof. Let G be a p-group and $\varphi : G \to H$ be a surjective homomorphism and let $h \in H$ be arbitrary. Since φ is surjective, let $g \in G$ be such that $\varphi(g) = h$. We now define $\psi : \langle g \rangle \to \langle h \rangle$ by $\psi(g^\alpha) = h^\alpha$. Note that ψ is a homomorphism since

$$\psi(g^\alpha g^\beta) = \psi(g^{alpha+\beta}) = h^{\alpha+\beta} = h^\alpha h^\beta = \psi(g^\alpha)\psi(g^\beta).$$

It is also immediate that ψ is surjective.

Now, $\ker(\psi)$ is a subgroup of $\langle g\rangle$ and $|\langle g\rangle| = p^m$ for some nonnegative integer m. By Lagrange's Theorem, it must be the case that $|\ker(\psi)| = p^k$ for some nonnegative integer $k \leq m$. By The First Isomorphism, we can conclude that

$$|\langle h\rangle| = \frac{|\langle g\rangle|}{|\ker\psi|} = \frac{p^m}{p^k} = p^{m-k}.$$

That is, the order of h is a power of p, and so H is a p-group. □

Lemma 8.0.4. A finite Abelian group G is a p-group if and only if the order of G is a power of p.

Proof. First, suppose that $|G| = p^k$ for some positive integer k. By Lagrange's Theorem, the order of any element of G must divide p^k, and hence must be itself a power of p.

Otherwise, suppose $|G|$ is not a power of p. It follows that $|G|$ must have some other prime divisor, q. By Lemma 8.0.1, G must have an element of order q, which means that G is not a p-group. □

The main structural fact about p-groups (Lemma 8.0.7) is that they can be decomposed into a direct product of cyclic groups, where each factor has some power of p as its order. Before we move to prove this result, we start with some motivating examples. The main idea is to find an element of maximal order and prove that the cyclic group generated by that element of maximal of order can be factored out of the group.

Example 8.0.5. If G is an Abelian group of order 4, then either $G \cong \mathbb{Z}_4$ or $G \cong \mathbb{Z}_2 \times \mathbb{Z}_2$. ⊣

Proof. If G has an element of order 4, then $G \cong \mathbb{Z}_4$, and we're done. So suppose G has no element of order 4. Since the identity element is the only element of order 1 and the order of any element must divide the order of the group, we know that every non-identity element of G must have order 2. So we can write $G = \{1, g_1, g_2, g_3\}$ where each g_j is an element of order 2. Note that $\langle g_1\rangle \cong \mathbb{Z}_2$ and consider $\langle g_2\rangle$. We claim that $G \cong \langle g_1\rangle \times \langle g_2\rangle \cong \mathbb{Z}_2 \times \mathbb{Z}_2$. Note that $\langle g_1\rangle \cap \langle g_2\rangle = \{1\}$ and that, by the uniqueness of the identity and the uniqueness of inverses, it must be the case that $g_1 g_2 = g_3$. It is thus straightforward to verify that the mapping $(g_1^p, g_2^q) \mapsto g_1^p g_2^q$, $\langle g_1\rangle \times \langle g_2\rangle \to G$, is an isomorphism. □

Remark. Note also here that $\mathbb{Z}_4 \ncong \mathbb{Z}_2 \times \mathbb{Z}_2$. So, even though $\mathbb{Z}_4$ has an element of order 2, we cannot factor out a cyclic group of order 2 from $\mathbb{Z}_4$ in the way we did in Example 8.0.5. This is why we must look for elements of maximal order.

The relatively simplicity of Example 8.0.5 doesn't give us a nuanced enough picture of the state of affairs. For example, once we picked out an element of maximal order (of which all non-identity elements were valid candidates), we simply chose some other arbitrary point from the group to decompose it. This may lead one to think that one can simply pick a sequence of distinct elements of relative maximal order to generate the cyclic subgroups in the direct product. This approach, however, does not work, in general.

Example 8.0.6. Let $G = \mathbb{Z}_8 \times \mathbb{Z}_4 \times \mathbb{Z}_4$. There are three elements $g_1, g_2, g_3 \in G$ of orders 8, 4, and 4, respectively, with the property that $g_j \notin \langle g_k \rangle$ when $j \neq k$ and for which the natural mapping

$$(ag_1, bg_2, cg_3) \overset{\varphi}{\mapsto} ag_1 + bg_2 + cg_3,\ \langle g_1 \rangle \times \langle g_2 \rangle \times \langle g_3 \rangle \to G,$$

is not an isomorphism. Indeed, consider $g_1 = (1, 1, 2)$, $g_2 = (2, 0, 0)$, and $g_3 = (2, 2, 2)$. These elements are as claimed, as we'll demonstrate below.

Note that the sequence of iterates of g_1 is

$$(1, 1, 2) \to (2, 2, 0) \to (3, 3, 2) \to (4, 0, 0) \to (5, 1, 2) \to (6, 2, 0) \to (7, 3, 2) \to (0, 0, 0),$$

the sequence of iterates of g_2 is

$$(2, 0, 0) \to (4, 0, 0) \to (6, 0, 0) \to (0, 0, 0),$$

and the sequence of iterates of g_3 is

$$(2, 2, 2) \to (4, 0, 0) \to (6, 2, 2) \to (0, 0, 0).$$

So $\text{ord}(g_1) = 8$, $\text{ord}(g_2) = 4$, and $\text{ord}(g_2) = 4$. Moreover, $g_j \notin \langle g_k \rangle$ when $j \neq k$.

Note also that

$$\langle g_1 \rangle \cap \langle g_2 \rangle \cap \langle g_3 \rangle = \{(0, 0, 0), (4, 0, 0)\}.$$

It follows that, after replacing the ordered triple (x, y, z) with the string xyz,

$$\ker(\varphi) = \{(000, 000, 000), (000, 400, 400), (400, 000, 400), (400, 400, 000)\}.$$

Hence,

$$[\langle g_1 \rangle \times \langle g_2 \rangle \times \langle g_3 \rangle : \ker(\varphi)] = \frac{|\langle g_1 \rangle \times \langle g_2 \rangle \times \langle g_3 \rangle|}{|\ker(\varphi)|} = \frac{8 \cdot 4 \cdot 4}{4} = 64 \neq 8 \cdot 4 \cdot 4 = |G|.$$

Therefore, φ is not an isomorphism. ⊣

As it will turn out, if we let g_1 be as it is (an element of maximal order in G), and we can find some $h \in G$ of order p for which $h \notin \langle g_1 \rangle$, then we get a workable path to the goal. That will be the general strategy in our proof of Lemma 8.0.7.

Lemma 8.0.7. If G is a finite Abelian p-group, then there is some sequence $a_1, a_2, \ldots, a_m$ of positive integers for which

$$G \cong \mathbb{Z}_{p^{a_1}} \times \mathbb{Z}_{p^{a_2}} \times \cdots \times \mathbb{Z}_{p^{a_m}}.$$

Proof. By Lemma 8.0.4, there must be some positive integer n such that $|G| = p^n$. Hence, we will proceed by induction. We will also prove an auxiliary assertion in conjunction with the statement of the lemma: under the assumption that g is an element of the finite Abelian p-group G of maximal order, then $G \cong \langle g \rangle \times H$ for some subgroup H of G where $(g^\alpha, h) \mapsto g^\alpha h$, $\langle g \rangle \times H \to G$, is a witnessing isomorphism. We proceed now by way of induction.

Note that, if $|G| = p$, then G is cyclic and, thus, $G \cong \mathbb{Z}_p \times \{e_G\} \cong \mathbb{Z}_p$ and both the auxiliary claim and the statement of the lemma are satisfied.

Let $k \geq 1$ and suppose the asserted auxiliary claim and the statement of the lemma hold for all groups of order p^n for each $n \leq k$. Suppose that $|G| = p^{k+1}$ and let $g \in G$ be of maximal order. Note that $\text{ord}(g) = p^r$ for some positive integer r, and so $\langle g \rangle \cong \mathbb{Z}_{p^r}$. If $r = k + 1$, then $G \cong \mathbb{Z}_{p^{k+1}} \times \{e_G\} \cong \mathbb{Z}_{p^{k+1}}$ and we are done.

So suppose $r < k + 1$. It follows that $\langle g \rangle$ is a proper subgroup of G.

We argue now that there is some $h \in G$ with $h \notin \langle g \rangle$ for which $\text{ord}(h) = p$. Since $\langle g \rangle$ is a proper subgroup of G, $G/\langle g \rangle$ is a nontrivial p-group by Lemma 8.0.3. Consequently, there is some element $x\langle g \rangle \in G/\langle g \rangle$, where $x \in G$, of order p. In particular, $x\langle g \rangle \neq \langle g \rangle$ and

$$(x\langle g \rangle)^p = x^p \langle g \rangle = \langle g \rangle.$$

It follows that $x \notin \langle g \rangle$ and that $x^p \in \langle g \rangle$. Hence, $x^p = g^j$ for some integer j. Let b be the positive integer for which $\text{ord}(x) = p^b$. Since g was assumed to be of maximal order,

$$p^b = \text{ord}(x) \leq \text{ord}(g) = p^r.$$

Moreover,

$$e_G = x^{p^b} = x^{p \cdot p^{b-1}} = (x^p)^{p^{b-1}},$$

and we can see that the order of x^p is $p^{b-1} < p^b \leq p^r$. It follows that the order of g^j is $p^{b-1} < p^r$, and so $\langle g^j \rangle$ is a proper subgroup of $\langle g \rangle$. Via the isomorphism $g^\alpha \mapsto \alpha$, $\langle g \rangle \to \mathbb{Z}_{p^r}$, we see that $\gcd(j, p^r) > 1$ by Lemma 4.4.13. Hence, p divides j and so we can write $j = pt$ where t is an integer. So let $h = xg^{-t}$. Since $x \notin \langle g \rangle$, $h \notin \langle g \rangle$. In particular, $h \neq e_G$. Note that

$$h^p = (xg^{-t})^p = x^p g^{-pt} = g^j g^{-j} = e_G.$$

Hence, $\text{ord}(h) = p$.

We now consider $\langle h \rangle$ and argue that $\langle g \rangle \cap \langle h \rangle = \{e_G\}$. Indeed, suppose $h^\alpha \in \langle g \rangle$ where $0 \leq \alpha < p$ and let β be an integer for which $h^\alpha = g^\beta$. Note then that

$$g^\beta = h^\alpha = \left(xg^{-t}\right)^\alpha = x^\alpha g^{-\alpha t} \implies x^\alpha = g^{\beta + \alpha t} \in \langle g \rangle.$$

Since the order of $x\langle g \rangle$ is p and $0 \leq \alpha < p$, it must be the case that $\alpha = 0$. Hence, $\langle h \rangle \cap \langle g \rangle = \{e_G\}$.

Now, since $\langle g \rangle \cap \langle h \rangle = \{e_G\}$ and the smallest positive integer α for which $g^\alpha = e_G \in \langle h \rangle$ is p^r, we see that $g^\beta \notin \langle h \rangle$ for every positive integer $\beta < p^r$. It follows that

$$\text{ord}\left(g\langle h \rangle\right) = p^r = \text{ord}(g).$$

Moreover, since $\text{ord}\left(y\langle h \rangle\right) \leq \text{ord}(y) \leq p^r$ for any $y \in G$, we see that $g\langle h \rangle$ is of maximal order in $G/\langle h \rangle$. Combined with the fact that

$$|G/\langle h \rangle| = \frac{|G|}{|\langle h \rangle|} = \frac{p^{k+1}}{p} = p^k,$$

we can apply the inductive hypothesis to conclude that

$$G/\langle h \rangle \cong \langle g\langle h \rangle \rangle \times K$$

where K is a subgroup of $G/\langle h\rangle$ and the natural mapping

$$(g^\alpha\langle h\rangle, A) \mapsto g^\alpha\langle h\rangle A,\ \langle g\langle h\rangle\rangle \times K \to G/\langle h\rangle,$$

is a witnessing isomorphism. It follows that $\langle g\langle h\rangle\rangle \cap K = \{\langle h\rangle\}$ since the kernel of the isomorphism is trivial by Exercise 4.3.8. Let H be the subgroup of G guaranteed by Exercise 4.3.5 for which $K = \{y\langle h\rangle : y \in H\}$. Note, in particular, that $\langle h\rangle \subseteq H$.

We now claim that $\langle g\rangle \cap H = \{e_G\}$. So consider $y \in \langle g\rangle \cap H$ and note that $y\langle h\rangle \in \langle g\langle h\rangle\rangle \cap K = \{\langle h\rangle\}$. Hence, $y\langle h\rangle = \langle h\rangle$ which implies that $y \in \langle h\rangle$. Now we have that $y \in \langle g\rangle \cap \langle h\rangle = \{e_G\}$, and so $y = e_G$. That is, $\langle g\rangle \cap H = \{e_G\}$.

To see that $G = \langle g\rangle H$, let $u \in G$ be arbitrary and note that there is some integer α and some $y \in H$ for which

$$u\langle h\rangle = g^\alpha\langle h\rangle \cdot y\langle h\rangle = g^\alpha y\langle h\rangle.$$

It follows that $u = g^\alpha y h^\beta$ for some integer β and thus, since $\langle h\rangle \subseteq H$, we see that $u \in \langle g\rangle H$. Hence, by Theorem 7.1.6,

$$G \cong \langle g\rangle \times H$$

and the natural mapping $(g^\alpha, \zeta) \mapsto g^\alpha\zeta$, $\langle g\rangle \times H \to G$, is a witnessing isomorphism.

Finally, since H is a proper subgroup of G and H is an Abelian p-group with order $\leq p^k$, the inductive hypothesis applies to assert that H is a direct product of cyclic groups where each factor has as its order some power of p. Since $\langle g\rangle$ is a cyclic group with order p^r and $G \cong \langle g\rangle \times H$, the proof is complete. □

Exercise 8.0.2. Let G be an Abelian group, p be a prime number, and

$$H = \{g \in G : \exists k \in \mathbb{Z}\ (g^{p^k} = e_G)\}.$$

Show that H is a subgroup of G.

Theorem 8.0.8 (The Fundamental Theorem of Finite Abelian Groups)**.** Let G be a finite Abelian group. Then there is a sequence $p_1, p_2, \ldots, p_n$ of (not necessarily distinct) primes and a sequence $a_1, a_2, \ldots, a_n$ of positive integers for which

$$G \cong \mathbb{Z}_{p_1^{a_1}} \times \mathbb{Z}_{p_2^{a_2}} \times \cdots \times \mathbb{Z}_{p_n^{a_n}}.$$

In other words, every finite Abelian group is isomorphic to a direct product of cyclic groups.

Proof. For every integer $n > 1$, consider the prime factorization

$$n = q_1^{b_1} q_2^{b_2} \cdots q_m^{b_m}$$

of n and let $\mathrm{pdl}(n) = m$. The function pdl is well-defined by The Fundamental Theorem of Arithmetic. We will use induction on $\mathrm{pdl}(n)$.

Suppose G is a finite Abelian group of order n with $\mathrm{pdl}(n) = 1$. Then $n = q^b$ for some positive inter b and some prime q. It follows that G is a finite Abelian q-group and Lemma 8.0.7 applies to guarantee the theorem statement.

Suppose now that, for $k \geq 1$, every finite Abelian group of order n with $\text{pdl}(n) \leq k$ satisfies the theorem statement. Then let G be a finite Abelian group of order n where $\text{pdl}(n) = k+1$ and let

$$n = q_1^{b_1} q_2^{b_2} \cdots q_k^{b_k} q_{k+1}^{b_{k+1}}$$

be the prime factorization of n. Set $m = q_1^{b_1} q_2^{b_2} \cdots q_k^{b_k}$ and $r = q_{k+1}^{b_{k+1}}$ for convenience, and let

$$H = \{g \in G : g^m = e_G\}$$

and

$$K = \left\{ g \in G : \exists j \in \mathbb{Z} \ \left(g^{q_{k+1}^j} = e_G \right) \right\}.$$

Note that K is a subgroup of G by Exercise 8.0.2 and is itself a finite Abelian q_{k+1}-group. By Lemma 8.0.7, there is a sequence $a_1, a_2, \ldots, a_\ell$ of positive integers for which

$$K \cong \mathbb{Z}_{q_{k+1}^{a_1}} \times \mathbb{Z}_{q_{k+1}^{a_2}} \times \cdots \times \mathbb{Z}_{q_{k+1}^{a_\ell}}.$$

So, to finish the proof, we will show that H is a subgroup of G with order m and that

$$G \cong H \times K.$$

Indeed, this will suffice since the order of H is m and $\text{pdl}(m) = k$, which will allow us to apply the inductive hypothesis on H.

To see that H is a subgroup of G, first observe that $e_G \in H$. Then suppose $g, h \in H$ and consider the fact that

$$(gh^{-1})^m = g^m h^{-m} = e_G \cdot h^{-m} = h^m \cdot h^{-m} = e_G.$$

Hence, $gh^{-1} \in H$, and Proposition 4.2.4 applies.

We now aim to show that $H \cap K = \{e_G\}$. First, note that any $g \in K$ has the property that $g^r = e_G$. Indeed, if $g \in K$ then we can let λ be the least positive integer for which

$$g^{q_{k+1}^\lambda} = e_G.$$

It follows that $\text{ord}(g) = q_{k+1}^\lambda$ and that, since the order of any element of a finite group must divide the order of the group, $q_{k+1}^\lambda \leq r = q_{k+1}^{b_{k+1}}$. Then,

$$g^r = g^{q_{k+1}^{b_{k+1}}} = \left(g^{q_{k+1}^\lambda} \right)^{q_{k+1}^{b_{k+1}-\lambda}} = e_G.$$

Now, suppose $g \in H \cap K$. Observe that $\gcd(m, r) = 1$, so Bézout's Identity informs us that there are integers x and y for which $mx + ry = 1$. Hence,

$$g = g^{mx+ry} = g^{mx} g^{ry} = (g^m)^x (g^r)^y = e_G.$$

That is, $H \cap K = \{e_G\}$.

To see that $G = HK$, let $g \in G$ be arbitrary and note that the order of g must be of the form

$$q_1^{w_1} q_2^{w_2} \cdots q_k^{w_k} q_{k+1}^{w_{k+1}},$$

where w_j is an integer with $0 \leq w_j \leq b_j$ for each positive integer $j \leq k+1$. Note that

$$\gcd(q_1^{w_1} q_2^{w_2} \cdots q_k^{w_k}, q_{k+1}^{w_{k+1}}) = 1,$$

so Bézout's Identity guarantees integers x and y with the property that

$$x q_1^{w_1} q_2^{w_2} \cdots q_k^{w_k} + y q_{k+1}^{w_{k+1}} = 1.$$

Let $h_1 = g^{y q_{k+1}^{w_{k+1}}}$ and $h_2 = g^{x q_1^{w_1} q_2^{w_2} \cdots q_k^{w_k}}$. Then

$$g = g^{x q_1^{w_1} q_2^{w_2} \cdots q_k^{w_k} + y q_{k+1}^{w_{k+1}}} = g^{x q_1^{w_1} q_2^{w_2} \cdots q_k^{w_k}} \cdot g^{y q_{k+1}^{w_{k+1}}} = h_2 h_1 = h_1 h_2.$$

To see that $h_1 \in H$, note that

$$\begin{aligned} h_1^m = \left(g^{y q_{k+1}^{w_{k+1}}}\right)^m &= g^{ym q_{k+1}^{w_{k+1}}} \\ &= g^{y q_1^{b_1} q_2^{b_2} \cdots q_k^{b_k} q_{k+1}^{w_{k+1}}} \\ &= \left(g^{q_1^{w_1} q_2^{w_2} \cdots q_k^{w_k} q_{k+1}^{w_{k+1}}}\right)^{y q_1^{b_1 - w_1} q_2^{b_2 - w_2} \cdots q_k^{b_k - w_k}} \\ &= e_G. \end{aligned}$$

To see that $h_2 \in K$, note that

$$h_2^{q_{k+1}^{w_{k+1}}} = \left(g^{x q_1^{w_1} q_2^{w_2} \cdots q_k^{w_k}}\right)^{q_{k+1}^{w_{k+1}}} = \left(g^{q_1^{w_1} q_2^{w_2} \cdots q_k^{w_k} q_{k+1}^{w_{k+1}}}\right)^x = e_G.$$

Thus, $G = HK$.

We can now apply Theorem 7.1.6 to establish that

$$G \cong H \times K.$$

The final thing to show is that the order of H is m. First, note that the order of any element of H must divide m by Lemma 4.4.9. Since $\gcd(m, q_{k+1}) = 1$, H can have no elements of order q_{k+1}. Hence, by Lemma 8.0.1, $q_{k+1} \nmid |H|$. A similar argument establishes the fact that $q_j \nmid |K|$ for every positive integer $j \leq k$. Now, as

$$q_1^{b_1} q_2^{b_2} \cdots q_k^{b_k} q_{k+1}^{b_{k+1}} = |G| = |H \times K| = |H| \cdot |K|,$$

$|K| = q_{k+1}^{b_{k+1}}$ and $|H| = q_1^{b_1} q_2^{b_2} \cdots q_k^{b_k} = m$. Finally, as $\text{pdl}(m) = k$, the inductive hypothesis applies to guarantee that H is isomorphic to a direct product of finite cyclic groups, finishing the proof. □

Part II

Rings & Fields

Chapter 9

Introducing Rings

As per their definitions, groups are objects with only one relevant binary operation. However, as we will note, standard mathematical objects of interest, like $\mathbb{Z}$ and $\mathbb{R}$, have another basic binary operation relevant to their arithmetic: multiplication. For the remainder of the course, we will study the resulting structure when two binary operations, which interact relatively well with each other, are present.

9.1 Rings

Definition 9.1.1. Let $(R, +)$ be an Abelian[1] group and suppose $\cdot : R^2 \to R$ is a binary operation. Then we say that the triple $(R, +, \cdot)$ is a *ring* if the following properties hold:

- (Associativity of $\cdot$) For all $x, y, z \in R$, $x(yz) = (xy)z$.
- (Distribution of $\cdot$ over $+$) For all $x, y, z \in R$, $x(y+z) = xy+xz$ and $(x+y)z = xz+yz$.

When $(R, +, \cdot)$ is a ring, we will refer to $+$ as the *group operation* and $\cdot$ as the *ring operation*.

Notation. Since the underlying group structure in a ring is an Abelian group, which we will typically write additively, we will generally let 0_R denote the identity element of $(R, +)$.

Remark. Any Abelian group can be turned into a ring with a trivial product. In particular, let $(G, +)$ be an Abelian group and define $\cdot$ on G by the rule $x \cdot y = 0_G$. Then $(G, +, \cdot)$ is a ring.

Exercise 9.1.1. Define $\diamond$ on $\mathbb{Z}_3$ via a Cayley table:

$\diamond$	0	1	2
0	0	0	0
1	0	2	0
2	0	0	0

Is $(\mathbb{Z}_3, +, \diamond)$ a ring? Why or why not?

[1] We will comment in Proposition 9.1.13 that, even if we were to drop this condition, in most cases of interest, it holds necessarily.

Proposition 9.1.2. Let $(R, +, \cdot)$ be a ring.

(a) For any $x \in R$, $x \cdot 0_R = 0_R \cdot x = 0_R$.

(b) For any $x, y \in R$, $(-x)y = x(-y) = -xy$.

(c) For any $x, y \in R$, $(-x)(-y) = xy$.

Proof. (a) Observe that

$$x \cdot 0_R = x \cdot (0_R + 0_R) = x \cdot 0_R + x \cdot 0_R$$

and that

$$0_R \cdot x = (0_R + 0_R) \cdot x = 0_R \cdot x + 0_R \cdot x.$$

In either case, by appealing to additive inverses, we obtain that

$$x \cdot 0_R = 0_R = 0_R \cdot x.$$

(b) Note that

$$xy + (-x)y = (x - x)y = 0_R y = 0_R$$

and that

$$xy + x(-y) = x(y - y) = x \cdot 0_R = 0_R.$$

By appealing to additive inverses, we see that

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

(c) Note that

$$\begin{aligned}(-x)(-y) - xy &= (-x)(-y) + (-xy) \\ &= (-x)(-y) + (-x)y \\ &= (-x)(-y + y) \\ &= (-x) \cdot 0_R \\ &= 0_R.\end{aligned}$$

Hence, $(-x)(-y) = xy$. □

Immediately, one can recognize that $(\mathbb{Z}, +, \cdot)$ and $(\mathbb{R}, +, \cdot)$ with the usual operations of addition $+$ and multiplication $\cdot$ are rings. These rings also have a collection of relatively special properties, which we'll give names to below.

Notation. For any given ring R, we will typically use R^* to denote the set $R \setminus \{0_R\}$.

Definition 9.1.3. Suppose $(R, +, \cdot)$ is a ring. We say that R is a ring *with unity* (or a ring *with identity*) if there is some element $1_R \in R^*$ which has the property that, for all $x \in R$, $x \cdot 1_R = 1_R \cdot x = x$.

Remark. Some authors do not impose the restriction that the identity element is nonzero, but there is only one ring satisfying the less restrictive condition. In particular, $R = \{0\}$ with the usual $+$ and $\cdot$ for real numbers is known as the *zero ring*. It does satisfy the property that $x \cdot 0 = 0 \cdot x = x$ for all $x \in R$ since R only consists of the zero element. So, in this way, 0 looks like a multiplicative identity.

To see that the zero ring is the only ring with the property that the additive identity and the multiplicative identity agree, suppose $(R, +, \cdot)$ is a ring where $0_R = 1_R$. Then let $x \in R$ be arbitrary and note that $x = x \cdot 1_R = x \cdot 0_R = 0_R$. That is, $R = \{0_R\}$.

Sanity Check 9.1.2. If R is a ring with unity, show that the multiplicative identity is unique.

Notation. As in Definition 9.1.3, we will typically use 1_R to denote the multiplicative identity of a ring R with unity.

Definition 9.1.4. A ring $(R, +, \cdot)$ is said to be *commutative* if, for all $x, y \in R$, $xy = yx$.

One will immediately note that $\mathbb{Z}$ and $\mathbb{R}$ are examples of commutative rings.

Example 9.1.5. Let $\mathsf{M}(2, \mathbb{R})$ be the set of all 2×2 matrices with real entries, let $+$ denote the usual entry-wise addition between matrices, and $\cdot$ be the usual matrix multiplication. Then $(\mathsf{M}(2, \mathbb{R}), +, \cdot)$ is a non-commutative ring. ⊣

In fact, matrix rings can be defined over any ring. For convenience, we'll limit our attention to 2×2 matrices.

Definition 9.1.6. Let $(R, +, \cdot)$ be a ring and define

$$\mathsf{M}(2, R) = \left\{ \begin{bmatrix} w & x \\ y & z \end{bmatrix} : w, x, y, z \in R \right\}.$$

Then define $+$ and $\cdot$ by the rules

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} + \begin{bmatrix} w & x \\ y & z \end{bmatrix} = \begin{bmatrix} a+w & b+x \\ c+y & d+z \end{bmatrix}$$

and

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \begin{bmatrix} w & x \\ y & z \end{bmatrix} = \begin{bmatrix} aw+by & ax+bz \\ cw+dy & cx+dz \end{bmatrix}.$$

Proposition 9.1.7. Let $(R, +, \cdot)$ be a ring. Then $(\mathsf{M}(2, R), +, \cdot)$ is a ring.

Proof. Note that $(\mathsf{M}(2, R), +)$ is isomorphic to a fourfold direct product of $(R, +)$, which is an Abelian group. Hence, $(\mathsf{M}(2, R), +)$ is an Abelian group with additive identity

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

To see that $\cdot$ is associative, observe that

$$\begin{aligned}
\begin{bmatrix} a & b \\ c & d \end{bmatrix}\left(\begin{bmatrix} s & t \\ u & v \end{bmatrix}\begin{bmatrix} w & x \\ y & z \end{bmatrix}\right) &= \begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} sw+ty & sx+tz \\ uw+vy & ux+vz \end{bmatrix} \\
&= \begin{bmatrix} a(sw+ty)+b(uw+vy) & a(sx+tz)+b(ux+vz) \\ c(sw+ty)+d(uw+vy) & c(sx+tz)+d(ux+vz) \end{bmatrix} \\
&= \begin{bmatrix} asw+aty+buw+bvy & asx+atz+bux+bvz \\ csw+cty+duw+dvy & csx+ctz+dux+dvz \end{bmatrix}
\end{aligned}$$

and that

$$\begin{aligned}
\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} s & t \\ u & v \end{bmatrix}\right)\begin{bmatrix} w & x \\ y & z \end{bmatrix} &= \begin{bmatrix} as+bu & at+bv \\ cs+du & ct+dv \end{bmatrix}\begin{bmatrix} w & x \\ y & z \end{bmatrix} \\
&= \begin{bmatrix} (as+bu)w+(at+bv)y & (as+bu)x+(at+bv)z \\ (cs+du)w+(ct+dv)y & (cs+du)x+(ct+dv)z \end{bmatrix} \\
&= \begin{bmatrix} asw+buw+aty+bvy & asx+bux+atz+bvz \\ csw+duw+cty+dvy & csx+dux+ctz+dvz \end{bmatrix} \\
&= \begin{bmatrix} asw+aty+buw+bvy & asx+atz+bux+bvz \\ csw+cty+duw+dvy & csx+ctz+dux+dvz \end{bmatrix} \\
&= \begin{bmatrix} a & b \\ c & d \end{bmatrix}\left(\begin{bmatrix} s & t \\ u & v \end{bmatrix}\begin{bmatrix} w & x \\ y & z \end{bmatrix}\right).
\end{aligned}$$

For the distributive properties, note that

$$\begin{aligned}
\begin{bmatrix} a & b \\ c & d \end{bmatrix}\left(\begin{bmatrix} s & t \\ u & v \end{bmatrix}+\begin{bmatrix} w & x \\ y & z \end{bmatrix}\right) &= \begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} s+w & t+x \\ u+y & v+z \end{bmatrix} \\
&= \begin{bmatrix} a(s+w)+b(u+y) & a(t+x)+b(v+z) \\ c(s+w)+d(u+y) & c(t+x)+d(v+z) \end{bmatrix} \\
&= \begin{bmatrix} as+aw+bu+by & at+ax+bv+bz \\ cs+cw+du+dy & ct+cx+dv+dz \end{bmatrix} \\
&= \begin{bmatrix} as+bu & at+bv \\ cs+du & ct+dv \end{bmatrix}+\begin{bmatrix} aw+by & ax+bz \\ cw+dy & cx+dz \end{bmatrix} \\
&= \begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} s & t \\ u & v \end{bmatrix}+\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} w & x \\ y & z \end{bmatrix}
\end{aligned}$$

and that

$$\begin{aligned}\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix} + \begin{bmatrix} s & t \\ u & v \end{bmatrix}\right)\begin{bmatrix} w & x \\ y & z \end{bmatrix} &= \begin{bmatrix} a+s & b+t \\ c+u & d+v \end{bmatrix}\begin{bmatrix} w & x \\ y & z \end{bmatrix} \\ &= \begin{bmatrix} (a+s)w+(b+t)y & (a+s)x+(b+t)z \\ (c+u)w+(d+v)y & (c+u)x+(d+v)z \end{bmatrix} \\ &= \begin{bmatrix} aw+sw+by+ty & ax+sx+bz+tz \\ cw+uw+dy+vy & cx+ux+dz+vz \end{bmatrix} \\ &= \begin{bmatrix} aw+by & ax+bz \\ cw+dy & cx+dz \end{bmatrix} + \begin{bmatrix} sw+ty & sx+tz \\ uw+vy & ux+vz \end{bmatrix} \\ &= \begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} w & x \\ y & z \end{bmatrix} + \begin{bmatrix} s & t \\ u & v \end{bmatrix}\begin{bmatrix} w & x \\ y & z \end{bmatrix}.\end{aligned}$$

Therefore, $(\mathsf{M}(2,R),+,\cdot)$ is a ring. □

Definition 9.1.8. Suppose $(R,+,\cdot)$ is a ring. If $x,y \in R^*$ are such that $xy = 0_R$, then we refer to x as a *left zero divisor* and y as a *right zero divisor*. We will say that an element $x \in R^*$ is a *zero divisor* if it is either a left zero divisor or a right zero divisor.

Definition 9.1.9. We say that a commutative ring $(R,+,\cdot)$ with unity is an *integral domain* if R has no zero divisors.

Remark. An integral domain is precisely the context where the so-called zero-product law holds. In particular, if $(R,+,\cdot)$ is an integral domain and $x,y \in R$ are such that $xy = 0_R$, then either $x = 0_R$ or $y = 0_R$.

Note that $\mathbb{Z}$ and $\mathbb{R}$ are standard examples of integral domains.

Example 9.1.10. With the usual operations of addition $\oplus$ and multiplication $*$ (as detailed in Example 4.1.8), $(\mathbb{Z}_{12},\oplus,*)$ is a commutative ring with unity which is not an integral domain. Indeed, note that $3 * 4 = 0$ and that neither 3 nor 4 are zero. ⊣

Definition 9.1.11. In a ring R with unity, we say that an element $x \in R$ is *left-invertible* if there exists $y \in R$ such that $yx = 1_R$; *right-invertible* if there exists $y \in R$ such that $xy = 1_R$; or *invertible* if there exists $y \in R$ such that $xy = yx = 1_R$.

Exercise 9.1.3. Show that, if an element x of a ring R with unity is both left- and right-invertible, then it is invertible. That is, show that, if $y,z \in R$ are such that $xy = 1_R$ and $zx = 1_R$, then $y = z$.

Sanity Check 9.1.4. Suppose R is a ring with unity and let $x \in R^*$. If there is some $y \in R^*$ for which $xy = yx = 1_R$, show that y is the unique element satisfying that property.

Proposition 9.1.12. If an element x of a ring with unity R is a left zero divisor (respectively, right zero divisor), then it is not left-invertible (respectively, right-invertible). Hence, any element x of a ring with unity which is a zero divisor is not invertible.

Proof. If x is a left zero divisor, then there is a nonzero $y \in R$ for which $xy = 0_R$. Note then that, for any $z \in R$,

$$(zx)y = z(xy) = z \cdot 0_R = 0_R,$$

and, since $y \neq 0_R$, $zx \neq 1_R$. Since z was arbitrary, x is not left-invertible.

In a similar way, suppose x is a right zero divisor and let $y \in R^*$ be such that $yx = 0_R$. Then, for any $z \in R$,

$$y(xz) = (yx)z = 0_R \cdot z = 0_R.$$

Since $y \neq 0_R$, we see that $xz \neq 1_R$. Thus, as z was arbitrary, x is not right-invertible.

Finally, since any invertible element is both left- and right-invertible, and a zero divisor is either a left zero divisor or a right zero divisor, we see that no zero divisor can be invertible. □

Recall that any Abelian group can be trivially turned into a ring by using the zero product as the ring operation. In such a context, every element is trivially seen to be a zero divisor. However, as we will show below with Example 9.1.21, there are rings in which every element is a zero divisor where the ring operation is not the zero product.

We also comment here that if a group is endowed with another binary operation which satisfies the distributivity properties over the group operation and has some element which is not a zero divisor, then the group must necessarily be Abelian.

Proposition 9.1.13. Let $(G, +)$ be a group and suppose $\cdot : G^2 \to G$ is such that $x(y+z) = xy + xz$ and $(x+y)z = xz + yz$ for each $x, y, z \in G$. Additionally suppose that there is some $g \in G$ such that, for every $x \in G \setminus \{0_G\}$, $gx \neq 0_G$. Then G is Abelian.

Proof. Let $g \in G$ be such that, for every $x \in G \setminus \{0_G\}$, $gx \neq 0_G$, and suppose $x, y \in G$ are arbitrary. By the distributivity properties, note that

$$\begin{aligned} gx + gy + gx + gy &= g(x+y) + g(x+y) \\ &= (g+g)(x+y) \\ &= (g+g)x + (g+g)y \\ &= gx + gx + gy + gy. \end{aligned}$$

It follows that

$$\begin{aligned} g(y+x) &= gy + gx \\ &= -gx + gx + gy + gx + gy - gy \\ &= -gx + gx + gx + gy + gy - gy \\ &= gx + gy \\ &= g(x+y) \end{aligned}$$

and that

$$g(y+x) - g(x+y) = 0_G.$$

One can verify that the results of Proposition 9.1.2 hold in this context, and so

$$0_g = g(y+x) - g(x+y) = g(y+x) + g(-(x+y)) = g((y+x) - (x+y)).$$

By the assumption on g, we see that $(y+x)-(x+y)=0_G$. That is, $y+x=x+y$, establishing that $(G,+)$ is an Abelian group. □

Definition 9.1.14. Suppose $(R,+,\cdot)$ is a ring with unity. We say that R is a *division ring* if every nonzero element is invertible; that is, if, for every $x\in R^*$, there exists $y\in R$ such that $xy=yx=1_R$. In such a case, we will let x^{-1} denote the multiplicative inverse of $x\in R^*$.

Note that $(\mathbb{Z},+,\cdot)$ is not a division ring since any element of absolute value greater than one does not have a multiplicative inverse. On the other hand, $(\mathbb{R},+,\cdot)$ is a division ring.

Exercise 9.1.5. Suppose $(R,+,\cdot)$ is a ring with unity and let

$$\mathrm{U}(R)=\{x\in R:\exists y\in R\ (xy=yx=1_R)\},$$

the set of invertible elements of R. Show that $(\mathrm{U}(R),\cdot)$ is a group.

Exercise 9.1.6. Suppose $(R,+,\cdot)$ is a ring. Show that $(R,+,\cdot)$ is a division ring if and only if $(R^*,\cdot)$ is a group.

Definition 9.1.15. A *field* is a commutative division ring.

A standard example of a field is $(\mathbb{R},+,\cdot)$. There are also plenty of finite fields.

Example 9.1.16. For any prime p, $(\mathbb{Z}_p,\oplus,*)$ is a field. In fact, by Theorem 4.1.12, for $n\geq 2$, $(\mathbb{Z}_n,\oplus,*)$ is a field if and only if n is prime. ⊣

Aside from $(\mathbb{R},+,\cdot)$, $(\mathbb{Q},+,\cdot)$ and $(\mathbb{C},+,\cdot)$ are other important examples of infinite fields. We will study $\mathbb{Q}$ a bit more in detail later.

We provide an example showing that not every division ring is a field.

Example 9.1.17 (The Quaternions)**.** Let $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ be symbols and define

$$\mathbb{H}=\{a+b\mathbf{i}+c\mathbf{j}+d\mathbf{k}:a,b,c,d\in\mathbb{R}\}.$$

Define $+$ on $\mathbb{H}$ by the rule

$$\begin{aligned}&(a_1+b_1\mathbf{i}+c_1\mathbf{j}+d_1\mathbf{k})+(a_2+b_2\mathbf{i}+c_2\mathbf{j}+d_2\mathbf{k})\\&=(a_1+a_2)+(b_1+b_2)\mathbf{i}+(c_1+c_2)\mathbf{j}+(d_1+d_2)\mathbf{k}\end{aligned}$$

and $\cdot$ on $\mathbb{H}$ by the rule

$$\begin{aligned}&(a_1+b_1\mathbf{i}+c_1\mathbf{j}+d_1\mathbf{k})(a_2+b_2\mathbf{i}+c_2\mathbf{j}+d_2\mathbf{k})\\&=(a_1a_2-b_1b_2-c_1c_2-d_1d_2)\\&\quad+(a_1b_2+a_2b_1+c_1d_2-c_2d_1)\mathbf{i}\\&\quad+(a_1c_2+a_2c_1-b_1d_2+b_2d_1)\mathbf{j}\\&\quad+(a_1d_2+a_2d_1+b_1c_2-b_2c_1)\mathbf{k}.\end{aligned}$$

We then define *the quaternions* to be $(\mathbb{H},+,\cdot)$. ⊣

For a brief history and some details on the importance of the quaternions, see the corresponding Wikipedia article.

Before we move on to proving that $(\mathbb{H}, +, \cdot)$ is a ring, we provide some computational motivation for the product and offer a matrix representation. The symbols $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ have the properties that

$$\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -1, \quad \begin{cases} \mathbf{ij} = \mathbf{k} \\ \mathbf{jk} = \mathbf{i} \\ \mathbf{ki} = \mathbf{j} \end{cases}, \quad \text{and} \quad \begin{cases} \mathbf{ji} = -\mathbf{k} \\ \mathbf{kj} = -\mathbf{i} \\ \mathbf{ik} = -\mathbf{j} \end{cases}.$$

The multiplicative properties outlined above can be summarized with the diagram in Figure 9.1. Then the real coefficients commute with the symbols $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$. So, the defined product

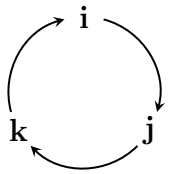


Figure 9.1: Basic Quaternion Product Diagram

is the standard distributive product of

$$(a_1 + b_1\mathbf{i} + c_1\mathbf{j} + d_1\mathbf{k})(a_2 + b_2\mathbf{i} + c_2\mathbf{j} + d_2\mathbf{k})$$

once one has simplified the expression using the rules outlined above. Explicitly,

$$\begin{aligned}
&(a_1 + b_1\mathbf{i} + c_1\mathbf{j} + d_1\mathbf{k})(a_2 + b_2\mathbf{i} + c_2\mathbf{j} + d_2\mathbf{k}) \\
&= a_1a_2 + a_1b_2\mathbf{i} + a_1c_2\mathbf{j} + a_1d_2\mathbf{k} \\
&\quad + b_1\mathbf{i}a_2 + b_1\mathbf{i}b_2\mathbf{i} + b_1\mathbf{i}c_2\mathbf{j} + b_1\mathbf{i}d_2\mathbf{k} \\
&\quad + c_1\mathbf{j}a_2 + c_1\mathbf{j}b_2\mathbf{i} + c_1\mathbf{j}c_2\mathbf{j} + c_1\mathbf{j}d_2\mathbf{k} \\
&\quad + d_1\mathbf{k}a_2 + d_1\mathbf{k}b_2\mathbf{i} + d_1\mathbf{k}c_2\mathbf{j} + d_1\mathbf{k}d_2\mathbf{k} \\
&= a_1a_2 + a_1b_2\mathbf{i} + a_1c_2\mathbf{j} + a_1d_2\mathbf{k} \\
&\quad + a_2b_1\mathbf{i} - b_1b_2 + b_1c_2\mathbf{k} - b_1d_2\mathbf{j} \\
&\quad + a_2c_1\mathbf{j} - b_2c_1\mathbf{k} - c_1c_2 + c_1d_2\mathbf{i} \\
&\quad + a_2d_1\mathbf{k} + b_2d_1\mathbf{j} - c_2d_1\mathbf{i} - d_1d_2 \\
&= a_1a_2 - b_1b_2 - c_1c_2 - d_1d_2 \\
&\quad + (a_1b_2 + a_2b_1 + c_1d_2 - c_2d_1)\mathbf{i} \\
&\quad + (a_1c_2 - b_1d_2 + a_2c_1 + b_2d_1)\mathbf{j} \\
&\quad + (a_1d_2 + b_1c_2 - b_2c_1 + a_2d_1)\mathbf{k}.
\end{aligned}$$

One important property of this product is the following:

$$\begin{aligned}
&(a+b\mathbf{i}+c\mathbf{j}+d\mathbf{k})(a-b\mathbf{i}-c\mathbf{j}-d\mathbf{k}) \\
&= a^2 - ab\mathbf{i} - ac\mathbf{j} - ad\mathbf{k} + ab\mathbf{i} + b^2 - bc\mathbf{k} + bd\mathbf{j} \\
&\quad + ac\mathbf{j} + bc\mathbf{k} + c^2 - cd\mathbf{i} + ad\mathbf{k} - bd\mathbf{j} + cd\mathbf{i} + d^2 \\
&= a^2 + b^2 + c^2 + d^2
\end{aligned}$$

and

$$\begin{aligned}
&(a-b\mathbf{i}-c\mathbf{j}-d\mathbf{k})(a+b\mathbf{i}+c\mathbf{j}+d\mathbf{k}) \\
&= a^2 + ab\mathbf{i} + ac\mathbf{j} + ad\mathbf{k} - ab\mathbf{i} + b^2 - bc\mathbf{k} + bd\mathbf{j} \\
&\quad - ac\mathbf{j} + bc\mathbf{k} + c^2 - cd\mathbf{i} - ad\mathbf{k} - bd\mathbf{j} + cd\mathbf{i} + d^2 \\
&= a^2 + b^2 + c^2 + d^2.
\end{aligned}$$

Now we discuss a matrix representation for the quaternions. Let

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

Then

$$\begin{aligned}
a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k} &= a\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + b\begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix} + c\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} + d\begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix} \\
&= \begin{bmatrix} a+bi & c+di \\ -c+di & a-bi \end{bmatrix}.
\end{aligned}$$

Recall the complex conjugate: $\overline{a+ib} = a - ib$. Hence, this matrix representation can be express more succinctly as

$$\begin{bmatrix} z & \zeta \\ -\overline{\zeta} & \overline{z} \end{bmatrix}$$

where $z = a + ib$ and $\zeta = c + id$. A property of the complex conjugate that we will use here is that the complex conjugate map preserves addition and multiplication:

$$\begin{aligned}
\overline{z+\zeta} &= \overline{a+ib+c+id} \\
&= \overline{(a+c)+i(b+d)} \\
&= (a+c)-i(b+d) \\
&= a+c-ib-id \\
&= a-ib+c-id \\
&= \overline{z}+\overline{\zeta}
\end{aligned}$$

and

$$\begin{aligned}
\overline{z\cdot\zeta} &= \overline{(a+ib)(c+id)} \\
&= \overline{(ac-bd)+i(ad+bc)} \\
&= (ac-bd)-i(ad+bc) \\
&= (a-ib)(c-id) \\
&= \overline{z}\cdot\overline{\zeta}.
\end{aligned}$$

To see that this matrix representation is a faithful representation of the quaternions, note that

$$\begin{aligned}
&(a_1\mathbf{1} + b_1\mathbf{i} + c_1\mathbf{j} + d_1\mathbf{k}) + (a_2\mathbf{1} + b_2\mathbf{i} + c_2\mathbf{j} + d_2\mathbf{k}) \\
&= \begin{bmatrix} a_1 + b_1 i & c_1 + d_1 i \\ -c_1 + d_1 i & a_1 - b_1 i \end{bmatrix} + \begin{bmatrix} a_2 + b_2 i & c_2 + d_2 i \\ -c_2 + d_2 i & a_2 - b_2 i \end{bmatrix} \\
&= \begin{bmatrix} (a_1 + a_2) + (b_1 + b_2)i & (c_1 + c_2) + (d_1 + d_2)i \\ -(c_1 + c_2) + (d_1 + d_2)i & (a_1 + a_2) - (b_1 + b_2)i \end{bmatrix} \\
&= (a_1 + a_2)\mathbf{1} + (b_1 + b_2)\mathbf{i} + (c_1 + c_2)\mathbf{j} + (d_1 + d_2)\mathbf{k}
\end{aligned}$$

and that,

$$\begin{aligned}
&(a_1\mathbf{1} + b_1\mathbf{i} + c_1\mathbf{j} + d_1\mathbf{k})(a_2\mathbf{1} + b_2\mathbf{i} + c_2\mathbf{j} + d_2\mathbf{k}) \\
&= \begin{bmatrix} a_1 + b_1 i & c_1 + d_1 i \\ -c_1 + d_1 i & a_1 - b_1 i \end{bmatrix} \begin{bmatrix} a_2 + b_2 i & c_2 + d_2 i \\ -c_2 + d_2 i & a_2 - b_2 i \end{bmatrix} \\
&= \begin{bmatrix} z_1 & \zeta_1 \\ -\overline{\zeta_1} & \overline{z_1} \end{bmatrix} \begin{bmatrix} z_2 & \zeta_2 \\ -\overline{\zeta_2} & \overline{z_2} \end{bmatrix} \\
&= \begin{bmatrix} z_1 z_2 - \zeta_1\overline{\zeta_2} & z_1\zeta_2 + \zeta_1\overline{z_2} \\ -\overline{\zeta_1} z_2 - \overline{z_1}\cdot\overline{\zeta_2} & -\overline{\zeta_1}\zeta_2 + \overline{z_1}\cdot\overline{z_2} \end{bmatrix} \\
&= \begin{bmatrix} z_1 z_2 - \zeta_1\overline{\zeta_2} & z_1\zeta_2 + \zeta_1\overline{z_2} \\ -\left(\overline{z_1}\cdot\overline{\zeta_2} + \overline{\zeta_1} z_2\right) & \overline{z_1}\cdot\overline{z_2} - \overline{\zeta_1}\zeta_2 \end{bmatrix} \\
&= \begin{bmatrix} z_1 z_2 - \zeta_1\overline{\zeta_2} & z_1\zeta_2 + \zeta_1\overline{z_2} \\ -\left(\overline{z_1\zeta_2 + \zeta_1\overline{z_2}}\right) & \overline{z_1 z_2 - \zeta_1\overline{\zeta_2}} \end{bmatrix}.
\end{aligned}$$

So, to finish showing the matrix product faithfully represents the quaternion product, we need only compute $z_1 z_2 - \zeta_1\overline{\zeta_2}$ and $z_1\zeta_2 + \zeta_1\overline{z_2}$. Note that

$$\begin{aligned}
&z_1 z_2 - \zeta_1\overline{\zeta_2} \\
&= (a_1 + b_1 i)(a_2 + b_2 i) - (c_1 + d_1 i)(c_2 - d_2 i) \\
&= a_1 a_2 - b_1 b_2 + i(a_1 b_2 + a_2 b_1) - (c_1 c_2 + d_1 d_2 + i(c_2 d_1 - c_1 d_2)) \\
&= (a_1 a_2 - b_1 b_2 - c_1 c_2 - d_1 d_2) + i(a_1 b_2 + a_2 b_1 - c_2 d_1 + c_1 d_2)
\end{aligned}$$

and that

$$\begin{aligned}
&z_1\zeta_2 + \zeta_1\overline{z_2} \\
&= (a_1 + b_1 i)(c_2 + d_2 i) + (c_1 + d_1 i)(a_2 - b_2 i) \\
&= a_1 c_2 - b_1 d_2 + i(a_1 d_2 + b_1 c_2) + a_2 c_1 + b_2 d_1 + i(a_2 d_1 - b_2 c_1) \\
&= (a_1 c_2 - b_1 d_2 + a_2 c_1 + b_2 d_1) + i(a_1 d_2 + b_1 c_2 + a_2 d_1 - b_2 c_1).
\end{aligned}$$

For easy reference, we recall the original definition of $\cdot$ for $\mathbb{H}$:

$$\begin{aligned}&(a_1 + b_1\mathbf{i} + c_1\mathbf{j} + d_1\mathbf{k})(a_2 + b_2\mathbf{i} + c_2\mathbf{j} + d_2\mathbf{k})\\ &= (a_1a_2 - b_1b_2 - c_1c_2 - d_1d_2)\\ &\quad + (a_1b_2 + a_2b_1 + c_1d_2 - c_2d_1)\mathbf{i}\\ &\quad + (a_1c_2 + a_2c_1 - b_1d_2 + b_2d_1)\mathbf{j}\\ &\quad + (a_1d_2 + a_2d_1 + b_1c_2 - b_2c_1)\mathbf{k}.\end{aligned}$$

Since the corresponding components agree throughout, we have verified that this matrix representation faithfully represents the quaternions as matrices.

Proposition 9.1.18. The quaternions $(\mathbb{H}, +, \cdot)$ constitute a non-commutative division ring.

Proof. First, note that $(\mathbb{H}, +)$ forms an Abelian group with identity $0 + 0\mathbf{i} + 0\mathbf{j} + 0\mathbf{k}$ since $(\mathbb{H}, +)$ can be seen as an isomorphic copy of the direct product of four copies of $(\mathbb{R}, +)$. By using the matrix representations for $\mathbb{H}$, we see that $\cdot$ is associative and satisfies the distributive property over addition by Proposition 9.1.7. Hence, $(\mathbb{H}, +, \cdot)$ is a ring.

We can immediately see that the ring operation is not commutative since

$$\mathbf{ij} = \mathbf{k} \neq -\mathbf{k} = \mathbf{ji}.$$

We can also see that the multiplicative identity of $\mathbb{H}$ is $\mathbf{1} = 1 + 0\mathbf{i} + 0\mathbf{j} + 0\mathbf{k}$. Then, given some $a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k} \neq 0$, note that

$$(a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}) \cdot \frac{a - b\mathbf{i} - c\mathbf{j} - d\mathbf{k}}{a^2 + b^2 + c^2 + d^2} = \frac{a - b\mathbf{i} - c\mathbf{j} - d\mathbf{k}}{a^2 + b^2 + c^2 + d^2} \cdot (a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}) = 1.$$

Therefore, $\mathbb{H}$ is a non-commutative division ring. □

The analogous substructure in the context of rings to the notion of subgroups in group theory is the following:

Definition 9.1.19. Let $(R, +, \cdot)$ be a ring. Then $S \subseteq R$ is called a *subring*[2] of R if S is a subgroup of $(R, +)$ and S is closed under the ring operation; that is, for any $x, y \in S$, $xy \in S$. In the case that S is a subring of R and S is a field, we will refer to S as a *subfield* of R.

It is important to note that subrings of rings with unity that don't contain the parent ring's multiplicative identity can still have some other multiplicative identity.

Example 9.1.20. Consider $(\mathbb{Z}_6, \oplus, *)$ and $R = \{0, 2, 4\}$. Note that $(R, \oplus)$ is a subgroup of $(\mathbb{Z}_6, \oplus)$ and consider the Cayley table for the restricted product:

$*$	0	2	4
0	0	0	0
2	0	4	2
4	0	2	4

[2]Note that, as mentioned in the corresponding Wikipedia article, some authors require subrings of rings with unity to contain the multiplicative identity. We presently will not elaborate on the pros and cons of this additional requirement.

Hence, R is a subring of $\mathbb{Z}_6$. Moreover, we see that 1 is the multiplicative identity of $\mathbb{Z}_6$, but 4 serves as a multiplicative identity for R. ⊣

Some subrings of rings with unity may fail to have a multiplicative identity at all. They may, moreover, also consist of zero divisors.

Example 9.1.21. Note that $R = \{0, 2, 4, 6\}$ is a subgroup of $\mathbb{Z}_8$ and consider the usual multiplication $*$ on $\mathbb{Z}_8$. Consider the Caley table for this operation restricted to R:

$*$	0	2	4	6
0	0	0	0	0
2	0	4	0	4
4	0	0	0	0
6	0	4	0	4

It follows that $(R, \oplus, *)$ is a commutative ring in which every element is a zero divisor and that R has no multiplicative identity. ⊣

The natural extension of Proposition 4.2.4 for subrings is

Proposition 9.1.22. Suppose $(R, +, \cdot)$ is a ring and that $S \subseteq R$. Then S is a subring of R if and only if the following three conditions hold:

- $S \neq \emptyset$.
- For every $x, y \in S$, $x - y \in S$.
- For every $x, y \in S$, $xy \in S$.

Note that $\mathbb{Z}$ is a subring of $\mathbb{Q}$, $\mathbb{Q}$ is a subring of $\mathbb{R}$, and $\mathbb{R}$ is a subring of $\mathbb{C}$. Since $\mathbb{Z}$ is not a field, we see that not every subring of a field is a field. Note also that $2\mathbb{Z}$ is an example of a proper subring of $\mathbb{Z}$ which does not contain a multiplicative identity.

Example 9.1.23. Not every subgroup of the underlying group constituting a ring is a subring. Let $G = \{n + in : n \in \mathbb{Z}\}$ and note that G is a subgroup of $(\mathbb{C}, +)$. Note also that $1 + i \in G$ and that $(1 + i)(1 + i) = 2i \notin G$. Since G is not closed under the ring operation, G is not a subring of $(\mathbb{C}, +, \cdot)$. ⊣

Exercise 9.1.7. Let R be a ring and $\mathcal{S}$ be a set of subrings of R. Show that $\bigcap \mathcal{S}$ is a subring of R.

Exercise 9.1.8. Provide a counterexample to the following statement: Let R be a ring and suppose that S_1 and S_2 are subrings of R. Then $S_1 \cup S_2$ is a subring of R.

Integral domains can also be defined in terms of a cancellation-style property.

Proposition 9.1.24. A commutative ring R with unity is an integral domain if and only if R satisfies the following property: for any $a, x, y \in R$ with $a \neq 0_R$, if $ax = ay$, then $x = y$.

Proof. Suppose R is a commutative ring with unity and that R is an integral domain. Then suppose that $a, x, y \in R$ are such that $a \neq 0_R$ and $ax = ay$. Then $a(x - y) = ax - ay = 0_R$. Since $a \neq 0_R$ and R is an integral domain, it must be the case that $x - y = 0_R$. That is, $x = y$.

For the other direction, suppose R is a commutative ring with unity, but that R is not an integral domain. Then there are nonzero elements $a, b \in R$ for which $ab = 0_R$. Since

$$a \cdot b = 0_R = a \cdot 0_R,$$

$a \neq 0_R$, and $b \neq 0_R$, we see that the cancellation property in the statement of the proposition fails for R. □

As an almost immediate consequence:

Theorem 9.1.25. Every finite integral domain is a field.

Proof. Suppose $(R, +, \cdot)$ is a finite integral domain and let $x \in R^*$. Consider $\mu : R^* \to R^*$ defined by $\mu(y) = xy$. Note that μ is defined since R has no zero divisors.

We will show now that μ is injective. So suppose $y_1, y_2 \in R^*$ are such that $\mu(y_1) = \mu(y_2)$. Then $xy_1 = xy_2$ and, by Proposition 9.1.24, $y_1 = y_2$.

Now, since R^* is a finite set and $\mu : R^* \to R^*$ is an injection, μ must also be a surjection. Hence, there must be some $y \in R^*$ for which $\mu(y) = 1_R$ since integral domains are assumed to have unity, and the multiplicative inverse is assumed to be nonzero. Integral domains are also assumed to be commutative, so we have that $xy = yx = 1_R$. That is, x has a multiplicative inverse.

Conclusively, R is a commutative ring with unity which is also a division ring. That is, R is a field. □

We comment here how the typical matrix multiplication imposes some important restrictions on the corresponding set of possible entries, if we restrict our attention to those matrices with a nonzero determinant.

Definition 9.1.26. For a ring R, define $\det : \mathsf{M}(2, R) \to R$ by the rule

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc.$$

Then let

$$\mathrm{NZD}(2, R) := \{A \in \mathsf{M}(2, R) : \det(A) \neq 0_R\}.$$

Though our definition of $\mathrm{NZD}(2, R)$ seems to correspond to what one would expect to be called $\mathrm{GL}(2, R)$, we have chosen to use this notation to emphasize the nonzero determinant condition. In particular, $\mathrm{GL}(2, R)$ would be defined as the set of invertible elements of $\mathsf{M}(2, R)$, which is a group by Exercise 9.1.5, and which, in very general contexts, might not correspond exactly to matrices with just a nonzero determinant.

Lemma 9.1.27. Let $(R, +, \cdot)$ be a ring with $0 = 0_R$ and consider $\cdot$ on $\mathsf{M}(2, R)$ as defined in Definition 9.1.6.

(a) If R has unity $1 = 1_R$, then $\mathrm{I}_2 := \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ has the property that $A \cdot \mathrm{I}_2 = \mathrm{I}_2 \cdot A = A$ for any $A \in \mathsf{M}(2, R)$.

(b) If R is a non-commutative ring with unity, then there exists $A \in \mathrm{NZD}(2, R)$ such that, for every $B \in \mathsf{M}(2, R)$, $AB \neq \mathrm{I}_2$.

(c) If R is a ring with unity and has at least one zero divisor, then there is some $A \in \mathrm{NZD}(2, R)$ such that, for every $B \in \mathsf{M}(2, R)$, $AB \neq \mathrm{I}_2$ or $BA \neq \mathrm{I}_2$.

(d) Suppose R has at least two elements and fix $y \in R^*$. Then let

$$\mathsf{D}_y = \left\{ \begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix} : x \in R^* \right\}.$$

If there is some $A \in \mathsf{M}(2, R)$ such that $AX = XA = X$ for every $X \in \mathsf{D}_y$, then R has unity.

Proof. (a) Suppose R has unity. Then note that

$$\begin{aligned} \begin{bmatrix} w & x \\ y & z \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} &= \begin{bmatrix} w \cdot 1 + x \cdot 0 & w \cdot 0 + x \cdot 1 \\ y \cdot 1 + z \cdot 0 & y \cdot 0 + z \cdot 1 \end{bmatrix} \\ &= \begin{bmatrix} w & x \\ y & z \end{bmatrix} \end{aligned}$$

and that

$$\begin{aligned} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} w & x \\ y & z \end{bmatrix} &= \begin{bmatrix} 1 \cdot w + 0 \cdot y & 1 \cdot x + 0 \cdot z \\ 0 \cdot w + 1 \cdot y & 0 \cdot x + 1 \cdot z \end{bmatrix} \\ &= \begin{bmatrix} w & x \\ y & z \end{bmatrix} \end{aligned}$$

(b) Suppose R is a non-commutative ring with unity and suppose $a, b \in R$ are such that $ab \neq ba$. Consider

$$A := \begin{bmatrix} a & b \\ a & b \end{bmatrix} \in \mathsf{M}(2, R)$$

and note that, since $ab - ba \neq 0_R$, $A \in \mathrm{NZD}(2, R)$. Then observe that, for

$$B = \begin{bmatrix} w & x \\ y & z \end{bmatrix} \in \mathsf{M}(2, R),$$

$$AB = \begin{bmatrix} a & b \\ a & b \end{bmatrix} \begin{bmatrix} w & x \\ y & z \end{bmatrix} = \begin{bmatrix} aw + by & ax + bz \\ aw + by & ax + bz \end{bmatrix}.$$

Since the first column in the resulting product has equal entries, $AB \neq \mathrm{I}_2$.

(c) Suppose R is a ring with unity and that $x \in R$ is a zero divisor. Then consider the fact that

$$A := \begin{bmatrix} x & 0 \\ 0 & 1 \end{bmatrix} \in \mathrm{NZD}(2, R).$$

Let

$$B = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathsf{M}(2, R)$$

be arbitrary. We proceed by cases.

First, suppose x is a left zero divisor. Note that

$$BA = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} ax & b \\ cx & d \end{bmatrix}.$$

Since x is a left zero divisor, x is not left-invertible by Proposition 9.1.12. Hence, $ax \neq 1$ and so $BA \notin \mathrm{I}_2$.

Now, suppose x is a right zero divisor. Then note that

$$AB = \begin{bmatrix} x & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} xa & xb \\ c & d \end{bmatrix}.$$

As above, $xa \neq 1$ by Proposition 9.1.12. Thus, $AB \neq \mathrm{I}_2$.

(d) Let

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

be such that $AX = XA = X$ for every $X \in \mathsf{D}_y$. We claim that a serves as an identity element for R.

Note that, for any $x \in R^*$,

$$\begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix} = \begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} xa & xb \\ yc & yd \end{bmatrix}$$

and

$$\begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix} = \begin{bmatrix} ax & by \\ cx & dy \end{bmatrix}.$$

In particular, $x = ax = xa$. Note that $a \neq 0$ since $x = xa$ and $x \cdot 0 = 0 \neq x$.

Since a is fixed, $a \neq 0$, and $x \in R^*$ was chosen to be arbitrary, we see that a serves as the multiplicative identity for R. That is, R is a ring with unity. □

Theorem 9.1.28. Let $(R, +, \cdot)$ be a ring with no zero divisors. Then $(\mathrm{NZD}(2, R), \cdot)$, where $\cdot$ is as defined in Definition 9.1.6, is a group if and only if R is a field.

Proof. ($\Leftarrow$) First, suppose R is a field with $0 = 0_R$ and $1 = 1_R$. By Proposition 9.1.7, $\cdot$ is associative and, by Lemma 9.1.27, I_2, as defined in the lemma, serves as the identity element. Hence, to establish that $(\mathrm{NZD}(2, R), \cdot)$ is a group, we need only show that $\mathrm{NZD}(2, R)$ has inverses. So, let

$$A = \begin{bmatrix} w & x \\ y & z \end{bmatrix} \in \mathrm{NZD}(2, R)$$

be arbitrary and note that $wz - xy \in R^*$ implies that $(wz - xy)^{-1} \in R^*$. Let $\alpha = (wz - xy)^{-1}$ and define

$$B = \begin{bmatrix} \alpha z & -\alpha x \\ -\alpha y & \alpha w \end{bmatrix} \in \mathsf{M}(2, R).$$

We first must show that $B \in \mathrm{NZD}(2, R)$. So consider the fact that

$$\begin{aligned} \alpha z \alpha w - \alpha x \alpha y &= \alpha^2 wz - \alpha^2 xy \\ &= \alpha^2 (wz - xy) \\ &= (wz - xy)^{-1} (wz - xy)^{-1} (wz - xy) \\ &= (wz - xy)^{-1} \neq 0. \end{aligned}$$

Hence, $B \in \mathrm{NZD}(2, R)$.

Now we show that B is the inverse element of A. Note that

$$\begin{aligned} AB &= \begin{bmatrix} w & x \\ y & z \end{bmatrix} \begin{bmatrix} \alpha z & -\alpha x \\ -\alpha y & \alpha w \end{bmatrix} \\ &= \begin{bmatrix} \alpha wz - \alpha xy & \alpha wx - \alpha wx \\ \alpha yz - \alpha yz & \alpha wz - \alpha xy \end{bmatrix} \\ &= \begin{bmatrix} \alpha(wz - xy) & 0 \\ 0 & \alpha(wz - xy) \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{aligned}$$

and that

$$\begin{aligned} BA &= \begin{bmatrix} \alpha z & -\alpha x \\ -\alpha y & \alpha w \end{bmatrix} \begin{bmatrix} w & x \\ y & z \end{bmatrix} \\ &= \begin{bmatrix} \alpha wz - \alpha xy & \alpha xz - \alpha xz \\ \alpha wy - \alpha wy & \alpha wz - \alpha xy \end{bmatrix} \\ &= \begin{bmatrix} \alpha(wz - xy) & 0 \\ 0 & \alpha(wz - xy) \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \end{aligned}$$

Hence, B is the inverse of A. Since A was arbitrary, we see that every element of $\mathrm{NZD}(2, R)$ has an inverse. Thus, $(\mathrm{NZD}(2, R), \cdot)$ is a group.

($\Rightarrow$) For the reverse direction, suppose $(\mathrm{NZD}(2, R), \cdot)$ is a group. Since a group must necessarily be nonempty, there is some

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{NZD}(2, R).$$

Note that, by the definition of $\mathrm{NZD}(2, R)$, $ad - bc \in R^*$. Hence, R has at least two elements since $0 \in R$ and $ad - bc \in R$ with $ad - bc \neq 0$. We can thus fix some $y \in R^*$ and define D_y as in Lemma 9.1.27. Since R is assumed to have no zero divisors, $\mathsf{D}_y \subseteq \mathrm{NZD}(2, R)$. Indeed, if $x \in R^*$, note that $xy \neq 0$ and so

$$\begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix} \in \mathrm{NZD}(2, R).$$

Since $\mathrm{NZD}(2,R)$ is assumed to be a group, it has an identity element. In particular, the hypothesis for Lemma 9.1.27(d) is satisfied and thus R must have unity. Say $1 = 1_R$.

Now, since the matrix I_2 as defined in Lemma 9.1.27 satisfies the properties of an identity in $(\mathrm{NZD}(2,R),\cdot)$ and identity elements in groups are unique, the identity element of $\mathrm{NZD}(2,R)$ must be I_2.

The fact that R must be commutative follows from Lemma 9.1.27(b). So the last thing to show is that R is a division ring. Let $x \in R^*$ be arbitrary and note that, since $(\mathrm{NZD}(2,R),\cdot)$ is a group and

$$\begin{bmatrix} x & 0 \\ 0 & 1 \end{bmatrix} \in \mathrm{NZD}(2,R),$$

we can let

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} x & 0 \\ 0 & 1 \end{bmatrix}^{-1}.$$

In particular,

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} x & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} xa & xb \\ c & d \end{bmatrix}$$

and

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} x & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} ax & b \\ cx & d \end{bmatrix}.$$

It follows that $1 = xa = ax$, and so x has a multiplicative inverse. □

Question 9.1. Can the assumption that R has no zero divisors be dropped from the hypothesis of Theorem 9.1.28, or is there an example of a (preferably commutative) ring (without unity by Lemma 9.1.27(c)) with at least one zero divisor for which $(\mathrm{NZD}(2,R),\cdot)$ forms a group?

One may think that appealing to properties of the determinant may be helpful, but we show here an example of a commutative ring R without unity and with zero divisors for which

$$\{\det(A) : A \in \mathrm{NZD}(2,R)\}$$

is a group, but $\mathrm{NZD}(2,R)$ is not. We'll start by showing that the determinant map always preserves multiplication for commutative rings.

Proposition 9.1.29. Suppose that R is a commutative ring. Then $\det : \mathsf{M}(2,R) \to R$ has the property that

$$\det(AB) = \det(A)\cdot\det(B).$$

Proof. First note that

$$\begin{aligned}
\det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} w & x \\ y & z \end{bmatrix}\right) &= \det\begin{bmatrix} aw+by & ax+bz \\ cw+dy & cx+dz \end{bmatrix} \\
&= (aw+by)(cx+dz) - (ax+bz)(cw+dy) \\
&= awcx + awdz + bycx + bydz - axcw - axdy - bzcw - bzdy \\
&= adwz + bcxy - adxy - bcwz
\end{aligned}$$

and that

$$\begin{aligned}\det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right)\det\left(\begin{bmatrix} w & x \\ y & z \end{bmatrix}\right) &= (ad-bc)(wz-xy)\\ &= adwz - adxy - bcwz + bcxy\\ &= adwz + bcxy - adxy - bcwz\\ &= \det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} w & x \\ y & z \end{bmatrix}\right).\end{aligned}$$

Thus, the proposition obtains. □

Example 9.1.30. Let $R = \{0, 2, 4, 6, 8, 10\}$ be viewed as a subring of $\mathbb{Z}_{12}$. Then R is a commutative ring without identity and with zero divisors. In particular, note that $2 * 6 = 0$. To see that R fails to have a multiplicative identity, consider the Cayley table for $*$:

$*$	0	2	4	6	8	10
0	0	0	0	0	0	0
2	0	4	8	0	4	8
4	0	8	4	0	8	4
6	0	0	0	0	0	0
8	0	4	8	0	4	8
10	0	8	4	0	8	4

Evidently, no element of R can serve as a multiplicative identity.

Now we show that

$$\{\det(A) : A \in \mathrm{NZD}(2, R)\} = \{4, 8\}$$

and that $(\{4, 8\}, *)$ is a group where $*$ is multiplication mod 12. From the Cayley table above, for $a, b, c, d \in R$, we see that $ab, -cd \in \{0, 4, 8\}$. Indeed, note that $-4 \,\%\, 12 = 8$ and $-8 \,\%\, 12 = 4$. Since $\{0, 4, 8\}$ is a subgroup of R, we see that $ad - bc \in \{0, 4, 8\}$. Hence,

$$\{\det(A) : A \in \mathrm{NZD}(2, R)\} \subseteq \{4, 8\}.$$

For equality, note that

$$\det\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} = 4$$

and that

$$\det\begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix} = 8.$$

Since $4 \cdot 4 \equiv 4 \pmod{12}$, $4 \cdot 8 \equiv 8 \pmod{12}$, and $8 \cdot 8 \equiv 4 \pmod{12}$, $(\{4, 8\}, *)$ is a group.

However, $(\mathrm{NZD}(2, R), \cdot)$ is not a group. Indeed, we will show that there is no multiplicative identity. Consider

$$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \in \mathrm{NZD}(2, R)$$

and

$$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} 2a & 2b \\ 2c & 2d \end{bmatrix}.$$

Note that $2a \in \{0, 4, 8\}$, so there is no $a \in R$ for which $2a = 2$. That is, $\mathrm{NZD}(2, R)$ has no identity element. ⊣

9.2 Ring Homomorphisms

Homomorphisms of groups preserve the group operations. The natural extension to rings then is to preserve both the group and the ring operations.

Definition 9.2.1. Suppose $(R, +, \cdot)$ and $(S, +, \cdot)$ are rings. A map $\varphi : R \to S$ is a *homomorphism* (of rings) if, for every $x, y \in R$, $\varphi(x + y) = \varphi(x) + \varphi(y)$ and $\varphi(xy) = \varphi(x)\varphi(y)$.

Note that every homomorphism of rings is also a homomorphism of the underlying groups. Hence, any properties enjoyed by homomorphisms of groups are inherited to homomorphisms of rings.

Definition 9.2.2. Suppose $(R, +, \cdot)$ and $(S, +, \cdot)$ are rings. A bijection $\varphi : R \to S$ is an *isomorphism* (of rings) if φ is a homomorphism of rings and its inverse $\varphi^{-1} : S \to R$ is a homomorphism of rings. We will use $R \cong S$ to denote that there exists an isomorphism (of rings) $\varphi : R \to S$.

Exercise 9.2.1. Show that, just as with groups, any bijective homomorphism of rings is necessarily an isomorphism of rings.

When context is understood, we simplify use the term *homomorphism*. For example, if we have a mapping between two objects that are identified as rings, then the term *homomorphism* will mean a homomorphism of rings. If we have a mapping between two objects that are only identified as groups and not as rings, then the term *homomorphism* will mean a homomorphism of groups. When the objects of consideration are rings and we wish to emphasize that a map is only necessarily a homomorphism of the underlying groups and not necessarily of the ring structure, we will specify explicitly that it is a homomorphism *of groups*. The same convention applies for *isomorphisms*.

Exercise 9.2.2. Consider $\varphi : \mathbb{Z} \to \mathbb{Z}$ defined by $\varphi(n) = 2n$. Is φ a homomorphism of rings? Why or why not?

Exercise 9.2.3. Suppose $\varphi : R \to S$ is a homomorphism of rings.

(a) Show that $\varphi[R]$ is a subring of S.

(b) If R is a commutative ring, show that $\varphi[R]$ is a commutative subring of S.

(c) If R is a ring with unity and φ is a surjection, show that, for every $y \in S$,

$$y = y \cdot \varphi(1_R) = \varphi(1_R) \cdot y.$$

(d) If R is a ring with unity and $\varphi(1_R) = 0_S$, show that $\varphi[R] = \{0_S\}$.

(e) If R is a ring with unity and $\varphi[R] \neq \{0_S\}$, show that $\varphi[R]$ is a ring with unity.

(f) If R is a division ring and $\varphi[R] \neq \{0_S\}$, show that $\varphi[R]$ is a division ring.

Exercise 9.2.4. Let $\varphi : R \to S$ be an isomorphism of the rings R and S.

(a) Show that, if R has zero divisors, then S has zero divisors.

(b) Show that, if R is a commutative ring with unity, then S is a commutative ring with unity.

(c) Show that, if R is an integral domain, then S is an integral domain.

(d) Show that, if R is a division ring, then S is a division ring.

(e) Show that, if R is a field, then S is a field.

Exercise 9.2.5. Prove or disprove the following statement: If $\varphi : R \to S$ is a homomorphism of rings R and S, and R has zero divisors, then S has zero divisors.

Remark. For a homomorphism $\varphi : R \to S$ of rings, the *kernel* is as it was defined for groups; in particular,

$$\ker(\varphi) = \{x \in R : \varphi(x) = 0_S\}.$$

Example 9.2.3. Let X be a nonempty set and let $\mathbb{R}^X$ be the set of all functions $X \to \mathbb{R}$. For $f, g \in \mathbb{R}^X$, let $(f + g) : X \to \mathbb{R}$ be defined by the rule $(f + g)(x) = f(x) + g(x)$ and let $(f \cdot g) : X \to \mathbb{R}$ be defined by the rule $(f \cdot g)(x) = f(x) \cdot g(x)$. These operations are known as the *pointwise sum* and *pointwise product*, respectively. Let $\mathbf{0}$ represent the constant zero function and $\mathbf{1}$ represent the constant one function. Then routine checks can confirm that $(\mathbb{R}^X, +, \cdot)$ is a commutative ring with unity where the additive identity is $\mathbf{0}$ and the multiplicative identity is $\mathbf{1}$.

Fix some $a \in X$ and define $E_a : \mathbb{R}^X \to \mathbb{R}$ by the rule $E_a(f) = f(a)$. Then E_a is a surjective homomorphism of rings and $\ker(E_a) = \{f \in \mathbb{R}^X : f(a) = 0\}$. ⊣

9.3 Ideals

In the context of groups, normal subgroups play an important role in quotient groups and not every subgroup of a group is a normal subgroup. In particular, every normal subgroup arises as the kernel of some homomorphism of groups. In the context of rings, we have an analogous class of subrings.

Definition 9.3.1. Suppose $(R, +, \cdot)$ is a ring and that $\mathcal{I}$ is a subgroup of $(R, +)$. Then $\mathcal{I}$ is a *left-ideal* of R if, for every $x \in R$ and $a \in \mathcal{I}$, $xa \in \mathcal{I}$; $\mathcal{I}$ is a *right-ideal* of R if, for every $x \in R$ and $a \in \mathcal{I}$, $ax \in \mathcal{I}$. If $\mathcal{I}$ is both a left-ideal and a right-ideal, then we say that $\mathcal{I}$ is a *two-sided ideal*. In the case that R is a commutative ring, then we refer to any left-ideal as, simply, an *ideal*.

Definition 9.3.2. In any ring R, both $\{0_R\}$ and R are two-sided ideals of R and we refer to both as *trivial* ideals. Any ideal which is a proper subset of R is a *proper* ideal.

Note that every type of ideal defined is necessarily a subring of the parent ring. However, like in the context of subgroups and normal subgroups, not every subring is an ideal.

Example 9.3.3. Note that $\mathbb{Z}$ is a subring of $\mathbb{R}$ which is not an ideal since $\frac{1}{2} \cdot 1 = \frac{1}{2} \notin \mathbb{Z}$. $\dashv$

Exercise 9.3.1. Here are two basic observations related to ideals.

(a) Show that any left- or right-ideal of a ring with unity which contains the multiplicative identity must be the whole ring.

(b) Show that no division ring has a proper nontrivial left- or right-ideal.

The simplest ideals are known as *principal ideals*.

Definition 9.3.4. For a ring R and an element $a \in R$, we define the *principal left-ideal generated by* a to be

$$Ra = \{ra : r \in R\}$$

and the *principal right-ideal generated by* a to be

$$aR = \{ar : r \in R\}.$$

In general, we will say an ideal $\mathcal{I}$ of R is a *principal ideal* if there is some $a \in R$ for which $\mathcal{I} = Ra$ or $\mathcal{I} = aR$. When R is a commutative ring, we will use the now overloaded[3] notation

$$\langle a \rangle = aR = Ra.$$

Proposition 9.3.5. Given $a \in R$, the principal left-ideal generated by a is a left-ideal, and the principal right-ideal generated by a is a right-ideal.

Proof. We first show that both aR and Ra are subgroups of $(R, +)$. Note that $0_R = a \cdot 0_R \in aR$ and $0_R = 0_R \cdot a \in Ra$ so both aR and Ra are nonempty. For $x, y \in R$, note that $ax - ay = a(x - y) \in aR$ and $xa - ya = (x - y)a \in Ra$. By Proposition 4.2.4, aR and Ra are both subgroups of $(R, +)$.

To see that Ra is a left-ideal, let $xa \in Ra$ and $y \in R$ be arbitrary. Note that $y(xa) = (yx)a \in Ra$. So Ra is a left-ideal.

For aR, let $ax \in aR$ and $y \in R$ be arbitrary. Then $(ax)y = a(xy) \in aR$. So aR is a right-ideal. $\square$

Note that, in the ring $\mathbb{Z}$, $n\mathbb{Z} = \{nk : k \in \mathbb{Z}\}$ is a principal ideal for any $n \in \mathbb{Z}$. In fact, since every ideal is necessarily a subgroup and, by Theorem 4.4.6, every subgroup of $\mathbb{Z}$ is of the form $n\mathbb{Z}$, we see that every ideal of $\mathbb{Z}$ must be of the form $n\mathbb{Z}$, $n \in \mathbb{Z}$.

In general, not every ideal is a principal ideal.

[3] Compare to the notation used for cyclic subgroups in Definition 4.4.1.

Example 9.3.6. Let R be the set of all binary sequences, which is, all functions from $\mathbb{N}$ to $\mathbb{Z}_2$. If $+$ denotes coordinate-wise addition and $\cdot$ denotes coordinate-wise multiplication, then $(R, +, \cdot)$ is a commutative ring with unity where the additive identity is the constant 0 function and the multiplicative identity is the constant 1 function. Then let

$$\mathcal{I} = \{f \in R : (\exists m \in \mathbb{N})(\forall n > m)\ \ [f(n) = 0]\}.$$

In words, $\mathcal{I}$ consists of all binary sequences that are eventually zero on some tail of $\mathbb{N}$. We prove below that $\mathcal{I}$ is a nonprincipal ideal of R.

First, we show that $\mathcal{I}$ is a subgroup of R. Note first that the constant zero function is a member of $\mathcal{I}$, so $\mathcal{I} \neq \emptyset$. Then let $f, g \in \mathcal{I}$ and let $m_f, m_g \in \mathbb{N}$ be such that, for all $n > m_f$, $f(n) = 0$ and all $n > m_g$, $g(n) = 0$. Let $m = \max\{m_f, m_g\}$ and note that, for any $n > m$, $f(n) = g(n) = 0$. Hence, for any $n > m$, $(f - g)(n) = f(n) - g(n) = 0$. So $f - g \in \mathcal{I}$.

Now we show that $\mathcal{I}$ absorbs multiplication. So let $f \in R$ and $g \in \mathcal{I}$. Let $m \in \mathbb{N}$ be such that, for every $n > m$, $g(n) = 0$. Note that, for every $n > m$, $(fg)(n) = f(n)g(n) = 0$. Hence, $fg \in \mathcal{I}$.

The last thing to show is that $\mathcal{I}$ is not principal. So consider any $g \in R$ such that $\langle g \rangle = gR \subseteq \mathcal{I}$. Since R has unity, note that $g \in gR \subseteq \mathcal{I}$ means that $g \in \mathcal{I}$. So let $m \in \mathbb{N}$ be such that, for every $n > m$, $g(n) = 0$. Define $f : \mathbb{N} \to \mathbb{Z}_2$ by

$$f(n) = \begin{cases} 0, & n \neq m+1 \\ 1, & n = m+1 \end{cases}$$

Note that $f \in \mathcal{I}$. Also note that, for any $h \in R$,

$$(gh)(m+1) = g(m+1)h(m+1) = 0 \neq 1 = f(m+1).$$

Since $h \in R$ was arbitrary, $f \notin gR$ and so $gR \neq \mathcal{I}$. Since g was arbitrary, this means that $\mathcal{I}$ is not a principal ideal. ⊣

Exercise 9.3.2. Let R be a ring and $\mathscr{F}$ be a set of left-ideals (respectively, right-ideals) of R. Show that $\bigcap \mathscr{F}$ is a left-ideal (respectively, right-ideal) of R.

Exercise 9.3.3. Provide a counterexample to the following statement: Let R be a ring and suppose that $\mathcal{I}_1$ and $\mathcal{I}_2$ are left-ideals of R. Then $\mathcal{I}_1 \cup \mathcal{I}_2$ is a left-ideal of R.

Proposition 9.3.7. If $\varphi : R \to S$ is a homomorphism of rings, then $\ker(\varphi)$ is a two-sided ideal of R.

Proof. We already know that $\ker(\varphi)$ is a subgroup of R by Exercise 4.3.4. So we need only show it satisfies the additional conditions for being a two-sided ideal. So suppose $x \in R$ and $h \in \ker(\varphi)$. Note that, by Proposition 9.1.2,

$$\varphi(xh) = \varphi(x)\varphi(h) = \varphi(x) \cdot 0_S = 0_S.$$

That is, $xh \in \ker(\varphi)$. In a similar way,

$$\varphi(hx) = \varphi(h)\varphi(x) = 0_S \cdot \varphi(x) = 0_S.$$

Thus, $hx \in \ker(\varphi)$. □

Just like how normal subgroups arise exactly as kernels of group homomorphisms, every two-sided ideal arises as the kernel of a ring homomorphism. Indeed, quotient groups over ideals inherit a ring structure from the originating ring.

Definition 9.3.8. Let $(R, +, \cdot)$ be a ring and $\mathcal{I}$ be a two-sided ideal of R. Define $\cdot$ on the quotient group $R/\mathcal{I}$ by the rule

$$(x + \mathcal{I}) \cdot (y + \mathcal{I}) = xy + \mathcal{I}.$$

As it will be shown below, $(R/\mathcal{I}, +, \cdot)$ constitutes a ring which is called the *quotient ring* or *factor ring*.

Proposition 9.3.9. For a ring $(R, +, \cdot)$ and a two-sided ideal $\mathcal{I}$ of R, $(R/\mathcal{I}, +, \cdot)$ is a ring. Moreover, the map $\varphi : R \to R/\mathcal{I}$ defined by $\varphi(x) = x + \mathcal{I}$ is a surjective homomorphism of rings with the property that $\mathcal{I} = \ker(\varphi)$.

Proof. Since $(R, +)$ is an Abelian group, every subgroup of R is normal, and so $(R/\mathcal{I}, +)$ is a group.

We must show that $\cdot$ is well-defined. So suppose $x_1, x_2, y_1, y_2 \in R$ are such that $x_1 + \mathcal{I} = x_2 + \mathcal{I}$ and $y_1 + \mathcal{I} = y_2 + \mathcal{I}$. We claim that

$$x_1y_1 + \mathcal{I} = x_2y_2 + \mathcal{I}.$$

Since $x_1 + \mathcal{I} = x_2 + \mathcal{I}$ and $y_1 + \mathcal{I} = y_2 + \mathcal{I}$, we can let $a_x, a_y \in \mathcal{I}$ be such that $x_1 = x_2 + a_x$ and $y_1 = y_2 + a_y$. Note then that

$$x_1y_1 = (x_2 + a_x)(y_2 + a_y) = x_2y_2 + x_2a_y + a_xy_2 + a_xa_y.$$

Since $a_x, a_y \in \mathcal{I}$ and $\mathcal{I}$ is a two-sided ideal of R, $x_2a_y, a_xy_2, a_xa_y \in \mathcal{I}$. Hence, $x_1y_1 - x_2y_2 \in \mathcal{I}$, which means that

$$x_1y_1 + \mathcal{I} = x_2y_2 + \mathcal{I}.$$

To finish showing that $(R/\mathcal{I}, +, \cdot)$ is a ring, we need only show that $\cdot$ is associative and satisfies the distributive properties. Immediately,

$$\begin{aligned}((x + \mathcal{I})(y + \mathcal{I})) \cdot (z + \mathcal{I}) &= (xy + \mathcal{I})(z + \mathcal{I}) \\ &= (xy)z + \mathcal{I} \\ &= x(yz) + \mathcal{I} \\ &= (x + \mathcal{I})(yz + \mathcal{I}) \\ &= (x + \mathcal{I})\left((y + \mathcal{I})(z + \mathcal{I})\right),\end{aligned}$$

$$\begin{aligned}(x + \mathcal{I})\left((y + \mathcal{I}) + (z + \mathcal{I})\right) &= (x + \mathcal{I})((y + z) + \mathcal{I}) \\ &= x(y + z) + \mathcal{I} \\ &= (xy + xz) + \mathcal{I} \\ &= (xy + \mathcal{I}) + (xz + \mathcal{I}) \\ &= (x + \mathcal{I})(y + \mathcal{I}) + (x + \mathcal{I})(z + \mathcal{I}),\end{aligned}$$

and

$$\begin{aligned}((x+\mathcal{I})+(y+\mathcal{I}))\,(z+\mathcal{I}) &= ((x+y)+\mathcal{I})(z+\mathcal{I}) \\ &= (x+y)z+\mathcal{I} \\ &= xz+yz+\mathcal{I} \\ &= (xz+\mathcal{I})+(yz+\mathcal{I}) \\ &= (x+\mathcal{I})(z+\mathcal{I})+(y+\mathcal{I})(z+\mathcal{I}).\end{aligned}$$

To finish the proof, we show that φ is a surjective homomorphism of rings with $\ker(\varphi) = \mathcal{I}$. Note that

$$\varphi(x+y) = (x+y)+\mathcal{I} = (x+\mathcal{I})+(y+\mathcal{I}) = \varphi(x)+\varphi(y)$$

and that

$$\varphi(xy) = xy+\mathcal{I} = (x+\mathcal{I})(y+\mathcal{I}) = \varphi(x)\varphi(y).$$

Hence, φ is a homomorphism of rings. It is immediate that φ is surjective.

Since, for any $a \in \mathcal{I}$, $\varphi(a) = a+\mathcal{I} = \mathcal{I}$, we see that $\mathcal{I} \subseteq \ker(\varphi)$. So suppose $x \in \ker(\varphi)$ and note that $x+\mathcal{I} = \varphi(x) = \mathcal{I}$. Hence, $x \in \mathcal{I}$ and so $\ker(\varphi) \subseteq \mathcal{I}$. This finishes the proof. □

Remark. As with quotient groups, if R is a ring and $\mathcal{I}$ is a two-sided ideal of R, then we call $\varphi : R \to R/\mathcal{I}$ defined by $\varphi(x) = x+\mathcal{I}$ the *canonical* homomorphism.

The analogous versions of The First and Second Isomorphism Theorems for rings hold and have nearly identical proofs, which we will leave as exercises for the reader.

Theorem 9.3.10 (The First Isomorphism Theorem)**.** Suppose $\varphi : R \to S$ is a surjective homomorphism (of rings) and that $K = \ker(\varphi)$. Let $\psi : R \to R/K$ be the canonical homomorphism. Then there exist a unique isomorphism (of rings) $\vartheta : R/K \to S$ such that $\varphi = \vartheta \circ \psi$.

Theorem 9.3.11 (The Second Isomorphism Theorem)**.** Suppose that S is a subring of a ring $(R, +, \cdot)$ and that $\mathcal{I}$ is a two-sided ideal of R. Then $S+\mathcal{I}$ is a subring of R, $S \cap \mathcal{I}$ is a two-sided ideal of S, and

$$S/(S \cap \mathcal{I}) \cong (S+\mathcal{I})/\mathcal{I}.$$

9.4 Prime and Maximal Ideals

We will define here special types of ideals in commutative rings with unity that actually characterize when certain properties are enjoyed by the corresponding quotient ring.

Definition 9.4.1. Suppose R is a commutative ring with unity and let $\mathcal{I}$ be an ideal of R. Then we say that $\mathcal{I}$ is *prime* if, for any $xy \in \mathcal{I}$, either $x \in \mathcal{I}$ or $y \in \mathcal{I}$.

Note the similarity of this definition to a property enjoyed by a prime number p; that is, for a given prime p, note that, if $p \mid ab$ for integers a and b, then either $p \mid a$ or $p \mid b$. Note that this criterion doesn't hold for composite numbers since $6 \mid (3 \cdot 4)$ but $6 \nmid 3$ and $6 \nmid 4$.

Theorem 9.4.2. Suppose R is a commutative ring with unity and that $\mathcal{I}$ is a proper ideal of R. Then $\mathcal{I}$ is a prime ideal if and only if $R/\mathcal{I}$ is an integral domain.

Proof. First, note that, by Exercise 9.2.3, $R/\mathcal{I}$ is a commutative ring with unity since R is a commutative ring with unity and $\mathcal{I}$ is a proper ideal. So, in either direction, we need only address the existence of zero divisors.

Suppose $\mathcal{I}$ is a prime ideal of R and consider $x+\mathcal{I}, y+\mathcal{I} \in R/\mathcal{I}$ for which

$$xy+\mathcal{I} = (x+\mathcal{I})(y+\mathcal{I}) = \mathcal{I}.$$

Note that, in particular, $xy \in \mathcal{I}$. Since $\mathcal{I}$ is assumed to be a prime ideal, either $x \in \mathcal{I}$ or $y \in \mathcal{I}$. This means that $R/\mathcal{I}$ has no zero divisors, and is, hence, an integral domain.

Now suppose $R/\mathcal{I}$ is an integral domain and let $x, y \in R$ be such that $xy \in \mathcal{I}$. Note that

$$(x+\mathcal{I})(y+\mathcal{I}) = xy+\mathcal{I} = \mathcal{I}.$$

If $x \in \mathcal{I}$, there is nothing to be shown, so suppose $x \notin \mathcal{I}$. Then $x+\mathcal{I} \neq \mathcal{I}$ and $(x+\mathcal{I})(y+\mathcal{I}) = \mathcal{I}$. Since $R/\mathcal{I}$ is assumed to be an integral domain, it must be the case that $y+\mathcal{I} = \mathcal{I}$. That is, $y \in \mathcal{I}$. Conclusively, $\mathcal{I}$ is a prime ideal. □

As noted above, every ideal of $\mathbb{Z}$ is of the form $n\mathbb{Z}$. Since, for $n \in \mathbb{Z}$ with $n \geq 2$, $\mathbb{Z}/n\mathbb{Z} \cong \mathbb{Z}_n$, we see that an ideal $n\mathbb{Z}$ of $\mathbb{Z}$ is prime if and only if n is prime.

Definition 9.4.3. Suppose R is a commutative ring with unity and let $\mathcal{I}$ be a proper ideal of R. Then we say that $\mathcal{I}$ is a *maximal* ideal of R if, for any ideal $\mathcal{J}$ of R with $\mathcal{I} \subseteq \mathcal{J}$, either $\mathcal{J} = \mathcal{I}$ or $\mathcal{J} = R$.

Theorem 9.4.4. Suppose R is a commutative ring with unity and that $\mathcal{I}$ is a proper ideal of R. Then $\mathcal{I}$ is a maximal ideal if and only if $R/\mathcal{I}$ is a field.

Proof. As noted in the proof of Theorem 9.4.2, $R/\mathcal{I}$ is a commutative ring with unity, where the multiplicative identity is $1+\mathcal{I}$ by Exercise 9.2.3(c). So, in either direction, we need only address the existence of multiplicative inverses.

First, suppose $\mathcal{I}$ is a maximal ideal and let $x+\mathcal{I} \neq \mathcal{I}$. Define $\mathcal{J} = \{xr+a : r \in R, a \in \mathcal{I}\}$. We show that $\mathcal{J}$ is an ideal of R, $\mathcal{I} \subseteq \mathcal{J}$, and $\mathcal{J} \setminus \mathcal{I} \neq \emptyset$. To see that $\mathcal{I} \subseteq \mathcal{J}$, note that, for any $a \in \mathcal{I}$,

$$a = x \cdot 0_R + a \in \mathcal{J}.$$

To see that $\mathcal{J} \setminus \mathcal{I} \neq \emptyset$, note that

$$x = x \cdot 1_R + 0_R \in \mathcal{J} \setminus \mathcal{I}.$$

To see that $\mathcal{J}$ is an ideal of R, we first must show that $\mathcal{J}$ is a subgroup of R. Since $0_R \in \mathcal{I} \subseteq \mathcal{J}$, we see that $\mathcal{J} \neq \emptyset$. Now consider $xr+a, xs+b \in \mathcal{J}$, $r, s \in R$ and $a, b \in \mathcal{I}$. Then

$$(xr+a)-(xs+b) = x(r-s)+(a-b) \in \mathcal{J}.$$

So $\mathcal{J}$ is a subgroup of R. To see that $\mathcal{J}$ is an ideal of R, let $s \in R$ and $xr+a \in \mathcal{J}$, $r \in R$ and $a \in \mathcal{I}$. Then

$$s(xr+a) = x(rs)+as \in \mathcal{J}.$$

Hence, $\mathcal{J}$ is an ideal of R.

Since $\mathcal{I}$ is maximal, $\mathcal{J} = R$. Hence, $1 \in \mathcal{J}$ so we can let $y \in R$ and $a \in \mathcal{I}$ be such that $1 = xy + a$. Note that $xy - 1 \in \mathcal{I}$ which shows that

$$1 + \mathcal{I} = xy + \mathcal{I} = (x + \mathcal{I})(y + \mathcal{I}).$$

Hence, $y + \mathcal{I}$ is a multiplicative inverse for $x + \mathcal{I}$ and, since $x + \mathcal{I} \neq \mathcal{I}$ was arbitrary, we see that $R/\mathcal{I}$ is a field.

Now suppose $R/\mathcal{I}$ is a field and suppose $\mathcal{J}$ is an ideal of R such that $\mathcal{I} \subseteq \mathcal{J}$ and $\mathcal{J} \setminus \mathcal{I} \neq \emptyset$. Let $x \in \mathcal{J} \setminus \mathcal{I}$ and note that $x + \mathcal{I} \neq \mathcal{I}$. Since $R/\mathcal{I}$ is a field, there is $y + \mathcal{I} \in R/\mathcal{I}$ such that

$$1 + \mathcal{I} = (x + \mathcal{I})(y + \mathcal{I}) = xy + \mathcal{I}.$$

Note then that $1 - xy \in \mathcal{I} \subseteq \mathcal{J}$. Since $x \in \mathcal{J}$ and $\mathcal{J}$ is an ideal, $xy \in \mathcal{J}$. It follows that, since ideals are subgroups of the underlying group structure,

$$1 = (1 - xy) + xy \in \mathcal{J}.$$

Hence, $\mathcal{J} = R$ by Exercise 9.3.1. Conclusively, $\mathcal{I}$ is a maximal ideal of R. □

Exercise 9.4.1. Show that a commutative ring with unity is a field if and only if it has no proper nontrivial ideals.

Since every field is an integral domain, we immediately have that

Corollary 9.4.5. Every maximal ideal of a commutative ring with unity is a prime ideal.

However, there are prime ideals which are not maximal.

Example 9.4.6. Let $R = \mathbb{Z} \times \mathbb{Z}_2$ where the group structure is the direct product and let $\cdot$ be defined by $(a, b) \cdot (c, d) = (ac, b * d)$. A routine check verifies that R is a commutative ring with unity, where the multiplicative identity is $(1, 1)$. Then let $\mathcal{I} = \{(0, 0), (0, 1)\}$ and note that $\mathcal{I}$ is a nontrivial proper ideal of R. Consider the map $\varphi : R \to \mathbb{Z}$ defined by $\varphi(x, y) = x$. Routine computations verify that φ is a surjective homomorphism of rings and that $\ker(\varphi) = \mathcal{I}$. By The First Isomorphism Theorem, $R/\mathcal{I} \cong \mathbb{Z}$, which is an integral domain but not a field. Hence, $\mathcal{I}$ is a prime ideal which is not maximal. Indeed, note that

$$\mathcal{J} = \{(2k, j) : k \in \mathbb{Z}, j \in \mathbb{Z}_2\}$$

is a proper ideal of R which properly contains $\mathcal{I}$. ⊣

Chapter 10

Polynomial Rings

10.1 Polynomials Over a Ring

The reader surely will recall polynomials from previous studies, and though the intuitions built from previous experience will certainly align in many ways with the current treatment of polynomials over arbitrary commutative rings with unity, we emphasize here a distinction between polynomials as abstract algebraic objects and polynomials as functions. We will provide examples after the introductory definitions to motivate our development and highlight the distinction.

Throughout, we will employ the notation

$$\mathbb{N}_0 = \{k \in \mathbb{Z} : k \geq 0\}$$

for the set of nonnegative integers.

Definition 10.1.1. Let R be a commutative ring with unity and x be a symbolic *indeterminate* which is not a member of R. We then consider the infinite sequence of symbols

$$(x^k : k \in \mathbb{N}_0)$$

with the property that $x^k = x^j$ if and only if $k = j$ and identify the expression

$$\sum_{k=0}^{\infty} a_k x^k$$

with each sequence $(a_k : k \in \mathbb{N}_0)$ of elements of R. Via the identification with infinite sequences, we note that

$$\sum_{k=0}^{\infty} a_k x^k = \sum_{k=0}^{\infty} b_k x^k$$

if and only if $a_k = b_k$ for all $k \in \mathbb{N}_0$. One can verify that the set $R^{\mathbb{N}_0}$ of infinite sequences of elements of R with coordinate-wise addition

$$(a_k : k \in \mathbb{N}_0) + (b_k : k \in \mathbb{N}_0) = (a_k + b_k : k \in \mathbb{N}_0)$$

is a countably infinite direct sum of the group $(R, +)$ and, hence, a group in its own right. We then define the set of *polynomials* over R to be

$$R[x] = \left\{ \sum_{k=0}^{\infty} a_k x^k : \exists d \in \mathbb{N}_0 \ \forall j \geq d+1 \ (a_j = 0_R) \right\}.$$

We define $\deg : R[x] \to \mathbb{N}_0 \cup \{\texttt{ND}\}$ by the rule

$$\deg\left(\sum_{k=0}^{\infty} a_k x^k\right) = \begin{cases} \min\{d \in \mathbb{N}_0 : \forall j \geq d+1 \ (a_j = 0_R)\}, & \exists k \in \mathbb{N}_0 \ (a_k \neq 0_R); \\ \texttt{ND}, & \text{otherwise.} \end{cases}$$

For each $p \in R[x]$, $\deg(p)$ is called the *degree* of p where the token ND represents the phrase *not defined* and the polynomial

$$\sum_{k=0}^{\infty} 0_R x^k$$

is called the *zero polynomial*. We also use 0 to denote the zero polynomial. For

$$p = \sum_{k=0}^{\infty} a_k x^k,$$

we refer to each $a_k \in R$ for $k \in \mathbb{N}_0$ as a *coefficient* of p and, for $p \neq 0$, distinguish the coefficient a_n, where $n = \deg(p)$, as the *leading coefficient* of p. Note that, by definition, if a_n is the leading coefficient of a nonzero polynomial $p \in R[x]$, then $a_n \neq 0_R$.

We set $\texttt{ND} < 0$, and so $\max\{j, k\}$ for $j, k \in \mathbb{N}_0 \cup \{\texttt{ND}\}$ is defined. We also extend the usual addition on $\mathbb{N}_0$ to $\mathbb{N}_0 \cup \{\texttt{ND}\}$ by setting $\texttt{ND} + k = k + \texttt{ND} = \texttt{ND}$ for any $k \in \mathbb{N}_0 \cup \{\texttt{ND}\}$.

We note that the addition on $R[x]$ as inherited from the correspondence with infinite sequences under coordinate-wise addition is defined as

$$\left(\sum_{k=0}^{\infty} a_k x^k\right) + \left(\sum_{k=0}^{\infty} b_k x^k\right) = \sum_{k=0}^{\infty} (a_k + b_k) x^k.$$

The polynomial product is defined as

$$\left(\sum_{k=0}^{\infty} a_k x^k\right)\left(\sum_{k=0}^{\infty} b_k x^k\right) = \sum_{k=0}^{\infty} \left(\sum_{j=0}^{k} a_j b_{k-j}\right) x^k.$$

To simultaneously simplify notation and adhere to previous experience with polynomials, for $a \in R$ and $n \in \mathbb{N}_0$, let

$$ax^n = \sum_{k=0}^{\infty} b_k x^k$$

where

$$b_k = \begin{cases} 0_R, & k \neq n; \\ a, & k = n. \end{cases}$$

We also let $a = ax^0$ and $ax = ax^1$ for $a \in R$. Note then that, with these notational conventions, any polynomial of $R[x]$ can be written as

$$a_0 + a_1x + a_2x^2 + \cdots + a_nx^n = \sum_{k=0}^{n} a_kx^k$$

for some $n \in \mathbb{N}_0$ and $a_0, a_1, \ldots, a_n \in R$.

Despite the technical formality of the definition, this polynomial multiplication is no different than what you have surely learned in the past, which will be elaborated on below.

Due to the notation, it is easy to confuse abstract polynomials as defined here with polynomials as functions. To elaborate on this point, let's consider R^R, the set of all functions $R \to R$, for a ring R and define $\texttt{fn} : R[x] \to R^R$ by

$$\texttt{fn}\left(\sum_{k=0}^{n} a_kx^k\right)(r) = \sum_{k=0}^{n} a_kr^k.$$

Note that the defining feature of $R[x]$ requires that there are only finitely many nonzero coefficients. So $\texttt{fn}$ is defined. As we will elaborate by example, the $\texttt{fn}$ map need not be an injection, in general.

Example 10.1.2. Consider $\mathbb{Z}_2[x]$ and note that $x \neq x^2$ as polynomials. However, $\texttt{fn}(x) = \texttt{fn}(x^2)$. Indeed, note that

$$\texttt{fn}(x)(0) = 0 = 0^2 = \texttt{fn}(x^2)(0)$$

and that

$$\texttt{fn}(x)(1) = 1 = 1^2 = \texttt{fn}(x^2)(1).$$

Since $\mathbb{Z}_2 = \{0, 1\}$, $\texttt{fn}(x) = \texttt{fn}(x^2)$. ⊣

For this reason, we treat polynomials as abstract algebraic objects in the general theory instead of as functions.

Before we show that $R[x]$ is a commutative ring with unity, we prove a fact about finite sums.

Lemma 10.1.3. Suppose R is a ring. Let $k \in \mathbb{N}_0$ and suppose that, for each $(j, \ell) \in \mathbb{N}_0^2$ with $0 \leq j \leq k$ and $0 \leq \ell \leq j$, we have $f(j, \ell) \in R$. Then

$$\sum_{j=0}^{k}\sum_{\ell=0}^{j} f(j,\ell) = \sum_{\ell=0}^{k}\sum_{j=\ell}^{k} f(j,\ell).$$

Proof. Note that, by the constraints, $0 \leq \ell \leq j \leq k$. Since j ranges up to k, the possible values for ℓ range from 0 to k; so $0 \leq \ell \leq k$. For a fixed ℓ, note that j ranges from ℓ to k; so $\ell \leq j \leq k$. Then

$$\sum_{j=\ell}^{k} f(j,\ell)$$

is the portion of the sum

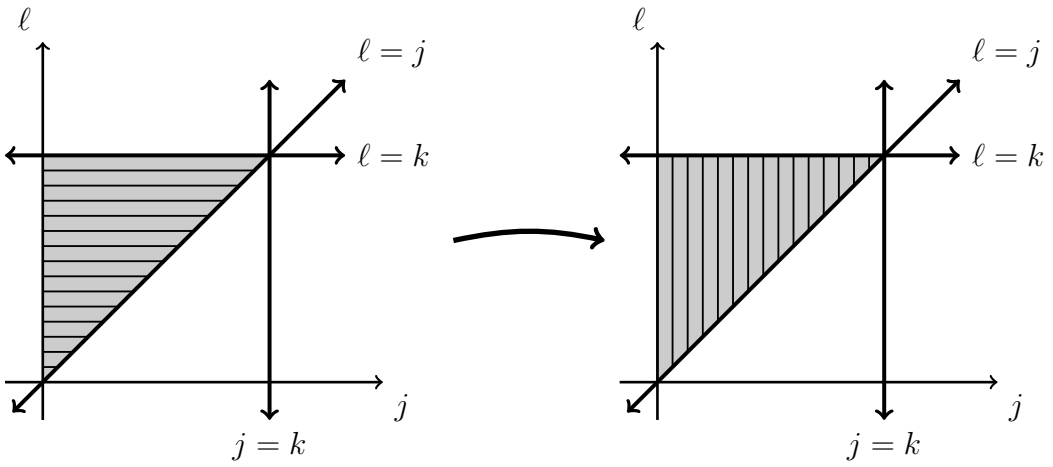
$$\sum_{j=0}^{k}\sum_{\ell=0}^{j} f(j,\ell)$$

corresponding to a fixed value of ℓ. Accumulating over ℓ, we obtain that

$$\sum_{\ell=0}^{k}\sum_{j=\ell}^{k} f(j,\ell) = \sum_{j=0}^{k}\sum_{\ell=0}^{j} f(j,\ell),$$

finishing the proof. □

Graphically, one can imagine the double sum in Lemma 10.1.3 as a region of integration of the form



Then, the swapping of the sums is seen graphically as changing the order of "integration" over the region.

Theorem 10.1.4. Let R be a commutative ring with unity. Then $(R[x], +, \cdot)$ for an indeterminate x is a commutative ring with unity.

Proof. Since we are formalizing $R[x]$ as a particular subset of the countably infinite direct sum $R^{\mathbb{N}_0}$, we will show that the infinite sequences of R which are eventually constantly zero form a subgroup. So suppose

$$\sum_{k=0}^{\infty} a_k x^k, \sum_{k=0}^{\infty} b_k x^k \in R[x]$$

and let $m_a, m_b \in \mathbb{N}_0$ be such that, for every $k \geq m_a + 1$, $a_k = 0_R$ and every $k \geq m_b + 1$, $b_k = 0_R$. Then let $m = \max\{m_a, m_b\}$ and observe that, for each $k \geq m + 1$, $a_k = 0_R$ and $b_k = 0_R$. Hence, for each $k \geq m + 1$, $a_k - b_k = 0_R$. It follows that

$$\left(\sum_{k=0}^{\infty} a_k x^k\right) - \left(\sum_{k=0}^{\infty} b_k x^k\right) = \sum_{k=0}^{\infty}(a_k - b_k)x^k \in R[x].$$

Hence, $(R[x], +)$ is a group. Note also that $R[x]$ is Abelian since direct sums of Abelian groups are Abelian.

We now show that the polynomial product is defined. So suppose

$$\sum_{k=0}^{\infty} a_k x^k, \sum_{k=0}^{\infty} b_k x^k \in R[x]$$

and let $m_a, m_b \in \mathbb{N}_0$ be such that, for every $k \geq m_a + 1$, $a_k = 0_R$ and every $k \geq m_b + 1$, $b_k = 0_R$. Then let $m = m_a + m_b$ and suppose $k \geq m + 1$. We will show that

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

So consider j where $0 \leq j \leq k$. We have two cases.

Suppose $j \leq m_a$. Then $-m_a \leq -j$ and so

$$m_b + 1 = m + 1 - m_a \leq k - m_a \leq k - j.$$

It follows that $b_{k-j} = 0_R$ and so $a_j b_{k-j} = 0_R$.

Otherwise, $j > m_a$ and $a_j = 0_R$ which implies that $a_j b_{k-j} = 0_R$.

Hence, for every $k \geq m + 1$,

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

and so

$$\left(\sum_{k=0}^{\infty} a_k x^k\right)\left(\sum_{k=0}^{\infty} b_k x^k\right) = \sum_{k=0}^{\infty}\left(\sum_{j=0}^{k} a_j b_{k-j}\right) x^k \in R[x].$$

Now, to see that $R[x]$ is commutative, note that

$$\begin{aligned}\left(\sum_{k=0}^{\infty} a_k x^k\right)\left(\sum_{k=0}^{\infty} b_k x^k\right) &= \sum_{k=0}^{\infty}\left(\sum_{j=0}^{k} a_j b_{k-j}\right) x^k \\ &= \sum_{k=0}^{\infty}\left(\sum_{j=0}^{k} a_{k-j} b_j\right) x^k \\ &= \sum_{k=0}^{\infty}\left(\sum_{j=0}^{k} b_j a_{k-j}\right) x^k \\ &= \left(\sum_{k=0}^{\infty} b_k x^k\right)\left(\sum_{k=0}^{\infty} a_k x^k\right).\end{aligned}$$

Since $R[x]$ is commutative, we need only show left-distributivity of $\cdot$ over $+$ to establish

the distributive properties of a ring for $R[x]$. Observe that

$$\begin{aligned}
&\left(\sum_{k=0}^{\infty} a_k x^k\right)\left(\sum_{k=0}^{\infty} b_k x^k\right) + \left(\sum_{k=0}^{\infty} a_k x^k\right)\left(\sum_{k=0}^{\infty} c_k x^k\right) \\
&= \left(\sum_{k=0}^{\infty}\left(\sum_{j=0}^{k} a_j b_{k-j}\right) x^k\right) + \left(\sum_{k=0}^{\infty}\left(\sum_{j=0}^{k} a_j c_{k-j}\right) x^k\right) \\
&= \sum_{k=0}^{\infty}\left[\left(\sum_{j=0}^{k} a_j b_{k-j}\right) + \left(\sum_{j=0}^{k} a_j c_{k-j}\right)\right] x^k \\
&= \sum_{k=0}^{\infty}\left[\sum_{j=0}^{k} (a_j b_{k-j} + a_j c_{k-j})\right] x^k \\
&= \sum_{k=0}^{\infty}\left[\sum_{j=0}^{k} a_j (b_{k-j} + c_{k-j})\right] x^k \\
&= \left(\sum_{k=0}^{\infty} a_k x^k\right)\left(\sum_{k=0}^{\infty} (b_k + c_k) x^k\right) \\
&= \left(\sum_{k=0}^{\infty} a_k x^k\right)\left[\left(\sum_{k=0}^{\infty} b_k x^k\right) + \left(\sum_{k=0}^{\infty} c_k x^k\right)\right].
\end{aligned}$$

To see that $R[x]$ has a multiplicative identity, note that, by our notational convention,

$$1_R = 1_R x^0 = \sum_{k=0}^{\infty} u_k x^k$$

where

$$u_k = \begin{cases} 1_R, & k = 0 \\ 0_R, & k > 0 \end{cases}$$

Then note that

$$\begin{aligned}
1_R \cdot \sum_{k=0}^{\infty} a_k x^k &= \left(\sum_{k=0}^{\infty} u_k x^k\right)\left(\sum_{k=0}^{\infty} a_k x^k\right) \\
&= \sum_{k=0}^{\infty}\left(\sum_{j=0}^{k} u_j a_{k-j}\right) x^k \\
&= \sum_{k=0}^{\infty} a_k x^k.
\end{aligned}$$

Since $R[x]$ is commutative, we see that 1_R satisfies the required properties of a multiplicative identity.

The final thing to show is that $\cdot$ is associative on $R[x]$. First consider

$$\left(\sum_{k=0}^{\infty} a_k x^k\right)\left(\sum_{k=0}^{\infty} b_k x^k\right) = \sum_{k=0}^{\infty}\left(\sum_{j=0}^{k} a_j b_{k-j}\right) x^k$$

and

$$\left(\sum_{k=0}^{\infty} b_k x^k\right)\left(\sum_{k=0}^{\infty} c_k x^k\right) = \sum_{k=0}^{\infty}\left(\sum_{j=0}^{k} b_j c_{k-j}\right) x^k,$$

and let, for $k \in \mathbb{N}_0$,

$$A_k = \sum_{j=0}^{k} a_j b_{k-j} \text{ and } B_k = \sum_{j=0}^{k} b_j c_{k-j}.$$

Observe that

$$\begin{aligned}\left(\sum_{k=0}^{\infty} a_k x^k\right)\left[\left(\sum_{k=0}^{\infty} b_k x^k\right)\left(\sum_{k=0}^{\infty} c_k x^k\right)\right] &= \left(\sum_{k=0}^{\infty} a_k x^k\right)\left(\sum_{k=0}^{\infty}\left(\sum_{j=0}^{k} b_j c_{k-j}\right) x^k\right)\\ &= \left(\sum_{k=0}^{\infty} a_k x^k\right)\left(\sum_{k=0}^{\infty} B_k x^k\right)\\ &= \sum_{k=0}^{\infty}\left(\sum_{j=0}^{k} a_j B_{k-j}\right) x^k\end{aligned}$$

and that

$$\begin{aligned}\left[\left(\sum_{k=0}^{\infty} a_k x^k\right)\left(\sum_{k=0}^{\infty} b_k x^k\right)\right]\left(\sum_{k=0}^{\infty} c_k x^k\right) &= \left(\sum_{k=0}^{\infty}\left(\sum_{j=0}^{k} a_j b_{k-j}\right) x^k\right)\left(\sum_{k=0}^{\infty} c_k x^k\right)\\ &= \left(\sum_{k=0}^{\infty} A_k x^k\right)\left(\sum_{k=0}^{\infty} c_k x^k\right)\\ &= \sum_{k=0}^{\infty}\left(\sum_{j=0}^{k} A_j c_{k-j}\right) x^k.\end{aligned}$$

So, to finish the proof, we will show that

$$\sum_{j=0}^{k} a_j B_{k-j} = \sum_{j=0}^{k} A_j c_{k-j}$$

for each $k \in \mathbb{N}_0$. So let $k \in \mathbb{N}_0$ and note that

$$\begin{aligned}
\sum_{j=0}^{k} a_j B_{k-j} &= \sum_{j=0}^{k} a_j \left(\sum_{\ell=0}^{k-j} b_\ell c_{k-j-\ell} \right) \\
&= \sum_{j=0}^{k} a_j \left(\sum_{\ell=j}^{k} b_{\ell-j} c_{k-j-(\ell-j)} \right) \\
&= \sum_{j=0}^{k} \left(\sum_{\ell=j}^{k} a_j b_{\ell-j} c_{k-\ell} \right) \\
&= \sum_{\ell=0}^{k} \left(\sum_{j=0}^{\ell} a_j b_{\ell-j} c_{k-\ell} \right) && \text{(by Lemma 10.1.3)} \\
&= \sum_{\ell=0}^{k} \left(\sum_{j=0}^{\ell} a_j b_{\ell-j} \right) c_{k-\ell} \\
&= \sum_{\ell=0}^{k} A_\ell c_{k-\ell} \\
&= \sum_{j=0}^{k} A_j c_{k-j}.
\end{aligned}$$

Thus, $\cdot$ is associative, and the proof is complete. □

Exercise 10.1.1. Let R be a commutative ring with unity. Using the formal definitions, show that, for $ax^k, bx^j \in R[x]$,

$$(ax^k)(bx^j) = abx^{k+j}.$$

Combined with the distributivity property of polynomial products, note that Exercise 10.1.1 shows that the standard polynomial product we have learned for polynomials with real coefficients is, at the core, the same polynomial product we have formally described here.

Exercise 10.1.2. Let R be a commutative ring with unity and let $p, q \in R[x]$.

(a) Show that $\deg(p+q) \leq \max\{\deg(p), \deg(q)\}$.

(b) Provide examples of two polynomials f and h such that

$$\deg(f+h) < \max\{\deg(f), \deg(h)\}.$$

Exercise 10.1.3. Let R be a commutative ring with unity and let $p, q \in R[x]$.

(a) Show that $\deg(pq) \leq \deg(p) + \deg(q)$.

(b) If R is an integral domain, show that $\deg(pq) = \deg(p) + \deg(q)$.

(c) If R is an integral domain, p is nonzero, and $\deg(pq) < \deg(p)$, show that $q = 0$.

(d) Find an example of a commutative ring with unity R which is not an integral domain and two polynomials $f, g \in R[x]$ such that $\deg(pq) < \deg(p) + \deg(q)$.

Remark. Note that, by Exercise 10.1.3(b), when R is an integral domain, $R[x]$ is also an integral domain.

Exercise 10.1.4. Show that, if $\mathbb{F}$ is a field, then $\mathbb{F}[x]$ is *not* a field.

Exercise 10.1.5. Show that, if $\varphi : R \to S$ is a homomorphism of rings, then $\Phi : R[x] \to S[x]$ defined by

$$\Phi(a_0 + a_1x + a_2x^2 + \cdots + a_nx^n) = \varphi(a_0) + \varphi(a_1)x + \varphi(a_2)x^2 + \cdots + \varphi(a_n)x^n$$

is a homomorphism.

We'll now discuss $\mathtt{fn} : R[x] \to R^R$ a bit more.

Proposition 10.1.5. Suppose R is a commutative ring with unity. Then the set R^R of all functions $R \to R$ with $+$ and $\cdot$ defined on R^R by

$$(f + g)(r) = f(r) + g(r)$$

and

$$(f \cdot g)(r) = f(r) \cdot g(r)$$

is a commutative ring with unity.

The proof of Proposition 10.1.5 is left as an exercise to the reader.

Proposition 10.1.6. Let R be a commutative ring with unity. The map $\mathtt{fn} : R[x] \to R^R$ is a homomorphism of rings.

Proof. Consider

$$p = \sum_{k=0}^{\infty} a_k x^k, q = \sum_{k=0}^{\infty} b_k x^k \in R[x].$$

We first show that

$$\mathtt{fn}(p + q) = \mathtt{fn}(p) + \mathtt{fn}(q)$$

as functions. So let $r \in R$ be arbitrary. Note that

$$\begin{aligned}
\texttt{fn}(p+q)(r) &= \texttt{fn}\left(\left(\sum_{k=0}^{\infty} a_k x^k\right) + \left(\sum_{k=0}^{\infty} b_k x^k\right)\right)(r) \\
&= \texttt{fn}\left(\sum_{k=0}^{\infty} (a_k + b_k) x^k\right)(r) \\
&= \sum_{k=0}^{\infty} (a_k + b_k) r^k \\
&= \sum_{k=0}^{\infty} (a_k r^k + b_k r^k) \\
&= \left(\sum_{k=0}^{\infty} a_k r^k\right) + \left(\sum_{k=0}^{\infty} b_k r^k\right) \\
&= \texttt{fn}\,(p)\,(r) + \texttt{fn}\,(q)\,(r).
\end{aligned}$$

Since r was arbitrary, we see that

$$\texttt{fn}(p+q) = \texttt{fn}(p) + \texttt{fn}(q).$$

Now we show that

$$\texttt{fn}(pq) = \texttt{fn}(p)\texttt{fn}(q).$$

So let $r \in R$ be arbitrary. Then

$$\begin{aligned}
\texttt{fn}(pq)(r) &= \texttt{fn}\left(\left(\sum_{k=0}^{\infty} a_k x^k\right)\left(\sum_{k=0}^{\infty} b_k x^k\right)\right)(r) \\
&= \texttt{fn}\left(\sum_{k=0}^{\infty}\left(\sum_{j=0}^{k} a_j b_{k-j}\right) x^k\right)(r) \\
&= \sum_{k=0}^{\infty}\left(\sum_{j=0}^{k} a_j b_{k-j}\right) r^k \\
&= \left(\sum_{k=0}^{\infty} a_k r^k\right)\left(\sum_{k=0}^{\infty} b_k r^k\right) \\
&= \texttt{fn}(p)(r) \cdot \texttt{fn}(q)(r).
\end{aligned}$$

Since r was arbitrary, we have that

$$\texttt{fn}(pq) = \texttt{fn}(p)\texttt{fn}(q),$$

which completes the proof. □

Proposition 10.1.7. Let R be a commutative ring with unity and $a \in R$. Then $E_a : R^R \to R$ defined by $E_a(f) = f(a)$ is a surjective homomorphism of rings.

Proof. For $f, g : R \to R$, note that

$$E_a(f+g) = (f+g)(a) = f(a) + g(a) = E_a(f) + E_a(g)$$

and that

$$E_a(fg) = (fg)(a) = f(a)g(a) = E_a(f)E_a(g).$$

Thus, E_a is a homomorphism of rings.

To see that E_a is surjective, simply note that, for any $r \in R$, the constant polynomial r has the property that $E_a(r) = r$. □

Consequently, as the composition of two homomorphisms, $f \mapsto f(a)$, $R[x] \to R$, is a homomorphism of rings.

Definition 10.1.8. We will refer to the homomorphism $f \mapsto f(a)$, $R[x] \to R$, as the *evaluation homomorphism at a*.

Ideals which are prime but not maximal arise quite naturally in the context of polynomial rings.

Example 10.1.9. In the ring $\mathbb{Z}[x]$, $\langle x \rangle$ is a prime ideal which is not maximal. To show this, we will argue that $\mathbb{Z}[x]/\langle x \rangle \cong \mathbb{Z}$. Consider the evaluation homomorphism $E_0 : \mathbb{Z}[x] \to \mathbb{Z}$ at 0 and note that $\langle x \rangle \subseteq \ker(E_0)$. Also note that E_0 is surjective since $x + k \in \mathbb{Z}[x]$ for each $k \in \mathbb{Z}$ and $E_0(x + k) = k$.

We claim that $\ker(E_0) = \langle x \rangle$. So let $p \in \ker(E_0)$ and write

$$p = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$

where each $a_j \in \mathbb{Z}$ for $0 \leq j \leq \mathbb{Z}$. Since $p(0) = 0$, we see that $a_0 = 0$ and so

$$p = a_1 x + a_2 x^2 + \cdots + a_n x^n = x(a_1 + a_2 x + \cdots + a_n x^{n-1}).$$

Since

$$q = a_1 + a_2 x + \cdots + a_n x^{n-1} \in \mathbb{Z}[x],$$

we see that $p = x \cdot q \in \langle x \rangle$. Then, by The First Isomorphism Theorem,

$$\mathbb{Z}[x]/\langle x \rangle = \mathbb{Z}[x]/\ker(E_0) \cong \mathbb{Z}.$$

Since $\mathbb{Z}$ is an integral domain but not a field, $\langle x \rangle$ is a prime ideal of $\mathbb{Z}[x]$ which is not maximal. In Example 10.2.12, we will provide an explicit proper ideal of $\mathbb{Z}[x]$ which properly contains $\langle x \rangle$. ⊣

10.2 The Polynomial Division Algorithm

In this section, we formalize polynomial long division in the context of polynomial rings over a given field.

Theorem 10.2.1. Let $\mathbb{F}$ be a field and suppose d is a nonzero polynomial of $\mathbb{F}[x]$. For any $p \in \mathbb{F}[x]$, there exist unique polynomials $q, r \in \mathbb{F}[x]$ such that $p = dq + r$ where $\deg(r) < \deg(d)$. In this context, q is called the *quotient* and r is called the *remainder*.

Proof. We will start by addressing the existence of $q, r \in \mathbb{F}[x]$. If $\deg(p) < \deg(d)$, then note that $q = 0$ and $r = p$ satisfy the equality $p = dq + r$ with $\deg(r) < \deg(d)$. Since $\deg(d) \geq 0$ and we have shown that the result holds for all $p \in \mathbb{F}[x]$ with $\deg(p) < \deg(d)$, we can use induction on $\deg(p)$. In particular, let $n \geq \deg(d)$ and suppose that, for each $p \in \mathbb{F}[x]$ with $\deg(p) < n$, there exist $q, r \in \mathbb{F}[x]$ such that $p = dq + r$ and $\deg(r) < \deg(d)$. Then let $p \in \mathbb{F}[x]$ be such that $\deg(p) = n$ and write

$$p = a_0 + a_1 x + \cdots + a_n x^n$$

and

$$d = b_0 + b_1 x + \cdots + b_m x^m$$

where $m = \deg(d) \leq n$. Then, since $0 \leq n - m$, note that

$$\deg(p - a_n b_m^{-1} x^{n-m} d) < \deg(p).$$

By the inductive hypothesis, there are $q', r \in \mathbb{F}[x]$ with

$$\left(p - a_n b_m^{-1} x^{n-m} d\right) = dq' + r$$

and $\deg(r) < \deg(d)$. Note then that

$$p = dq' + a_n b_m^{-1} x^{n-m} d + r = d\left(q' + a_n b_m^{-1} x^{n-m}\right) + r.$$

So $q = q' + a_n b_m^{-1} x^{n-m}$ and r satisfy $p = dq + r$ with $\deg(r) < \deg(d)$. Hence, we have established the existence of q and r for each $p \in \mathbb{F}[x]$.

To finish the proof, we prove the uniqueness of q and r. So suppose $q_1, q_2, r_1, r_2 \in \mathbb{F}[x]$ are such that

$$p = dq_1 + r_1 = dq_2 + r_2$$

with $\deg(r_1) < \deg(d)$ and $\deg(r_2) < \deg(d)$. Then note that

$$d(q_1 - q_2) = r_2 - r_1.$$

By Exercise 10.1.2,

$$\deg(r_2 - r_1) \leq \max\{\deg(r_2), \deg(-r_1)\} = \max\{\deg(r_2), \deg(r_1)\} < \deg(d).$$

Since

$$\deg(d(q_1 - q_2)) = \deg(r_2 - r_1) < \deg(d),$$

we see that $q_1 - q_2$ is the zero polynomial by Exercise 10.1.3. That is, $q_1 = q_2$ and, consequently, $r_1 = r_2$. □

As a refresher on polynomial long division, let's divide $4x^3 - 2x + 7$ by $2x^2 + x + 1$.

$$\begin{array}{r|llll} & 2x & -1 & & \\ 2x^2+x+1 & 4x^3 & +0x^2 & -2x & +7 \\ & 4x^3 & +2x^2 & +2x & \\ \hline & & -2x^2 & -4x & +7 \\ & & -2x^2 & -x & -1 \\ \hline & & & -3x & +8 \end{array}$$

Then

$$4x^3 - 2x + 7 = (2x^2 + x + 1)(2x - 1) + (-3x + 8).$$

Corollary 10.2.2. Let $\mathbb{F}$ be a field, $a \in \mathbb{F}$, and $p \in \mathbb{F}[x]$. Then there exists $q \in \mathbb{F}[x]$ such that

$$p = (x - a)q + p(a).$$

Proof. By Theorem 10.2.1, there are $q, r \in \mathbb{F}[x]$ such that $p = (x - a)q + r$ and $\deg(r) < \deg(x - a) = 1$. Hence, $\deg(r) \leq 0$, which means that r must be a constant function. Since $p(a) = r(a)$, r must be the constant polynomial taking value $p(a)$. □

With Theorem 10.2.1, we extend the usual notions of integer arithmetic to rings of polynomials.

Definition 10.2.3. For a commutative ring R with unity and $p, d \in R[x]$, we say that d *is a factor of* p, or *divides* p, denoted by $d \mid p$, if there is $q \in R[x]$ such that $p = dq$. Otherwise, we write $d \nmid p$.

Exercise 10.2.1. Let $\mathbb{F}$ be a field and $p, q \in \mathbb{F}[x]$. Show that $p \mid q$ and $q \mid p$ if and only if there is some $\lambda \in \mathbb{F}^*$ such that $p = \lambda q$.

Definition 10.2.4. A polynomial is said to be *monic* if its leading coefficient is 1.

Definition 10.2.5. For a field $\mathbb{F}$, a polynomial $g \in \mathbb{F}[x]$ is said to be a *greatest common divisor* of the polynomials $p, q \in \mathbb{F}[x]$, if g divides both p and q and, for any other polynomial $d \in \mathbb{F}[x]$ which divides both p and q, $d \mid g$. The polynomials p and q are said to be *relatively prime* if the constant 1 polynomial is a greatest common divisor of p and q. Since we will show in Theorem 10.2.13 that a greatest common divisor can either be the zero polynomial or be taken to be a unique monic polynomial after scaling by an element of $\mathbb{F}$, we will let $\gcd(p, q)$ denote the zero or monic greatest common divisor of p and q.

We first note that, in the context of $\mathbb{F}[x]$ for a field $\mathbb{F}$, $\gcd(0, 0) = 0$. Indeed, note that $0 \mid 0$ and that, for any $d \in \mathbb{F}[x]$, $d \mid 0$. So 0 satisfies the definition of being a greatest common divisor for 0 and 0.

On the other hand, for any nonzero polynomial p, note that, even though $p \cdot 0 = 0$, which implies that $p \mid 0$, $0 \nmid p$ since $\mathbb{F}[x]$ is an integral domain. So p is not a greatest common divisor of 0 and 0.

Hence, the only greatest common divisor of 0 and 0 is 0, itself. That is, $\gcd(0, 0) = 0$.

Definition 10.2.6. For a commutative ring R with unity and $p \in R[x]$, we say that $a \in R$ is a *zero*, or a *root*, of p if $p(a) = 0$; that is, if p is in the kernel of the evaluation homomorphism at a.

Corollary 10.2.7. Let $\mathbb{F}$ be a field and $p \in \mathbb{F}[x]$. Then $a \in \mathbb{F}$ is a zero of p if and only if $x - a$ is a factor of p.

Proof. First, suppose $x-a$ is a factor of p. Then, there is some $q \in \mathbb{F}[x]$ such that $p = (x-a)q$. Note that $p(a) = 0 \cdot q(a) = 0$.

On the other hand, suppose $p(a) = 0$. Then, by Corollary 10.2.2, there is some $q \in \mathbb{F}[x]$ such that $p = (x - a)q + p(a) = (x - a)q$. That is, $x - a$ is a factor of p. □

Corollary 10.2.8. Let $\mathbb{F}$ be a field and $p \in \mathbb{F}[x]$ be a nonzero polynomial with degree n. Then p can have at most n distinct zeros in $\mathbb{F}$.

Proof. Note that any polynomial of degree 0 is a nonzero constant polynomial and, hence, has no zeros. So we proceed by induction.

Suppose, for $k \geq 0$, that every polynomial of degree $\leq k$ has at most k distinct zeros in $\mathbb{F}$. Then let p be a polynomial of degree $k + 1$.

If p has no zeros, we are done. So suppose p has at least one zero, $a \in \mathbb{F}$. By Corollary 10.2.7, $x - a$ is a factor of p so we can write $p = (x - a)q$ for a polynomial $q \in \mathbb{F}[x]$. Note that $\deg(q) < \deg(p) = k + 1$, and so the inductive hypothesis applies to assert that q can have at most k distinct zeros in $\mathbb{F}$. Appending a to that list of zeros shows that p can have at most $k + 1$ distinct zeros in $\mathbb{F}$. □

As we continue to elaborate on algebraic similarities between the ring of the integers and polynomial rings over a given field, we introduce a new category of rings.

Definition 10.2.9. A ring R is a *principal ideal domain*, abbreviated as *PID*, if R is an integral domain in which every ideal is a principal ideal.

As mentioned above, since ideals are subgroups, a consequence of Theorem 4.4.6 is that every ideal of $\mathbb{Z}$ is a principal ideal. So $\mathbb{Z}$ is a PID.

Theorem 10.2.10. For a field $\mathbb{F}$, $\mathbb{F}[x]$ is a PID. Moreover, every nontrivial ideal of $\mathbb{F}[x]$ is generated by a nonzero polynomial of minimal degree.

Proof. Let $\mathcal{I}$ be an ideal of $\mathbb{F}[x]$. If $\mathcal{I} = \{0\}$, then $\mathcal{I} = \langle 0 \rangle$. If $\mathcal{I}$ contains any nonzero $a \in \mathbb{F}$, then $1 \in \mathcal{I}$ which implies that $\mathcal{I} = \mathbb{F}[x] = \langle 1 \rangle$.

So, suppose $\mathcal{I}$ is a nontrivial proper ideal. Then $\mathcal{I}$ does not contain any constant polynomials other than the zero polynomial, so choose $g \in \mathcal{I}$ to have minimal positive degree. Note that, by the definition of ideals, $\langle g \rangle \subseteq \mathcal{I}$. Then, for any $p \in \mathcal{I}$, by Theorem 10.2.1, there are $q, r \in \mathbb{F}[x]$ such that $p = gq + r$ and $\deg(r) < \deg(g)$. Since $g \in \mathcal{I}$, $gq \in \mathcal{I}$. Hence, $r = p - gq \in \mathcal{I}$. Since g was assumed to have minimal positive degree in $\mathcal{I}$ and $\deg(r) < \deg(g)$, $r = 0$. Hence, $p = gq$. Since $p \in \mathcal{I}$ was arbitrary, we see that $\mathcal{I} \subseteq \langle g \rangle$, establishing that $\mathcal{I} = \langle g \rangle$. □

Example 10.2.11. $\mathbb{R}[x]/\langle x^2+1\rangle \cong \mathbb{C}$. Indeed, note that $\mathbb{R}[x]$ is a subring of $\mathbb{C}[x]$ and so the restricted evaluation $E_i : \mathbb{R}[x] \to \mathbb{C}$ defined by $E_i(p) = p(i)$ is a homomorphism of rings. Note that E_i is also surjective since, for $a, b \in \mathbb{R}$, $bx - ax^2 \in \mathbb{R}[x]$ and

$$E_i(bx - ax^2) = bi - a(i^2) = a + bi.$$

By Theorem 10.2.10, $\ker(E_i)$ must be a principal ideal generated by a polynomial of minimal positive degree. Note that $x^2 + 1 \in \ker(E_i)$ and that no nonzero polynomial of lesser degree is in $\ker(E_i)$. It follows that $\ker(E_i) = \langle x^2 + 1\rangle$ and so, by The First Isomorphism Theorem,

$$\mathbb{R}[x]/\langle x^2+1\rangle = \mathbb{R}[x]/\ker(E_i) \cong \mathbb{C}.$$

⊣

Example 10.2.12. The polynomial ring $\mathbb{Z}[x]$ is an integral domain but not a PID. As remarked above, a polynomial ring over an integral domain is, itself, an integral domain. So $\mathbb{Z}[x]$ is an integral domain. To see that $\mathbb{Z}[x]$ is not a PID, consider

$$\mathcal{I} = \{p \in \mathbb{Z}[x] : \exists k \in \mathbb{Z} \ \ (p(0) = 2k)\}.$$

To see that $\mathcal{I}$ is an ideal of $\mathbb{Z}[x]$, we first show that it is a subgroup of $\mathbb{Z}[x]$. First, observe that $0 \in \mathcal{I}$, so $\mathcal{I} \neq \emptyset$. Then, let $p, q \in \mathcal{I}$ and let $j, k \in \mathbb{Z}$ be such that $p(0) = 2j$ and $q(0) = 2k$. Note that

$$(p - q)(0) = p(0) - q(0) = 2j - 2k = 2(j - k).$$

So $p - q \in \mathcal{I}$.

To see that $\mathcal{I}$ absorbs multiplication, let $p \in \mathbb{Z}[x]$ and $q \in \mathcal{I}$. Then let $k \in \mathbb{Z}$ be such that $q(0) = 2k$. Note that

$$(pq)(0) = p(0)q(0) = 2\left(k \cdot p(0)\right).$$

So $pq \in \mathcal{I}$.

Now we show that $\mathcal{I}$ is not principal. So let $g \in \mathbb{Z}[x]$ be such that $\langle g\rangle \subseteq \mathcal{I}$. Note that $g \in \mathcal{I}$. If g is a constant polynomial, then $g = 2k$ for some $k \in \mathbb{Z}$. Note then that, for any $p \in \mathbb{Z}[x]$, $x \neq gp$ since every coefficient of gp is even. That is, $x \in \mathcal{I} \setminus \langle g\rangle$ and so $\mathcal{I} \neq \langle g\rangle$.

On the other hand, if g is a non-constant polynomial, then, for any $p \in \mathbb{Z}[x]$, gp either has strictly positive degree or $gp = 0$. In particular, $2 \neq gp$ which implies that $2 \in \mathcal{I} \setminus \langle g\rangle$. Hence, $\mathcal{I} \neq \langle g\rangle$.

Since the g above was chosen to be arbitrary, $\mathcal{I}$ is not principal. ⊣

The following is the polynomial analog to Bézout's Identity (Theorem 4.4.11).

Theorem 10.2.13. Let $\mathbb{F}$ be a field and suppose d is a greatest common divisor for two polynomials p and q in $\mathbb{F}[x]$. Then there are polynomials r and s for which

$$d = pr + qs.$$

Moreover, two polynomials, where at least one is nonzero, have a unique monic greatest common divisor.

Proof. As noted above, there is a unique greatest common divisor for $p = 0$ and $q = 0$; that is, $\gcd(0, 0) = 0$. Also note that $0 = 0 \cdot 0 + 0 \cdot 0$, so the conclusion of the theorem is satisfied for $p = 0$ and $q = 0$.

So suppose at least one of p or q is nonzero. Without loss of generality, suppose p is nonzero and consider the set

$$X = \{pa + qb : a, b \in \mathbb{F}[x]\}.$$

Note that, since $d \mid p$ and $d \mid q$, we can write $p = dt$ and $q = du$ for $t, u \in \mathbb{F}[x]$. It follows that, for any $a, b \in \mathbb{F}[x]$,

$$pa + qb = dta + dub = d(ta + ub).$$

Hence, $d \mid (pa + qb)$.

Now, since p is assumed to be nonzero, $p \cdot 1 + q \cdot 0 = p \in X \setminus \{0\}$, and so $X \setminus \{0\} \neq \emptyset$. Then let $m \in X \setminus \{0\}$ have minimal degree and set $a, b \in \mathbb{F}[x]$ to be such that $m = pa + qb$. Note that, by our choice of m, $m \neq 0$, so we can apply Theorem 10.2.1 to obtain $Q_1, Q_2, R_1, R_2 \in \mathbb{F}[x]$ such that $p = mQ_1 + R_1$ and $q = mQ_2 + R_2$ where $\deg(R_j) < \deg(m)$, for $j = 1, 2$. Note that

$$R_1 = p - mQ_1 = p - (pa + qb)Q_1 = p(1 - aQ_1) + q(-bQ_1) \in X$$

and that

$$R_2 = q - mQ_2 = q - (pa + qb)Q_2 = p(-aQ_2) + q(1 - bQ_2) \in X.$$

Since $R_j \in X$, for $j = 1, 2$, but $\deg(R_j) < \deg(m)$ and m was chosen to have minimal degree in $X \setminus \{0\}$, it must be the case that $R_j = 0$. That is, $p = mQ_1$ and $q = mQ_2$. Hence, $m \mid p$ and $m \mid q$. By the definition of greatest common divisors, $m \mid d$. Since $d \mid m$ by the comment above, Exercise 10.2.1 asserts that there is some $\lambda \in \mathbb{F}$ such that $m = \lambda d$. Since $m \neq 0$, $\lambda \neq 0$. Then, since $d = \lambda^{-1}m$,

$$d = \lambda^{-1}m = \lambda^{-1}(pa + qb) = p(\lambda^{-1}a) + q(\lambda^{-1}b).$$

Setting $r = \lambda^{-1}a$ and $s = \lambda^{-1}b$ thus provides that $d = pr + qs$.

For the final assertion, note that scaling m by the multiplicative inverse of its leading coefficient yields a monic greatest common divisor of p and q. Since the argument above establishes that every greatest common divisor of p and q differs from m by a scaling constant, this monic greatest common divisor is unique. □

In the proof of Theorem 10.2.13, note that the original greatest common divisor d necessarily divides all expressions of the form $pr + qs$. So we can use this result to compute greatest common divisors for polynomials.

Example 10.2.14. Using the polynomial version of the extended Euclidean algorithm, find the greatest common divisor of

$$p = 2x^4 - 10x^3 + 14x^2 - 10x + 12$$

and

$$q = 3x^3 + 6x^2 + 3x + 6.$$

First compute

$$\begin{array}{r|llll} & \frac{2}{3}x & -\frac{14}{3} & & \\ 3x^3+6x^2+3x+6 & 2x^4 & -10x^3 & +14x^2 & -10x & +12 \\ & 2x^4 & +4x^3 & +2x^2 & +4x & \\ \hline & & -14x^3 & +12x^2 & -14x & +12 \\ & & -14x^3 & -28x^2 & -14x & -28 \\ \hline & & & 40x^2 & & +40 \end{array}$$

Rewrite the remainder as $40(x^2+1)$ and then compute

$$\begin{array}{r|llll} & 3x & +6 & & \\ x^2+1 & 3x^3 & +6x^2 & +3x & +6 \\ & 3x^3 & & +3x & \\ \hline & & 6x^2 & & +6 \\ & & 6x^2 & & +6 \\ \hline & & & & 0 \end{array}$$

Note then that, since $(x^2+1) \mid q$ and

$$p = \left(\frac{2}{3}x - \frac{14}{3}\right) q + 40(x^2+1),$$

$(x^2+1) \mid p$. That is, x^2+1 is a common divisor for p and q. Then observe that

$$\begin{aligned} 40(x^2+1) &= p + q\left(-\frac{2}{3}x + \frac{14}{3}\right) \\ x^2+1 &= p \cdot \frac{1}{40} + q\left(-\frac{2}{120}x + \frac{14}{120}\right) \\ x^2+1 &= p \cdot \frac{1}{40} + q\left(-\frac{1}{60}x + \frac{7}{60}\right). \end{aligned}$$

Conclusively,

$$x^2+1 = \gcd(p,q).$$

⊣

10.3 Irreducible Polynomials

Definition 10.3.1. A non-constant polynomial $p \in \mathbb{F}[x]$, where $\mathbb{F}$ is a field, is *irreducible* (*over* $\mathbb{F}$) if it cannot be expressed as the product of two polynomials $q, r \in \mathbb{F}[x]$ with degrees that are smaller than the degree of p.[1] Otherwise, p is *reducible* (*over* $\mathbb{F}$).

[1] A more general context is explored by other texts where the ring of coefficients is only assumed to be an integral domain and the defining characteristic is that, if p can be written as the product qr, then one of q or r must be invertible in the ring. In such a context, the polynomial $2x^2-6$ would be reducible over $\mathbb{Z}$ since $2x^2-6 = 2(x^2-3)$ and neither 2 nor x^2-3 is invertible in $\mathbb{Z}[x]$.

Note immediately that every polynomial of the form $ax + b \in \mathbb{F}[x]$, where $a, b \in \mathbb{F}$, is irreducible. However, we are typically interested in irreducible polynomials of larger degree.

Remark. Note that any polynomial of degree 2 can only be factored into polynomials of lesser degrees if both factors are of degree 1. Similarly, a polynomial of degree 3 can only be factored into two polynomials of lesser degrees if one factor is of degree 2 and the other factor is of degree 1. Hence, for polynomials of degree 2 or 3, the question of reducibility becomes a question of the existence of zeros.

Example 10.3.2. Note that $x^2 + 1$ is irreducible over $\mathbb{R}$. ⊣

Example 10.3.3. Note that $x^2 - 2$ is irreducible over $\mathbb{Q}$ but reducible over $\mathbb{R}$. ⊣

Example 10.3.4. Note that $1 + x - x^2$ is irreducible over $\mathbb{Z}_2$ since it is a quadratic with no zeros in $\mathbb{Z}_2$, but that $1 + x - x^2$ is reducible over $\mathbb{R}$. ⊣

Example 10.3.5. Note that x^2+1 is reducible over $\mathbb{Z}_5$ since $2^2+1 \equiv 0 \pmod 5$. Completing the division algorithm yields that

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

⊣

However, in higher degrees, the existence of zeros becomes less relevant to the question of reducibility.

Example 10.3.6. The polynomial $x^4 - 5x^2 + 6$ is reducible over $\mathbb{Q}$, but has no zeros in $\mathbb{Q}$. Indeed, note that

$$x^4 - 5x^2 + 6 = (x^2 - 2)(x^2 - 3).$$

⊣

The irreducibility of a non-constant polynomial over a field can also be characterized in terms of the maximality of the corresponding principal ideal.

Theorem 10.3.7. For a field $\mathbb{F}$ and a non-constant polynomial $p \in \mathbb{F}[x]$, $\langle p \rangle$ is a maximal ideal of $\mathbb{F}[x]$ if and only if p is irreducible.

Proof. First, note that, if p is reducible, then $p = qr$ where q and r both have degrees strictly less than the degree of p. It follows that $\langle p \rangle \subseteq \langle q \rangle$. Moreover, since p is of greater degree than q, $q \in \langle q \rangle \setminus \langle p \rangle$. That is, $\langle p \rangle \subsetneq \langle q \rangle$. Now, note that q is non-constant since, otherwise, $\deg(qr) = \deg(r)$. Thus, no non-constant polynomial is a member of $\langle q \rangle$. That is, $\langle q \rangle$ is a proper ideal of $\mathbb{F}[x]$. Consequently, $\langle p \rangle$ is not a maximal ideal.

On the other hand, suppose $\langle p \rangle$ is not a maximal ideal and note that $\langle p \rangle \neq \langle 0 \rangle$ since p is non-constant. Then let $\mathcal{I}$ be a proper and necessarily nontrivial ideal of $\mathbb{F}[x]$ that properly contains $\langle p \rangle$. By Theorem 10.2.10, $\mathcal{I} = \langle q \rangle$ for some $q \in \mathbb{F}[x]$. Note that q must be non-constant since, otherwise, $\langle q \rangle$ would be trivial ideal. Since $p \in \langle p \rangle \subseteq \langle q \rangle$, there is some $r \in \mathbb{F}[x]$ such that $p = qr$. If r were a constant polynomial, then r would be invertible and $\langle q \rangle = \langle p \rangle$, which is contradictory to our assumptions. So r is of positive degree, which implies that both q and r are of degrees both strictly less than the degree of p. Hence, p is reducible. □

Exercise 10.3.1. Show that, for a field $\mathbb{F}$, if $p \in \mathbb{F}[x]$ is irreducible and $q, r \in \mathbb{F}[x]$ are such that $p \mid qr$, then either $p \mid q$ or $p \mid r$.

Another important application of Theorem 10.3.7 is the construction of finite fields of prime-power order.

Example 10.3.8 (A field of order 4)**.** Note that $x^2 + x + 1$ is irreducible over $\mathbb{Z}_2$ since it has no zeros in $\mathbb{Z}_2$. Hence, $\mathbb{Z}_2[x]/\langle x^2 + x + 1\rangle$ is a field of cardinality 4 since it is in bijective correspondence with $\{(a, b) : a, b \in \mathbb{Z}_2\}$. ⊣

We will now turn our attention to relationships between properties of polynomials in $\mathbb{Z}[x]$ and the reducibility of related polynomials over $\mathbb{Q}$. The following lemma shows that any polynomial in $\mathbb{Q}[x]$ can be obtained as a rational multiple of some polynomial of $\mathbb{Z}[x]$.

Lemma 10.3.9. Let $p \in \mathbb{Q}[x]$ be of degree n. Then there exist $a_0, a_1, \ldots, a_n, b, d \in \mathbb{Z}$ such that

$$\gcd(a_0, a_1, \ldots, a_n) = \gcd(b, d) = 1$$

and

$$p = \frac{b}{d}\left(a_0 + a_1 x + \cdots + a_n x^n\right).$$

Proof. Let

$$p = \frac{\alpha_0}{\beta_0} + \frac{\alpha_1}{\beta_1}x + \frac{\alpha_2}{\beta_2}x^2 + \cdots + \frac{\alpha_n}{\beta_n}x^n$$

and, for each $1 \le j \le n$,

$$\gamma_j = \frac{\alpha_j \beta_0 \beta_1 \cdots \beta_n}{\beta_j}$$

and note that each $\gamma_j \in \mathbb{Z}$ and that

$$\begin{aligned} p &= \frac{\beta_0 \beta_1 \cdots \beta_n}{\beta_0 \beta_1 \cdots \beta_n} \cdot \left(\frac{\alpha_0}{\beta_0} + \frac{\alpha_1}{\beta_1}x + \frac{\alpha_2}{\beta_2}x^2 + \cdots + \frac{\alpha_n}{\beta_n}x^n\right) \\ &= \frac{1}{\beta_0 \beta_1 \cdots \beta_n} \cdot \left(\gamma_0 + \gamma_1 x + \gamma_2 x^2 + \cdots + \gamma_n x^n\right). \end{aligned}$$

Let $\delta = \gcd(\gamma_0, \gamma_1, \ldots, \gamma_n)$ and, for each $1 \le j \le n$, $a_j = \dfrac{\gamma_j}{\delta}$. Finally, set

$$b = \frac{\delta}{\gcd(\delta, \beta_0 \beta_1 \cdots \beta_n)}$$

and

$$d = \frac{\beta_0 \beta_1 \cdots \beta_n}{\gcd(\delta, \beta_0 \beta_1 \cdots \beta_n)}$$

and observe that

$$\begin{aligned} p &= \frac{1}{\beta_0\beta_1\cdots\beta_n}\cdot\left(\gamma_0+\gamma_1x+\gamma_2x^2+\cdots+\gamma_nx^n\right)\\ &= \frac{1}{\beta_0\beta_1\cdots\beta_n}\cdot\left(a_0\delta+a_1\delta x+a_2\delta x^2+\cdots+a_n\delta x^n\right)\\ &= \frac{\delta}{\beta_0\beta_1\cdots\beta_n}\cdot\left(a_0+a_1x+a_2x^2+\cdots+a_nx^n\right)\\ &= \frac{b}{d}\left(a_0+a_1x+a_2x^2+\cdots+a_nx^n\right). \end{aligned}$$

Note that all of the desired properties hold by construction. □

Definition 10.3.10. Given a polynomial $p \in \mathbb{Z}[x]$,

$$p = a_0 + a_1x + \cdots + a_nx^n,$$

the *content* of p is $\gcd(a_0, a_1, \ldots, a_n)$. We say that p is *primitive* if its content is 1.

Note that Lemma 10.3.9 states that any $p \in \mathbb{Q}[x]$ can be expressed as the rational multiple of a primitive polynomial in $\mathbb{Z}[x]$.

Lemma 10.3.11 (Gauss's Lemma)**.** The product of two primitive polynomials is primitive.

Proof. Let P and Q be polynomials in $\mathbb{Z}[x]$ with the property that PQ is not primitive. Then let p be a prime number such that p divides each of the coefficients of PQ. Let $\varphi : \mathbb{Z}[x] \to \mathbb{Z}_p[x]$ be the map defined by

$$\varphi(a_0 + a_1x + \cdots a_nx^n) = (a_0 \,\%\, p) + (a_1 \,\%\, p)x + \cdots + (a_n \,\%\, p)x^n.$$

By Exercise 10.1.5, φ is a homomorphism of rings. Note that $\varphi(P)\varphi(Q) = \varphi(PQ) = 0$. Since $\mathbb{Z}_p[x]$ is an integral domain, either $\varphi(P) = 0$ or $\varphi(Q) = 0$; that is, either P or Q is not primitive, finishing the proof. □

Lemma 10.3.12. If a polynomial $P \in \mathbb{Z}[x]$ is reducible over $\mathbb{Q}$, then there exist non-constant polynomials $Q, R \in \mathbb{Z}[x]$ with degrees strictly less than the degree of P such that $P = QR$.

Proof. Let λ be the content of P and let $P' \in \mathbb{Z}[x]$ be the primitive polynomial such that $P = \lambda P'$. Since P is reducible over $\mathbb{Q}$, let $Q', R' \in \mathbb{Q}[x]$ be such that $P = Q'R'$ where the degrees of Q' and R' are strictly less than the degree of P. By Lemma 10.3.9, we can let $b_1, b_2, d_1, d_2 \in \mathbb{Z}$, with $d_1 \neq 0$ and $d_2 \neq 0$, and $Q, R \in \mathbb{Z}[x]$ be such that $Q' = \frac{b_1}{d_1}Q$, $R' = \frac{b_2}{d_2}R$, both Q and R are primitive, and $\gcd(b_1, d_1) = \gcd(b_2, d_2) = 1$. Note now that

$$\lambda P' = P = Q'R' = \frac{b_1b_2}{d_1d_2}QR \implies d_1d_2\lambda P' = b_1b_2QR.$$

Now, $d_1d_2\lambda P' \in \mathbb{Z}[x]$ and $b_1b_2QR \in \mathbb{Z}[x]$. Note that the content of $d_1d_2\lambda P'$ is $d_1d_2\lambda$ and that, since QR is a primitive polynomial by Gauss's Lemma, the content of b_1b_2QR is b_1b_2. Since $d_1d_2\lambda P' = b_1b_2QR$, this implies that $d_1d_2\lambda = b_1b_2$. Hence, $P' = QR$. The conclusion of the theorem is proved by noting that $P = \lambda P' = \lambda QR = (\lambda Q)R$. □

Example 10.3.13. We show that $p = 24x^2 + 13x - 2$ can be factored in the way of Lemma 10.3.12. By the quadratic formula, we know that that roots of p are $x = \frac{1}{8}$ and $x = -\frac{2}{3}$. Then it must be the case that

$$\begin{aligned} 24x^2 + 13x - 2 &= 24\left(x - \frac{1}{8}\right)\left(x + \frac{2}{3}\right) \\ &= 8 \cdot 3\left(x - \frac{1}{8}\right)\left(x + \frac{2}{3}\right) \\ &= (8x - 1)(3x + 2). \end{aligned}$$

⊣

Theorem 10.3.14 (The Schönemann–Eisenstein Criterion)**.** Let $P \in \mathbb{Z}[x]$ be written as

$$P = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$$

and suppose p is a prime number with the property that

- $p \mid a_j$ for each $0 \leq j < n$,
- $p \nmid a_n$, and
- $p^2 \nmid a_0$.

Then P is irreducible over $\mathbb{Q}$.

Proof. To prove this theorem, we will suppose that P is reducible over $\mathbb{Q}$, that $p \mid a_j$ for each $0 \leq j < n$, and that $p^2 \nmid a_0$. We will conclude that, necessarily, $p \mid a_n$. By Lemma 10.3.12, we can let $Q, R \in \mathbb{Z}[x]$ be non-constant polynomials with degrees strictly less than the degree of P such that $P = QR$. Let

$$Q = b_0 + b_1x + \cdots + b_mx^m$$

and

$$R = c_0 + c_1x + \cdots + c_\ell x^\ell.$$

Since $a_0 = b_0c_0$, $p \mid a_0$, and $p^2 \nmid a_0$, we conclude that p divides one of b_0 or c_0, but not both. So, without loss of generality, suppose $p \mid b_0$ but that $p \nmid c_0$.

We will now show, via induction, that $p \mid b_k$ for all $0 \leq k \leq m$. So, suppose that, for a given $k \geq 0$, $k < m$, we have shown that $p \mid b_j$ for all $j \leq k$. Consider the fact that, since $k + 1 \leq m < n$,

$$p \mid a_{k+1} = \sum_{j=0}^{k+1} b_jc_{k+1-j} = \left(\sum_{j=0}^{k} b_jc_{k+1-j}\right) + b_{k+1}c_0.$$

By the inductive hypothesis, $p \mid b_j$ for all $j \leq k$, so

$$p \mid \left(\sum_{j=0}^{k} b_jc_{k+1-j}\right)$$

which implies that

$$p \mid \left(a_{k+1} - \sum_{j=0}^{k} b_j c_{k+1-j} \right) = b_{k+1} c_0.$$

Since $p \nmid c_0$, it must be the case that $p \mid b_{k+1}$.

Consequently, $p \mid b_m$ and, since $a_n = b_m c_\ell$, we have that $p \mid a_n$. □

Theorem 10.3.14 is typically referred to as *Eisenstein's Criterion*, but the first proof, due to Schönemann, appeared (coincidentally in the same journal as Eisenstein's) four years prior to Eisenstein's. Though Eisenstein's Criterion has a somewhat limited scope of direct applicability, it does help in the generation of polynomials with integer coefficients that are irreducible over $\mathbb{Q}$ with arbitrarily large degree.

Example 10.3.15. The polynomial

$$p = x^9 + 3x^6 - 9x^5 + 21x^2 - 12x + 6$$

is irreducible over $\mathbb{Q}$ since 3 is a prime that divides each of the non-leading coefficients, $3^2 = 9$ does not divide the constant term, and 3 does not divide the leading coefficient. ⊣

Note that, without Eisenstein's Criterion, determining that p in Example 10.3.15 is irreducible over $\mathbb{Q}$ does not reduce to checking whether or not p has zeros over $\mathbb{Q}$, for a polynomial of degree 9 could fail to have a linear factor while still having a quadratic, cubic, or quartic factor. Note, moreover, that p *is* reducible over $\mathbb{R}$ as an odd-degree polynomial (consider end-behavior and the Intermediate Value Theorem).

Chapter 11

More on Integral Domains

11.1 Factorization in Integral Domains

The standard notion of divisibility in the integers which we expanded to the context of polynomial rings is a particular kind of relation which we will extend to the general context of commutative rings with unity (Definition 11.1.3).

Definition 11.1.1. For a set X, a relation $\lesssim$ on X is a *preorder* if $\lesssim$ is reflexive and transitive; that is, if the following two conditions hold:

- for every $x \in X$, $x \lesssim x$ and
- for all $x, y, z \in X$, $x \lesssim y \wedge y \lesssim z \implies x \lesssim z$.

Note that a symmetric preorder is an equivalence relation. Another common type of preorder is a partial order.

Definition 11.1.2. For a set X, a relation $\leq$ on X is a *partial order* if $\leq$ is reflexive, antisymmetric, and transitive; that is, if the following three conditions hold:

- for every $x \in X$, $x \leq x$,
- for all $x, y \in X$, $x \leq y \wedge y \leq x \implies x = y$, and
- for all $x, y, z \in X$, $x \leq y \wedge y \leq z \implies x \leq z$.

Definition 11.1.3. Let R be a commutative ring with unity. For $d, a \in R$ we say that d *divides* a, denoted by $d \mid a$, if there exists $b \in R$ such that $a = db$. In this case, we will also call d a *factor* of a.

Elements $a, b \in R$ are said to be *associates* if there exists an invertible element[1] $u \in R$ such that $a = ub$.

Exercise 11.1.1. Let R be a commutative ring with unity.

(a) Show that the *divides* relation is a preorder on R.

[1] Invertible elements are often called *units*.

(b) Show that 0_R is the greatest element of the *divides* preorder; that is, show that, for all $r \in R$, $r \mid 0_R$.

(c) Show that 0_R is also a maximal element of the *divides* preorder; that is, show that, if $0_R \mid r$, then $r = 0_R$.

(d) Show that 1_R is the least element of the *divides* preorder; that is, show that, for all $r \in R$, $1_R \mid r$.

(e) Show that, for $r \in R$, $r \mid 1_R$ if and only if r is invertible.

(f) Suppose R is an integral domain. Then show that a and b are associates if and only if both $a \mid b$ and $b \mid a$.

In fact, in the context of integral domains, we can capture the divisibility relation in terms of ideals.

Exercise 11.1.2. Let R be an integral domain and $a, b \in R$.

(a) Show that $a \mid b$ if and only if $\langle b \rangle \subseteq \langle a \rangle$.

(b) Show that a and b are associates if and only if $\langle a \rangle = \langle b \rangle$.

(c) Show that a is invertible if and only if $\langle a \rangle = R$.

(d) Show that, if $a \mid b$ and b is invertible, then a is invertible.

Exercise 11.1.3. For a commutative ring R with unity, and set $a \simeq b$ if a and b are associates.

(a) Show that $\simeq$ is an equivalence relation.

(b) Consider the set $R/_{\simeq} = \{[r] : r \in R\}$ of equivalences classes where

$$[r] = \{a \in R : r \simeq a\}.$$

Set $[a] \leq [b]$ if $a \mid b$. Show that $\leq$ is well-defined and that it is a partial order on $R/_{\simeq}$.

(c) Show that $[1]$ consists exactly of all invertible elements of R, that $[1]$ is minimal with respect to $\leq$ in the sense that, if $[r] \leq [1]$, then $[r] = [1]$, and that $[1]$ is least with respect to $\leq$ in the sense that $[1] \leq [r]$ for all $r \in R$.

(d) Show that $[0] = \{0\}$ is maximal with respect to $\leq$ in that sense that, if $[0] \leq [r]$, then $[r] = [0]$ and that $[0]$ is greatest with respect to $\leq$ in the sense that $[r] \leq [0]$ for all $r \in R$.

Definition 11.1.4. Let R be an integral domain. A nonzero and non-invertible element $r \in R$ is said to be *irreducible* if, whenever $r = ab$, either a or b is invertible. Note that the negation of this is that there are non-invertible and non-zero elements $a, b \in R$ such that $r = ab$.

Irreducible elements are thus those which cannot be factored into "simpler nontrivial" factors. In the partial order context of Exercise 11.1.3, irreducible elements correspond to minimal nontrivial equivalence classes.

Exercise 11.1.4. Let R be an integral domain and let $\simeq$ and $\leq$ be as in Exercise 11.1.3. Let $X = R/_{\simeq} \setminus \{[1], [0]\}$.[2] Then show that $r \in R$ is irreducible if and only if $[r]$ is minimal in X with respect to $\leq$ in the sense that, if $[a] \in X$ is such that $[a] \leq [r]$, then $[a] = [r]$.

Definition 11.1.5. Let R be an integral domain. A nonzero and non-invertible element $p \in R$ is said to be *prime* if, whenever $p \mid ab$, then either $p \mid a$ or $p \mid b$.

Exercise 11.1.5. Let R be an integral domain and show that an element $p \in R$ is prime if and only if $\langle p \rangle$ is a proper nontrivial prime ideal of R.

Exercise 11.1.6. Show that, if an element p of an integral domain R is prime and $p \mid a_1 a_2 \cdots a_n$ for an integer $n \geq 2$, then there is some positive integer $j \leq n$ such that $p \mid a_j$.

Lemma 11.1.6. Let R be an integral domain and $p \in R$ be prime. Then p is irreducible.

Proof. Suppose $p = ab$ for $a, b \in R$. Note that $p \mid ab$ and so either $p \mid a$ or $p \mid b$. Without loss of generality, suppose $p \mid a$. Then $a = pr$ for some $r \in R$. It follows that

$$p = ab = prb \implies p(1 - rb) = p - prb = 0.$$

Since p is nonzero and R is an integral domain, $1 - rb = 0 \implies 1 = rb$. That is, b is invertible. Therefore, p is irreducible. □

Theorem 11.1.7. Let R be a PID and $r \in R$. Then the following are equivalent:

(i) $\langle r \rangle$ is a maximal ideal of R.

(ii) $\langle r \rangle$ is a proper nontrivial prime ideal of R.

(iii) r is prime.

(iv) r is irreducible.

Proof. For commutative rings with unity, we have Corollary 9.4.5 which asserts (i) $\implies$ (ii). Then, since every PID is an integral domain, Exercise 11.1.5 guarantees that (ii) $\implies$ (iii) and Lemma 11.1.6 guarantees that (iii) $\implies$ (iv). So, to finish the proof, we show that (iv) $\implies$ (i).

Suppose r is irreducible and suppose that $\mathcal{I}$ is a proper ideal of R with $\langle r \rangle \subseteq \mathcal{I}$. Since R is a PID, $\mathcal{I} = \langle a \rangle$ for some $a \in R$. Hence, $\langle r \rangle \subseteq \langle a \rangle$ which implies that $a \mid r$ by Exercise 11.1.2. Also by Exercise 11.1.2, since $\mathcal{I} = \langle a \rangle$ was assumed to be proper, a is not invertible. Now, there must be some $b \in R$ such that $ab = r$ since $a \mid r$. Since a is not invertible and r is irreducible, b must be invertible. Hence, $a = rb^{-1}$ and so $r \mid a$ which implies that $\langle a \rangle \subseteq \langle r \rangle$. That is, $\langle r \rangle$ is a maximal ideal of R. □

[2] One can think of the removed equivalence classes as "trivial" since they serve as a global minimum and maximum, respectively.

Recall Example 10.1.9, where it was demonstrated that $\langle x \rangle$ is a prime ideal of the integral domain $\mathbb{Z}[x]$ which is not maximal. Hence, Theorem 11.1.7 reaffirms Example 10.2.12, which asserted that $\mathbb{Z}[x]$ is not a PID. However, $\mathbb{Z}[x]$ is not the correct context to find an irreducible element which is not prime.[3] For this, we turn to a certain subring of $\mathbb{C}$.

Example 11.1.8. Let $\mathbb{Z}[\sqrt{-5}] = \left\{a + b\sqrt{-5} : a, b \in \mathbb{Z}\right\}$. It is routine to verify that $\mathbb{Z}[\sqrt{-5}]$ is a subring of $\mathbb{C}$ with unity and, hence, an integral domain. Our claim is that 2 is an irreducible element of $\mathbb{Z}[\sqrt{-5}]$ which is not prime.

First, note that

$$\left(1 + \sqrt{-5}\right)\left(1 - \sqrt{-5}\right) = 1 + 5 = 6 = 2 \cdot 3.$$

So $2 \mid \left(1 + \sqrt{-5}\right)\left(1 - \sqrt{-5}\right)$. To justify our asserted claim, we must show that 2 is irreducible and that $2 \nmid \left(1 + \sqrt{-5}\right)$ and $2 \nmid \left(1 - \sqrt{-5}\right)$.

We define[4] $N : \mathbb{Z}[\sqrt{-5}] \to \mathbb{N}_0$ by

$$N(a + b\sqrt{-5}) = a^2 + 5b^2.$$

We observe that N is multiplicative; that is, note that

$$\begin{aligned}
N\left(\left(a + b\sqrt{-5}\right)\left(c + d\sqrt{-5}\right)\right) &= N\left((ac - 5bd) + (ad + bc)\sqrt{-5}\right) \\
&= (ac - 5bd)^2 + 5(ad + bc)^2 \\
&= a^2c^2 - 10abcd + 25b^2d^2 + 5a^2d^2 + 10abcd + 5b^2c^2 \\
&= a^2c^2 + 5b^2c^2 + 5a^2d^2 + 25b^2d^2 \\
&= (a^2 + 5b^2)c^2 + 5d^2(a^2 + 5b^2) \\
&= (a^2 + 5b^2)(c^2 + 5d^2) \\
&= N\left(a + b\sqrt{-5}\right) \cdot N\left(c + d\sqrt{-5}\right).
\end{aligned}$$

To see that 2 is irreducible, consider nonzero elements $a + b\sqrt{-5}, c + d\sqrt{-5} \in \mathbb{Z}[\sqrt{-5}]$ and note that, if either $b > 0$ or $d > 0$, then

$$N\left(\left(a + b\sqrt{-5}\right)\left(c + d\sqrt{-5}\right)\right) \geq 5 > 4 = N(2).$$

Hence, in the case that either $b > 0$ or $d > 0$,

$$\left(a + b\sqrt{-5}\right)\left(c + d\sqrt{-5}\right) \neq 2.$$

Then, for integers a, c, it is clear that, if $2 = ac$, then either $a = 1$ or $c = 1$. That is, 2 is irreducible.

We now show that $2 \nmid \left(1 + \sqrt{-5}\right)$ and $2 \nmid \left(1 - \sqrt{-5}\right)$. For any $\zeta \in \mathbb{Z}[\sqrt{-5}]$, note that $N(2\zeta) = N(2) \cdot N(\zeta) = 4 \cdot N(\zeta)$. So, if $\xi \in \mathbb{Z}[\sqrt{-5}]$ is such that $2 \mid \xi$, $4 \mid N(\xi)$. Note that

$$N\left(1 + \sqrt{-5}\right) = N\left(1 - \sqrt{-5}\right) = 6.$$

Since $4 \nmid 6$, we see that $2 \nmid \left(1 + \sqrt{-5}\right)$ and $2 \nmid \left(1 - \sqrt{-5}\right)$. That is, 2 is not prime. ⊣

[3] We will elaborate on this in §11.2.

[4] This N is typically known as the *norm*.

Recall that the Fundamental Theorem of Arithmetic asserts that any positive integer has a unique, up to ordering, prime factorization. Since $\mathbb{Z}$ is a PID, the prime numbers coincide precisely with the irreducible elements. In the general context of integral domains, the definition of irreducibility corresponds to the inability to be factored into "simpler" terms. Consequently, we will wish to discuss factorizations in terms of irreducible elements rather than prime elements, in general.

11.2 Unique Factorization Domains

Definition 11.2.1. We say that an integral domain R is a *unique factorization domain*, abbreviated as *UFD*, if, for every nonzero non-invertible element $r \in R$, there is a unique, up to order and units, finite sequence $p_1, p_2, \ldots, p_n$ of irreducible elements of R such that $r = p_1 p_2 \cdots p_n$. Formally, the uniqueness up to order and units is captured as follows. Let $p_1, p_2, \ldots, p_n$ be irreducible elements of R with the property that $r = p_1 p_2 \cdots p_n$. Then $p_1, p_2, \ldots, p_n$ are said to be a *unique, up to order and units, factorization of* r if, for any finite sequence $q_1, q_2, \ldots, q_m$ or irreducible elements of R such that $r = q_1 q_2 \cdots q_m$, $m = n$ and there is a bijection $b : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ such that, for each $1 \leq j \leq n$, q_j and $p_{b(j)}$ are associates.

Example 11.2.2. By an application of the Fundamental Theorem of Arithmetic, $\mathbb{Z}$ is a UFD. Note first that the only invertible elements of $\mathbb{Z}$ are 1 and -1. Then, suppose $n \in \mathbb{Z}$ is nonzero. Use the Fundamental Theorem of Arithmetic to find a prime factorization of $|n| > 0$. Such a factorization consists of irreducible elements and is unique up to order and units. Since $n = \pm|n|$ and -1 is a unit, this finishes the argument. ⊣

We will show here (Theorem 11.2.5) that, in fact, every PID is a UFD. Since, in the context of PIDs, irreducibility of an element corresponds to the maximality of the corresponding principal ideal, we will make use of some facts regarding ideals.

Lemma 11.2.3. Suppose R is a ring and that $\{\mathcal{I}_j : j \in \mathbb{N}\}$ is a sequence of two-sided ideals of R such that $\mathcal{I}_j \subseteq \mathcal{I}_{j+1}$ for each $j \in \mathbb{N}$. Then $\bigcup_{j=1}^{\infty} \mathcal{I}_j$ is a two-sided ideal of R.

Proof. First, let $\mathcal{J} = \bigcup_{j=1}^{\infty} \mathcal{I}_j$ and note that $0 \in \mathcal{I}_1 \subseteq \mathcal{J}$. So $\mathcal{J} \neq \emptyset$. To establish that $\mathcal{J}$ is a subgroup of R, let $a, b \in \mathcal{J}$. Then we can let $j_a, j_b \in \mathbb{N}$ be such that $a \in \mathcal{I}_{j_a}$ and $b \in \mathcal{I}_{j_b}$. Let $k = \max\{j_a, j_b\}$ and note that, by the hypothesis, $\mathcal{I}_{j_a} \cup \mathcal{I}_{j_b} \subseteq \mathcal{I}_k$. So $a, b \in \mathcal{I}_k$ and, since $\mathcal{I}_k$ is an ideal of R, $a - b \in \mathcal{I}_k \subseteq \mathcal{J}$. Hence, $\mathcal{J}$ is a subgroup of R.

Now, to show that $\mathcal{J}$ is a two-sided ideal of R, let $a \in \mathcal{J}$ and $r \in R$. Then let $j \in \mathbb{N}$ be such that $a \in \mathcal{I}_j$ and note that $ra \in \mathcal{I}_j \subseteq \mathcal{J}$ and $ar \in \mathcal{I}_j \subseteq \mathcal{J}$ since $\mathcal{I}_j$ is a two-sided ideal of R. Therefore, $\mathcal{J}$ is a two-sided ideal of R. □

Compare Lemma 11.2.3 to Exercise 9.3.3, which asks you to provide an example of two left-ideals of a ring such that their union is not an ideal. Such an example can be found in the context of a standard PID, and, thus, consist of two-sided ideals. The crucial difference between Exercise 9.3.3 and Lemma 11.2.3 is clearly the requirement in the lemma that the ideals be *ascending* with respect to the subset relation.

Lemma 11.2.4. Suppose R is a PID and that $\{\mathcal{I}_j : j \in \mathbb{N}\}$ is a sequence of ideals of R such that $\mathcal{I}_j \subseteq \mathcal{I}_{j+1}$ for each $j \in \mathbb{N}$. Then there exists some $m \in \mathbb{N}$ such that, for every $j \geq m$, $\mathcal{I}_j = \mathcal{I}_m$.

Proof. Let $\mathcal{J} = \bigcup_{j=1}^{\infty} \mathcal{I}_j$ and, by Lemma 11.2.3, note that $\mathcal{J}$ is an ideal of R. Since R is a PID, there is some $r \in R$ such that $\mathcal{J} = \langle r \rangle$. Note that $r \in \mathcal{J}$ and so there must be some $m \in \mathbb{N}$ such that $r \in \mathcal{I}_m$. It follows that, for each $j \geq m$,

$$\mathcal{J} = \langle r \rangle \subseteq \mathcal{I}_m \subseteq \mathcal{I}_k \subseteq \mathcal{J},$$

finishing the proof. □

Theorem 11.2.5. Every PID is a UFD.

Proof. We will first argue that any nonzero non-invertible element of a PID has an irreducible factor by way of the contrapositive. So suppose R is an integral domain and $r \in R$ is a nonzero non-invertible element with no irreducible factors. If r were itself irreducible, it would be an irreducible factor of itself, so r is not irreducible. Thus, there exist nonzero non-invertible elements $a_1, b_1 \in R$ such that $r = a_1 b_1$.

For $k \in \mathbb{N}$, suppose we have defined nonzero non-invertible non-irreducible elements

$$\{a_j : j \leq k\} \cup \{b_j : j \leq k\} \subseteq R$$

such that, for $j < k$, $b_j = a_{j+1} b_{j+1}$. Since b_k is not irreducible, we can find non-invertible $a_{k+1}, b_{k+1} \in R$ such that $b_k = a_{k+1} b_{k+1}$. If either a_{k+1} or b_{k+1} were irreducible, then r would have an irreducible factor, contrary to our assumptions. So a_{k+1} and b_{k+1} are nonzero, non-invertible, and not irreducible.

Now consider the sequence $\{b_n : n \in \mathbb{N}\}$ and note that, by construction, $b_{k+1} \mid b_k$ for each $k \in \mathbb{N}$. Verify also that, since a_{k+1} is non-invertible and R is an integral domain, $b_k \nmid b_{k+1}$. Hence, by Exercise 11.1.2, $\langle b_k \rangle \subsetneq \langle b_{k+1} \rangle$ for each $k \in \mathbb{N}$. Since we now have an infinite sequence of strictly increasing ideals, the conclusion of Lemma 11.2.4 fails which guarantees that R is not a PID. Thus, if R is a PID, then any nonzero non-invertible element of R has an irreducible factor.

Now we show the existence of a factorization in terms of irreducible elements of any nonzero non-invertible element of a PID. To accomplish this, we will suppose that R is an integral domain in which every nonzero non-invertible element has an irreducible factor but that there is some element of R which fails to have a finite factorization in terms of irreducible elements, and conclude that R is not a PID.[5] So let $r \in R$ be such that no finite sequence of irreducible elements has r as their product. In particular, r itself is not irreducible but has an irreducible factor p_1 by the hypothesis. We can then write $r = p_1 q_1$ for some $q_1 \in R$.

[5]To dissect the logic here, let ϕ be the statement "R is a PID", ψ be the statement "every nonzero non-invertible element of R has an irreducible factor," and ρ be the statement, "for every element r of R, there is a finite sequence of irreducible elements of R whose product is r." We have already shown that $\phi \to \psi$ and claim that, to show that $\phi \to \rho$, it suffices to show that $(\psi \wedge \neg\rho) \to \neg\phi$, where $\neg$ is logical negation. Note that the contrapositive is $\phi \to \neg(\psi \wedge \neg\rho)$, which is equivalent to $\phi \to (\neg\psi \vee \rho)$. Since $\phi \to \psi$, it will follow that ρ is necessarily true.

For $k \in \mathbb{N}$, suppose we have defined nonzero elements

$$\{p_j : j \leq k\} \cup \{q_j : j \leq k\} \subseteq R$$

where every p_j, $j \leq k$, is irreducible, every q_j, $j \leq k$, is nonzero and non-invertible, and, for each $j < k$, $q_j = p_{j+1}q_{j+1}$. Note that

$$r = p_1 p_2 \cdots p_k q_k,$$

and so q_k is not irreducible by the hypothesis. Since q_k is nonzero and non-invertible, q_k has an irreducible factor by the hypothesis. So we can write $q_k = p_{k+1}q_{k+1}$ for some irreducible p_{k+1} and some nonzero non-invertible q_{k+1}.

Now consider the sequence $\{q_n : n \in \mathbb{N}\}$ and note that, by construction, $q_{k+1} \mid q_k$ for each $k \in \mathbb{N}$, which establishes that $\langle q_k \rangle \subseteq \langle q_{k+1} \rangle$ for each $k \in \mathbb{N}$. Since p_{k+1} is irreducible, $q_k \nmid q_{k+1}$. Hence, $\langle q_k \rangle \subsetneq \langle q_{k+1} \rangle$ for each $k \in \mathbb{N}$. Since we now have an infinite sequence of strictly increasing ideals, the conclusion of Lemma 11.2.4 fails which guarantees that R is not a PID.

The last thing to show is that, in a PID, the uniqueness of the factorization up to order and units. To accomplish this, we will use induction on the length of the factorization and, throughout the argument, will invoke both Theorem 11.1.7 and Exercise 11.1.6.

So let R be a PID and, for the base case, suppose we have an element that has an irreducible factorization consisting of one irreducible term. That is, consider an irreducible element $p \in R$ and suppose that we have a sequence $q_1, q_2, \ldots, q_m$ of elements of R such that

$$p = q_1 q_2 \cdots q_m$$

and each q_j is either irreducible or invertible. Then $p \mid q_j$ for some $1 \leq j \leq m$ and, consequently, q_j cannot be invertible. By the assumptions, q_j must be irreducible and, hence, $pu_0 = q_j$ where u_0 is invertible. By reindexing the q_k's, we can assume, without loss of generality, that $j = 1$ and write

$$p = q_1 q_2 \cdots q_m = puq_2 \cdots q_m \implies 1 = u_0 q_2 \cdots q_m.$$

It follows that $q_2, \ldots, q_m$ are all invertible. Thus, $p = q_1 u$ where $u \in R$ is invertible.

Now suppose that, for $k \in \mathbb{N}$, we have shown that any element of R for which there exists a factorization in terms of no more than k-many irreducible elements has a unique irreducible factorization up to order and units. Then suppose $r \in R$ can be written as

$$r = p_1 p_2 \cdots p_k p_{k+1}$$

where each p_j is irreducible and let $q_1, q_2, \ldots, q_m$ be elements of R such that

$$r = q_1 q_2 \cdots q_m$$

and each q_j is either irreducible or invertible. Since $p_{k+1} \mid q_1 q_2 \cdots q_m$, there must be some j for which $p_{k+1} \mid q_j$. As in the base case, up to reindexing, we may, without loss of generality, suppose $j = 1$. Then, note that $p_{k+1} \mid q_1$ implies that q_1 is not invertible and, since q_1 must

thus be irreducible by the assumptions. Moreover, since $q_1 = p_{k+1}u_0$ for $u_0 \in R$, u_0 must be invertible. It follows that

$$\begin{aligned} &p_1p_2\cdots p_kp_{k+1} = p_{k+1}p_1p_2\cdots p_k = q_1q_2\cdots q_m = p_{k+1}u_0q_2\cdots q_m \\ \implies\ &p_1p_2\cdots p_k = u_0q_2\cdots q_m. \end{aligned}$$

Since we now have $s = u_0q_2\cdots q_m$ written as the product of k-many irreducible elements or R, we can apply the inductive hypothesis to assert that any factorization of s in terms of irreducible elements must consist of exactly k-many irreducible elements and be equivalent to $p_1p_2\cdots p_k$ up to scaling by an invertible element. Therefore, the factorization of r into irreducible elements is unique up to order and units. □

There are, however, examples of integral domains which are not UFDs. Indeed, we have already been acquainted with such an example in Example 11.1.8.

Example 11.2.6. Consider again

$$\left(1+\sqrt{-5}\right)\left(1-\sqrt{-5}\right) = 1+5 = 6 = 2\cdot 3$$

for $6, 2, 3, 1+\sqrt{-5}, 1-\sqrt{-5} \in \mathbb{Z}[\sqrt{-5}]$, as in Example 11.1.8, where it was shown that 2 is irreducible and not prime. In particular, note that 2 is neither an associate of $1+\sqrt{-5}$ nor $1-\sqrt{-5}$.

To see that 3 is irreducible, consider the fact that $N(3) = 9$ and that the possible values for $N(\zeta)$ below 9 for $\zeta \in \mathbb{Z}[\sqrt{-5}]$ are $0, 1, 4, 5, 6$. It follows that, for $\zeta, \xi \in \mathbb{Z}[\sqrt{-5}]$ with $N(\zeta) < 9$ and $N(\xi) < 9$,

$$N(\zeta\xi) = N(\zeta)\cdot N(\xi) \in \{0, 1, 4, 5, 6, 20, 24, 25, 30, 36\}.$$

Since

$$N(3) = 9 \notin \{0, 1, 4, 5, 6, 20, 24, 25, 30, 36\},$$

we see that 3 is irreducible.

A similar consideration shows that both $1+\sqrt{-5}$ and $1-\sqrt{-5}$ are irreducible.

To establish that none of the four elements $2, 3, 1+\sqrt{-5}, 1-\sqrt{-5}$ are associates in a single argument, we show that the only invertible elements of $\mathbb{Z}[\sqrt{-5}]$ are 1 and -1. So suppose $\zeta, \xi \in \mathbb{Z}[\sqrt{-5}]$ are such that $\zeta\xi = 1$. Then

$$1 = N(1) = N(\zeta\xi) = N(\zeta)\cdot N(\xi).$$

It follows that $N(\zeta) = N(\xi) = 1$. It is straightforward then to verify that $\zeta, \xi \in \mathbb{Z}$ and, moreover, that either $\zeta = \xi = 1$ or $\zeta = \xi = -1$.

Finally, since

$$6 = 2\cdot 3 = \left(1+\sqrt{-5}\right)\left(1-\sqrt{-5}\right)$$

and the two factorizations are not unique up to order and units, we see that $\mathbb{Z}[\sqrt{-5}]$ is not a UFD. ⊣